Design and Study of Evidence-Linked AI Support for Reviewing Multi-Source Student Synthesis

Software Engineering
Master's Degree Programme in Information and Communication Technology
Master of Science in Technology Thesis

Author:
Farah Tahir

Supervisors:
Mr.Tuomas Mäkilä
Ms. Xiaoran Han

May 2026

Master of Science in Technology Thesis
Department of Computing, Faculty of Technology
University of Turku

Subject: Software Engineering
Degree: Master's Degree Programme in Information and Communication Technology
Author: Farah Tahir
Title: Design and Study of Evidence-Linked AI Support for Reviewing Multi-Source Student Synthesis
Number of pages: 121 pages, 02 appendix pages
Date: May 2026

Reviewing student synthesis tasks in multi-source reading environments places a demanding cognitive load on teachers. When students work with several sources of varying quality, the synthesis text alone does not reveal which sources were used, whether critical ideas were covered, or whether the student's reading strategy was systematic. This thesis presents the design and evaluation of an augmented teacher panel for the LearnNet multi-source reading environment, intended to make student process data and AI-generated coverage assessments accessible to teachers alongside the synthesis text.

The panel was designed and implemented following a Design Science Research approach and evaluated through a qualitative case study with three teacher participants, who reviewed student cases in both manual and augmented conditions using a think-aloud protocol. Four research questions examined transparency of student understanding, alignment between AI and teacher judgments, panel efficiency and usefulness, and technical feasibility of the underlying large language model service.

Results showed that traceability and the specific student text passage that grounds it were the primary determinant of perceived AI adequacy. The efficiency benefit of AI-generated summaries was mediated by teacher domain familiarity. Narrative group reports were more interpretable for instructional planning than quantitative aggregation displays. The LLM service was technically feasible, with 55 of 74 (74.3%) session calls completing within 30 seconds.

A cross-cutting finding was that all three teachers engaged with the AI support in a verification-first mode, checking AI claims against student text evidence before accepting them. This pattern is interpreted as evidence that effective teacher-facing AI must be designed around conditional trust and inspectable reasoning - the teacher-in-the-loop principle - rather than accuracy alone.

Keywords: AI-assisted teacher support, multi-source synthesis, learning analytics, evidence-linked feedback, teacher-in-the-loop, design science research

UNIVERSITY OF TURKU

Table of Contents

UNIVERSITY OF TURKU

Glossary

Term	Definition
DSR	Design Science Research - a research paradigm in which the design and construction of an artifact constitutes a form of knowledge contribution, evaluated against practical and theoretical criteria.
LearnNet / Read2Learn	A controlled multi-source reading environment developed at the University of Turku, in which students read, annotate, and synthesise multiple sources on a given topic.
LLM	Large Language Model - a neural language model trained on large text corpora and capable of generating fluent text in response to structured prompts
.FeedbackAPI	A dedicated Python service developed for this thesis that assembles LLM input, calls the GPT-5 API, and applies post-processing validation before returning structured AI output to the teacher panel
Firebase	A cloud-hosted real-time database (Google) used in LearnNet to capture timestamped teacher interaction events (clicks, tab switches, panel activations)
Langfuse	An open-source LLM observability platform used to trace and log every API call made by the FeedbackAPI, including latency, token usage, and generation outputs.
P-codes	Predefined main ideas (12 in total) against which student syntheses are assessed; derived from the four relevant source texts in the biodiversity loss task (three per source)
Coverage assessment	An AI-generated judgment, per p-code, of whether and how well a student's synthesis addresses that predefined main idea, accompanied by a quoted evidence preview.
Think-aloud protocol	A data-collection technique in which participants verbalise their thoughts in real time while performing a task, enabling researchers to trace cognitive and interpretive processes.
NVivo	Qualitative data analysis software used to organise, code, and retrieve segments of transcribed think-aloud and interview data.
Source verification	A deterministic post-processing step that checks each AI-generated evidence preview against the closed-source corpus to confirm that the quoted or paraphrased passage is traceable to an original text
Teacher-in-the-loop	The empirically observed pattern, documented in this thesis, is one in which teachers engage with AI support in a verification-first mode - accepting AI claims provisionally and checking them against student text evidence before acting on them.

UNIVERSITY OF TURKU

List of Figures

List of Tables

UNIVERSITY OF TURKU

1 Introduction

Primary and lower-secondary pupils increasingly search for, judge, and synthesise information online. Yet many struggle to evaluate sources and to integrate ideas across texts in a coherent way. This thesis contributes to an ongoing University of Turku project - LearnNet - that builds a controlled web environment for practising “internet reading”: searching within a closed corpus, evaluating reliability and relevance, extracting main ideas, and composing a synthesis from multiple sources. The prior system work demonstrates the feasibility and value of such an environment for both research and instruction, and it motivates a new teacher-facing augmentation layer that surfaces process summaries and model-checked evidence to support faster, clearer judgments of students’ synthesis work.

1.1 Motivation and Scope

In multi-source reading and synthesis tasks, students work across several texts to produce a written response that integrates ideas from different sources. Classroom teachers need rapid, trustworthy insight into how students formed a synthesis, not just the final text. Existing project systems support the core learner workflow in a restricted search space and log-rich behavioural data; however, teachers must still open many artefacts and pages to infer coverage of main ideas and cross-text linking. This thesis targets that gap by designing and evaluating an augmented teacher panel that (i) summarises the student’s process, (ii) previews evidence mapped to reference ideas (“P-codes”), and (iii) reports link structure within and across texts. The scope includes the design and implementation of the augmentation service, the instrumentation needed for trustworthy logs, and an empirical study of its usefulness, transparency, and alignment with teacher judgments.

1.2 Research Questions

The study is structured around four research questions, each addressing a distinct dimension of the augmented teacher panel. The first two questions concern the quality and reliability of the AI-generated information; the third concerns whether that information actually changes how teachers work; and the fourth concerns whether the underlying service can be made robust enough for real classroom use.

RQ1 addresses transparency: **what process summaries and evidence previews help teachers form a quick and confident understanding of a student's synthesis process?** A teacher reviewing a multi-source synthesis task needs to reconstruct, from a written text and its associated activity log, how a student engaged with the available sources and which ideas were understood, missed, or misrepresented. The question is therefore not whether the panel provides information, but whether the specific elements it surfaces: synthesis evaluations, coverage indicators, evidence previews, and process data that reduce the cognitive effort required to reach a reliable judgment.

RQ2 addresses alignment: **to what extent do model-produced evidence previews overlap with teacher-identified passages or ideas in the provided sources?** This question treats the AI output as a candidate set of claims and asks how well that set corresponds to what an informed teacher would independently identify as relevant. Partial or full alignment is expected to correlate with teachers' willingness to act on AI suggestions without additional manual checking; low alignment signals that the AI is attending to different cues than the teacher and that the preview may mislead rather than support.

RQ3 addresses usefulness and efficiency: **whether the augmented panel reduces time-to-first-insight and the number of clicks required to reach key information, compared to the unaugmented panel, and how teachers rate the clarity and perceived effort involved.** Efficiency here is not an end in itself; it matters because a panel that requires more navigational effort than unassisted review does not improve the teacher's situation, even if the AI's assessments are accurate. The question, therefore, combines behavioural measures that are derived from interaction logs with teacher self-reports of clarity and cognitive demand.

RQ4 addresses technical feasibility and integrity: **whether the suggestion and evidence service can meet the latency and reliability constraints of a classroom setting, and whether it can prevent fabricated or unverifiable citations**. The integrity dimension is particularly important in an educational context. An AI service that occasionally presents plausible-sounding but groundless evidence previews would undermine teacher trust and could lead to inaccurate assessments of student work. The question, therefore, asks both whether the service is fast enough to be usable in practice and whether its outputs are anchored to verifiable passages in the closed-source corpus.

Taken together, the four questions reflect a design-oriented evaluation logic: RQ1 and RQ2 concern whether the artifact produces good information, RQ3 concerns whether it produces a better

experience, and RQ4 concerns whether it can be trusted as infrastructure. All four must receive satisfactory answers for the design to be considered a viable basis for further development.

1.3 Research Method

The study adopts the Design Science Research Methodology (DSRM) of Peffers et al. [35], supported by the foundational DSR principles of Hevner et al. [34], as its overarching framework. This positions the design and construction of the augmented teacher panel as a form of knowledge contribution in its own right, with decisions about the panel's architecture, data pipeline, and interface treated as design knowledge to be documented and justified alongside the empirical findings. The background chapter was developed through a structured literature review organised around four areas: multi-source reading and synthesis, teacher-facing dashboard design, large language models in educational settings, and AI ethics in K-12.

The empirical evaluation of the artifact follows a qualitative case study design [33], with three teacher participants. A within-subjects design was used in which each participant reviewed student synthesis cases under both a manual and an augmented condition within a single session. Data were collected across three streams: system interaction logs capturing teacher navigation and LLM service performance, and qualitative data comprising concurrent think-aloud protocols, micro-survey responses, and semi-structured retrospective interviews. Qualitative transcripts were analysed thematically using an inductive-deductive approach, with survey data summarised descriptively and triangulated against the qualitative findings. The study builds on the established instrumentation infrastructure of the LearnNet project and extends it with an LLM observability layer and an integrity verification pipeline. Full details of the research design, data collection, and analysis procedures are presented in Chapter 4.

1.4 Declaration of Generative AI

Generative AI tools were used during the preparation of this thesis in a limited and supervised capacity. Specifically, AI assistance was used to support language editing and the improvement of academic tone in selected passages, to aid in structuring arguments, and to assist in discussing relevant literature in a better way. To ensure absolute confidentiality, all sensitive project and research data were anonymized prior to being processed by the AI (only if absolutely required). All

ideas, arguments, interpretations, and conclusions presented in this thesis are my own. AI-generated text was not used directly or reproduced without review and revision.

1.5 Thesis Structure

The remainder of this thesis is organised as follows.

Chapter 2 reviews the related literature and prior work that inform the present design. It covers research on multi-source reading and synthesis in educational contexts, the theoretical basis for teacher feedback and formative assessment, prior iterations of the LearnNet platform and related project studies, and the broader landscape of learning analytics dashboards and AI-assisted teacher support tools. The chapter situates the augmented teacher panel within this body of work and identifies the gaps the design is intended to address.

Chapter 3 describes the context and problem in detail. It specifies the student workflow in the LearnNet environment, the data that the platform collects for each student, the transparency problem that the existing teacher view creates, and the design requirements that follow from that problem. The chapter introduces the teacher panel concept and defines the functional scope of the augmentation.

Chapter 4 presents the study design. It describes the Design Science Research and case study frameworks that structure the evaluation, the participant selection criteria and profiles, the session materials and task design, the data collection instruments, and the analytical procedures applied to each data stream. The chapter includes a discussion of the ethical arrangements and the measures taken to protect participant confidentiality.

Chapter 5 presents the technical architecture of the augmented panel. It describes the processing pipeline that transforms raw student activity into teacher-facing summaries, the integration with the GPT-5 API, the post-generation validation and source verification steps, and the caching model that balances responsiveness against computational cost. The chapter also covers the Langfuse observability integration and the data management arrangements that support trustworthy analytics.

Chapter 6 reports the results, organised by research question. Section 6.2 addresses RQ1 on transparency and panel element usefulness. Section 6.3 addresses RQ2 on alignment between AI assessments and teacher judgments. Section 6.4 addresses RQ3 on efficiency, usefulness, and

perceived effort. Section 6.5 addresses RQ4 on technical feasibility and integrity. Section 6.6 presents a thematic analysis of the think-aloud protocols across all three participants.

Chapter 7 discusses the findings in relation to the research questions, the related literature, and the broader question of how teacher-facing AI tools should be designed. It introduces the teacher-in-the-loop principle as an interpretive frame for the cross-cutting pattern observed in the data, draws design implications for future iterations of the panel and for similar systems in other contexts, and addresses the study's limitations and threats to validity.

Chapter 8 concludes the thesis. It summarises the main contributions at three levels: the design artifact, the empirical findings, and the design knowledge produced, and identifies the most productive directions for future work, including larger-scale evaluations, authentic classroom deployment, and the extension of the panel to group-level instructional planning.

2 Background

This chapter reviews the literature that underpins the design and evaluation of the teacher-facing support examined in this thesis. It is organised around four interconnected areas. The first is multi-source reading and synthesis, which establishes the educational task context and explains why this form of literacy is both important and difficult for learners. The second is teacher dashboards and process indicators, which covers what kinds of data open-ended learning tasks produce and what teachers need from a dashboard in order to use that data for instructional decisions. The third is large language models, which introduces the technical role LLMs can play in transforming unstructured student and source text into compact, interpretable support for teacher review. The fourth is the ethical and pedagogical principles that apply when AI is used in school settings, with particular attention to transparency, human oversight, and teacher-in-the-loop design. The chapter closes with a summary of the research gap that the thesis addresses.

2.1 Multi-source Reading and Synthesis

In contemporary education, students are increasingly expected to work with several texts on the same topic instead of relying on a single source. This is especially common in science education, inquiry-based learning, digital literacy, and other tasks where learners must locate relevant information, compare claims, and combine ideas into a written response [1]. Research on multiple-document comprehension describes this as a distinct form of literacy because readers must not only understand each text individually, but also evaluate how texts relate to one another and build an integrated understanding of the topic across sources [1].

A central product of this work is synthesis writing. In educational settings, synthesis tasks are used when students are asked to explain a phenomenon, compare perspectives, justify a position, or prepare a concise account from multiple sources [1]. Such tasks are important because they reflect the kinds of information practices required beyond school, where knowledge is often distributed across documents of varying relevance, reliability, and perspective [3]. At the same time, research shows that learners often find these tasks difficult. They may focus on isolated facts, fail to identify the most important ideas, or treat documents separately instead of integrating them into a coherent whole [3].

For this reason, multi-source reading and synthesis have become important topics both in educational research and in the design of digital learning environments. They are relevant not only for understanding student performance, but also for supporting teachers who must interpret how students selected, connected, and transformed ideas across texts. This makes multi-source reading and synthesis a suitable foundation for the present thesis, which examines how teacher-facing support can make these processes more visible and easier to evaluate in practice [5].

2.1.1 What is Multi-source Reading

Multi-source reading refers to the process of working with several texts that address the same topic or problem and constructing understanding across them instead of reading each text separately [1]. It differs from single-text comprehension because the reader must not only understand the content of each document, but also compare claims, notice overlaps and contradictions, evaluate how texts relate to one another, and combine relevant ideas into a shared representation of the topic [4]. In this sense, multi-source reading is not simply “reading many texts.” It is a coordinated activity that requires the reader to move between documents while maintaining a broader task goal [1].

Research on multiple-document comprehension describes this process as involving both content-level and source-level understanding [1]. At the content level, the reader identifies central ideas, detects relationships among them, and decides which ideas are relevant for the task. At the source level, the reader monitors where each idea came from, considers the credibility and usefulness of the source, and evaluates whether different documents reinforce, extend, or challenge one another [1]. These operations are especially important in educational tasks where the learner is expected to justify a claim, explain a phenomenon, or produce a synthesis from a set of provided materials [2].

Multi-source reading is, therefore, closely tied to strategic processing. Readers need to plan how to approach the task, allocate attention across documents, and revise their understanding as new information is encountered [3]. They must also keep track of how separate pieces of information fit together, which places demands on working memory and self-regulation [4]. If these strategies are weak, students may focus on surface details, overuse a single source, or fail to build meaningful links across texts [3].

This makes multi-source reading both educationally important and practically difficult. It is educationally important because it reflects the way knowledge is encountered in digital and academic environments, where information is distributed across several sources rather than presented as one complete answer [1]. It is difficult because success depends on a combination of comprehension, evaluation, and integration skills that do not automatically emerge from ordinary reading practice [2]. For this reason, researchers increasingly treat multi-source reading as a distinct competence that requires targeted instructional support and careful assessment [1].

2.1.2 Synthesis as an Outcome

Synthesis is commonly understood as the outcome of successful multi-source reading. It is a written product in which the reader transforms and integrates ideas from several texts into a coherent response that serves the purpose of the task [1]. This means that synthesis is more than summarising individual sources one after another. Instead, the student is expected to select relevant ideas, connect them meaningfully, and express them in a unified way in their own words [6].

A synthesis differs from simple note collection or copy-based writing because it requires the writer to reorganise information around relationships rather than around source boundaries [1]. For example, when students write a synthesis, they may group ideas by cause and effect, compare explanations from different texts, or combine complementary claims into a broader account of a phenomenon. In this way, synthesis becomes a sign that the student has moved beyond isolated comprehension toward integration [1].

Research has shown that this transformation is one of the most difficult parts of multiple-document work [3]. Students often retain a source-by-source structure in their writing, even when the task requires integration. They may reproduce phrases from the texts, repeat similar points, or omit the links that make the response coherent [2]. As a result, the final product may contain relevant content but still fail to function as a synthesis in the stronger sense used in educational research [1].

For teachers, synthesis is important because it makes visible not only what a student has understood, but also how the student has combined and prioritised information from several documents [6]. A synthesis can therefore reveal whether the student has identified the central ideas, whether the student recognises relationships among them, and whether the student can

communicate those relationships clearly. This is one reason why synthesis tasks are widely used in inquiry-oriented and text-based learning environments [1].

In the context of this thesis, synthesis is relevant because the teacher panel is designed to support the review of such products. To understand what kind of support is useful, it is first necessary to establish that synthesis is not only a writing task, but also the visible outcome of deeper processes of selection, evaluation, and integration across texts. [1]

2.1.3 Typical Difficulties and Instructional Needs

Although multi-source reading and synthesis are important educational goals, research shows that they are also demanding for learners. Students often identify information from individual texts but struggle to integrate those ideas into a coherent written response [3]. A common pattern is that learners reproduce ideas source by source instead of reorganising them around the demands of the task. This can lead to summaries that contain relevant content but still lack integration, comparison, or explanation [1].

Other typical difficulties include weak monitoring of source information, limited attention to conflicting claims, and problems in deciding which ideas are central and which are only peripheral [3]. These challenges are not only a matter of reading comprehension. They also involve self-regulation and task management. Students need to define the problem, decide what kind of information they need, evaluate whether a source is useful, and monitor whether their draft is moving toward a coherent outcome [2]. In open-ended digital tasks, this becomes more difficult because learners must divide their attention between reading, selecting, and writing, often while switching repeatedly between multiple documents. This means that even when students understand parts of the content, they may still fail to produce a strong synthesis because they do not manage the process effectively [6].

For this reason, prior research emphasizes the value of explicit support. Embedded instruction can guide learners through information problem-solving step by step, for example, by helping them clarify the task, judge source relevance, and organise ideas into a structured response [39]. Inquiry-based learning research similarly stresses that learners benefit when tasks are scaffolded through phases such as orientation, conceptualisation, investigation, and conclusion, because these phases help make the process visible and manageable [38]. Recent work on multiple-document

literacy also shows that digital scaffolds can improve performance when they make source information and relationships between documents more visible [7].

These findings are particularly relevant for teacher-facing support. If students vary widely in how they search, select, and connect ideas, then teachers need tools that reveal those differences in a concise and interpretable way [5]. The goal is not only to evaluate the final answer, but also to understand whether a student struggled with selecting ideas, connecting ideas, or turning those connections into a coherent written synthesis. This need for visibility provides an important rationale for dashboards, process indicators, and evidence-linked support, which are discussed in the following sections [5].

2.2 Teacher Dashboards and Process Indicators

2.2.1 Process Data and Indicators in Open-ended Tasks

Open-ended learning tasks produce rich traces of student activity. When students search, open texts, select information, and write responses in digital environments, these actions leave timestamped records that can be analysed as **process data** [5]. Such data differ from product-level outcomes because they describe *how* the learner approached the task rather than only *what* the learner finally produced. In multiple-source reading and synthesis tasks, the process data may include searches, source openings, bookmark and snippet actions, revisions to drafts, and the timing between these events. These traces have become an important basis for learning analytics because they provide a way to study strategic behaviour during complex tasks that unfold over time [5].

To make these traces useful for teachers or researchers, raw events are transformed into **process indicators**. Process indicators are interpretable measures derived from logs that summarise relevant aspects of a learner's activity [10]. Examples include the number of sources opened, the number of snippets selected from each source, the time to first draft, and the balance between reading and writing activity. Sequence-based indicators can also be used to describe behavioural paths, such as repeated switching between texts or long pauses before writing [9]. The purpose of these indicators is not to reduce learning to simple numbers, but to create concise representations of patterns that would otherwise remain hidden in event logs [9]

In open-ended tasks, such indicators matter because learners often follow very different paths even when they are given the same assignment [5]. Some students search widely before selecting

information, while others commit early to one or two sources. Some begin drafting quickly and revise heavily, whereas others delay writing until they have collected many snippets. Based on this variation, one can identify learners who may need different kinds of instructional support [5]. A learner who repeatedly searches without selecting useful information may need support in source evaluation. A learner who collects many ideas but struggles to start writing may need support in integration and organisation. Process indicators, therefore, help shift attention from general outcome judgments to more specific interpretations of where a student may be succeeding or struggling [5].

At the same time, the literature warns against over-interpreting process data. A count or timing measure does not by itself reveal understanding, motivation, or quality [5]. For example, longer time on task does not necessarily indicate deeper learning, because it may reflect confusion just as easily as persistence [11]. Similarly, many clicks can signal active exploration, but they can also reflect inefficient navigation. This means that process indicators need to be grounded in theory and interpreted in context [10]. Research on inquiry-based learning and information problem solving is useful here because it provides process models that connect observable events to meaningful task phases, such as orientation, investigation, evaluation, and conclusion [8]. When indicators are tied to such models, they become easier to interpret and more relevant for teacher-facing dashboards [10].

For this thesis, the importance of process indicators lies in their role as a foundation for teacher support. Teachers working with several student syntheses do not have time to reconstruct every path from raw logs. A compact set of indicators can help them notice patterns quickly and decide where to look more closely. This creates a bridge from open-ended learning analytics to dashboard design, which is the focus of the next subsection. [12]

2.2.2 What Teachers Need from Dashboards

Teacher dashboards are useful only when they reduce the effort required to turn student data into an instructional judgment. In practice, this means that dashboards should provide a quick overview, present information in a form that teachers can interpret without extra processing, and support a short path from observation to action [5]. Prior research on teacher-facing dashboards shows that teachers value systems that help them notice meaningful patterns at a glance, especially in environments where many actions and artefacts are generated over a short period of time [14]. A

dashboard that simply exposes raw records or disconnected numbers does not meet this need well, because the teacher must still reconstruct the story of the student's work manually [14].

A first design requirement is therefore an **overview**. Teachers need a concise picture of what the student did, how far the student progressed, and where the student may have struggled. This should not mean reducing everything to one score. Rather, it means selecting a small number of interpretable indicators that reveal useful patterns, such as whether the student searched widely, relied on very few sources, or delayed writing after collecting information [9]. Human-centred dashboard research in K-12 contexts similarly emphasizes that dashboards must be grounded in teacher goals and classroom realities, not only in what data happen to be available [13]. Teachers are more likely to use dashboards when the information is closely connected to their instructional decision-making and when the display is not overloaded with detail [13].

A second requirement is **traceability to evidence**. Teachers need to know where a claim comes from and whether they can verify it. Earlier dashboard research has shown that awareness alone is often not enough. If teachers cannot connect an indicator back to the underlying student activity or to the source material, the system risks becoming informative but not actionable [14]. For this reason, dashboards should allow the teacher to move from a compact summary to the relevant supporting information without losing context. In educational settings, this kind of traceability is also important for trust, because teachers remain responsible for explaining their interpretations to students and for justifying any feedback they give [14].

A third requirement is a **minimal click-path to action**. Teachers often use dashboards under time pressure, so the usefulness of a dashboard depends not only on what it shows but also on how efficiently it can be used [15]. If the route from noticing a pattern to checking the relevant evidence requires many interface steps, the value of the dashboard decreases. Research on teacher-facing dashboards, therefore, highlights the importance of actionable views, where the teacher can quickly move from overview to detail and from detail to a practical instructional response [14]. In this thesis, that requirement is especially relevant because the panel is intended to support the review of short syntheses rather than long-term analytics monitoring.

Taken together, the literature suggests that teacher dashboards should be selective, interpretable, and closely tied to instructional use. They should support quick understanding, preserve traceability to

underlying evidence, and keep the path from insight to action short. These principles form the background for the dashboard support studied in this thesis.

2.3 Large Language Models for Teacher Support

2.3.1 LLMs: A Brief Definition for Software Engineers

Large language models (LLMs) are neural language models trained to predict the next token in a sequence from a large amount of text data [20]. In practical terms, this means that the model learns statistical regularities of language and can then generate responses that follow an input prompt. Most current LLMs are based on the transformer architecture [17], which replaces sequential recurrence with self-attention, allowing more efficient parallel training and shorter dependency paths across positions in a sequence [17]. Because of this architecture, LLMs can process instructions, source excerpts, and draft text within a single context window and generate outputs conditioned on all of them together [18].

From a software engineering perspective, LLMs are often used as general-purpose inference services rather than as fixed task-specific models. The same underlying model can support several functions depending on the prompt, the surrounding system, and the constraints placed on the output [19]. In educational systems, this flexibility is relevant because one service can be used to produce evidence previews, concise summaries, or candidate feedback text without requiring separate models for each task. However, the broad capability of foundation models also introduces risks such as hallucinated content, unstable wording, and outputs that do not naturally follow application-specific formats [19]. For this reason, production use typically depends on constraining and validating the model's output rather than relying on free-form generation alone.

In engineering terms, prompting defines the role, task, and input context for the model, while generation parameters such as temperature affect the variability of the response. Lower temperature settings generally produce more repeatable outputs, which is useful in interfaces where consistency and auditability matter more than creativity [20]. In addition, many applied systems require the model to return structured data rather than unrestricted prose. This has led to the growing use of schema-constrained outputs, where the model is asked to return JSON-like responses that can be validated before use in the interface [19].

In this thesis, LLMs are treated as assistive components inside a larger teacher-facing system rather than as autonomous judges. The key point for the background is that LLMs make it possible to transform unstructured student and source text into machine-usable suggestions and summaries, but only when their outputs are bounded by appropriate interface, validation, and oversight mechanisms. This makes them relevant for dashboard support, but not sufficient on their own. Their value depends on how they are integrated into a trustworthy workflow, which is addressed in the following subsections. [19]

2.3.2 Where LLMs Add Value in Dashboards

The value of LLMs in dashboards lies in their ability to transform large amounts of unstructured text into compact, interpretable support for human decision-making [19]. Teacher dashboards often combine several kinds of information, such as written student responses, source texts, process traces, and summary indicators. While process indicators help describe what a learner did, they do not directly explain how ideas appear in the student’s writing or how those ideas relate to the source material. LLMs are useful in this gap because they can generate short, context-sensitive summaries and identify likely relationships between a draft and the texts it draws on [19]. This makes them relevant for dashboards that aim to help teachers move quickly from overview to informed interpretation.

One important contribution of LLMs is the generation of evidence-linked previews. In this role, the model does not replace teacher judgment. Instead, it proposes likely matches between parts of a student’s answer and relevant parts of the provided sources. This kind of support is valuable because it reduces the effort required to inspect many texts and to check whether a student's response appears to use the expected ideas [13]. In teacher-facing systems, such previews can support faster orientation by making the link between product and evidence more visible.This aligns with teacher-facing dashboard research, which highlights the value of making student activity visible to support timely teacher awareness [14].

A second contribution of LLMs is the generation of lightweight, editable suggestions. Because LLMs can summarise patterns across text and process information, they can assist in drafting teacher-facing prompts, concise feedback messages, or short descriptions of likely strengths and gaps [19]. In dashboard contexts, such suggestions are useful when they remain tentative and revisable. Their value is not that they automate the teacher's role, but that they can lower the burden

of producing a first response or highlight a point that deserves teacher attention. This is especially relevant in settings where teachers must review many student texts under time pressure [13].

LLMs may also add value at a group level. When information from several student cases needs to be summarised, teachers often need to identify patterns across the class rather than inspect every detail individually. In principle, LLMs can assist by producing concise descriptions of common strengths, recurring gaps, or typical evidence patterns across a set of texts [19]. This kind of support is potentially useful in class-level dashboards because it can connect individual cases to broader instructional needs. However, the value of such summaries depends on careful constraints and verification, since aggregation can amplify any errors already present at the individual level [19].

At the same time, the literature makes it clear that the usefulness of LLMs depends on how they are embedded in the system. Foundation models are powerful because they are flexible, but this same flexibility means they can produce unstable or misleading outputs if used without clear boundaries [19]. For dashboard use, the most useful role of LLMs is therefore not autonomous judgment, but bounded assistance that complements process indicators and teacher expertise. In this sense, LLMs add value when they improve visibility, reduce effort, and keep the teacher in control of interpretation and action [13].

2.3.3 Guardrails and Verification

The educational value of LLM support depends not only on what the model can generate, but also on how safely and transparently those outputs are handled. In teacher-facing dashboards, guardrails are necessary because even plausible model outputs may contain unsupported claims, incorrect citations, or wording that appears more certain than the evidence justifies [19]. This is especially important in educational contexts, where the system is used to support human judgment rather than to replace it. Recent work on explainable and trustworthy AI in education argues that AI-based support should make its reasoning traceable, expose uncertainty where relevant, and allow users to inspect the basis of a suggestion instead of accepting it as a black-box result [22].

One important guardrail is structured output. When LLM responses are constrained to a predefined format, the system can validate whether required fields are present and whether the output can be rendered consistently in the interface [19]. This matters for dashboards because teachers need stable

and interpretable views rather than free-form model text that may vary in structure from one response to the next

A second guardrail is verification against the original sources. If a dashboard presents evidence-linked support, then the quoted evidence should be checkable against the source material itself. Without such verification, teachers cannot know whether the system is surfacing real support from the texts or only plausible-sounding language [22].

A further requirement is source traceability. In multiple-source learning tasks, the system should indicate which source a suggested idea or quote comes from and ensure that the citation refers to a valid source in the task context. This supports teacher trust because it makes the model output auditable in the same way that other dashboard information should be auditable [13]. More broadly, these design choices align with the teacher-in-the-loop view of AI in education. The system should provide bounded and inspectable assistance, while the teacher remains responsible for interpretation and instructional response [22]. In this thesis, such guardrails are not only design principles but also part of the technical feasibility question addressed later in the study.

2.4 Ethics and Teacher-in-the-Loop in K-12

2.4.1 Core Principles for AI in Schools

The use of AI in school settings raises ethical questions that go beyond technical performance. In K-12 education, systems are used with minors, within institutional settings, and often in tasks that shape feedback, classroom participation, or perceptions of competence. For this reason, the literature emphasizes a small set of core principles that should guide AI design in schools: transparency, human oversight, age-appropriate design, and accountability [24].

Transparency means that teachers and learners should be able to understand what role the system plays, what kind of output it produces, and what its limitations are [24]. Human oversight means that AI should support professional judgement rather than replace it, especially in contexts related to assessment or instructional decisions [23]. Age-appropriate design requires interfaces and outputs that are understandable, non-deceptive, and suitable for school use, while accountability requires that system behaviour can be inspected, justified, and challenged when needed [22].

These principles are especially important in teacher-facing dashboards. A dashboard that combines analytics and AI-generated support can easily appear authoritative, even when its outputs are probabilistic or partial. If the system presents AI suggestions as facts, teachers may over-trust the output or find it difficult to explain its basis to students [22]. This is why many authors in AI in education argue for a teacher-in-the-loop approach, where the model supports interpretation but the teacher remains responsible for judgment, feedback, and instructional action [22]. In such an approach, AI is framed as *support, not scoring*. It can help surface candidate patterns, summarize evidence, or draft tentative suggestions, but it should not make final evaluative decisions on behalf of the teacher [23].

A related ethical issue is the treatment of student data. Learning analytics systems often rely on detailed behavioural traces, which can create tensions between pedagogical usefulness and data minimisation [23]. In school environments, this makes it especially important to use only the data needed for the stated purpose, to present indicators in forms that are understandable and contestable, and to avoid turning process traces into opaque judgements about students [24]. The literature therefore supports designs in which AI outputs remain bounded, inspectable, and clearly tied to visible evidence [22].

In the context of this thesis, these principles justify a design where AI-generated outputs are kept explanatory and assistive rather than decisive. Evidence previews are shown to support teacher review, not to replace it. Suggestions are editable and ignorable. Quotes and source links remain visible so that teachers can check the basis of the output. In this way, the system aligns with the broader ethical position that AI in K–12 should strengthen teacher agency and transparency rather than reduce them [24].

2.4.2 Why Teacher-in-the-loop Matters in Educational AI

The principle of keeping teachers in the loop is especially important when AI is used in assessment-related or feedback-related educational settings. In such contexts, the system may influence how student work is interpreted, what kinds of weaknesses are noticed, and what next steps are suggested. If these functions are automated too strongly, there is a risk that probabilistic model outputs begin to function as judgments rather than as support [22]. For this reason, recent literature in AI in education argues that teachers should remain the primary interpreters of evidence and the final decision-makers in matters that affect pedagogy, feedback, and evaluation [22].

Teacher-in-the-loop design matters for several reasons. First, teachers bring contextual knowledge that the system does not have. They understand the task, the classroom context, the learning goals, and the level of the students. A model may detect patterns in text, but it cannot fully account for the teacher's professional understanding of what counts as a meaningful response in a given educational situation [22]. Second, teacher involvement is important for fairness and explainability. If a student or another teacher asks why a particular interpretation or next-step suggestion was made, the human reviewer must be able to justify that decision in terms of the visible evidence rather than hidden model behavior [23].

This issue is also reflected in emerging European regulation. The EU AI Act treats several uses of AI in education as high-risk, especially where systems may influence access, evaluation, or educational opportunities. Such uses require strong human oversight, transparency, and the ability to intervene or override system outputs [25]. Although the present thesis does not build an automated assessment system, the same principle is relevant here. If an AI-enhanced dashboard is used in school contexts, the outputs should remain advisory, reviewable, and easy for the teacher to ignore or modify [25].

A teacher-in-the-loop approach also affects interface design. It suggests that AI outputs should be phrased as tentative support, not as final verdicts. It also means that the basis of an output should remain visible, for example, through source-linked evidence, interpretable process summaries, and concise suggestions that teachers can revise. Based on these principles, the author interprets this as a requirement for the system to remain advisory rather than evaluative [22]. In this thesis, the role of the AI is therefore supportive. It helps reduce effort in reviewing complex student work, but it does not replace teacher judgment. This positioning is essential both ethically and pedagogically, and it forms part of the rationale for the system design introduced later in the thesis [22].

2.5 Summary of the Background and Research Gap

The background literature shows that multi-source reading and synthesis are demanding forms of literacy because learners must identify relevant ideas, evaluate sources, and integrate information across texts into a coherent written response [1]. These demands are not limited to comprehension alone. They also involve planning, monitoring, and regulation during open-ended work, which makes process visibility important for both teaching and research [5].

At the same time, studies on teacher-facing dashboards suggest that useful support must be selective, interpretable, and closely tied to instructional action. Teachers need concise overviews, traceability to evidence, and interfaces that help them form judgments quickly without forcing them to reconstruct the entire student process from raw data [13], [14].

The literature on large language models suggests that they can add value in such settings when they are used in bounded and supportive ways. LLMs can help transform unstructured text into summaries, evidence-linked previews, and editable suggestions, but their usefulness depends on clear guardrails, source traceability, and human oversight [19]. In educational settings, especially with minors, these requirements are strengthened by ethical concerns around transparency, accountability, and the continued centrality of teacher judgement [22], [24]. This supports a teacher-in-the-loop design in which AI helps surface candidate interpretations while teachers remain responsible for evaluation and feedback.

Despite these advances, an important gap remains. Research has established the importance of multiple-document literacy, process-oriented support, dashboard interpretability, and trustworthy AI in education, but there is still limited work on how these strands can be combined in a teacher-facing review environment for short multi-source syntheses. In particular, there is little evidence on how process summaries and evidence-linked AI support can help teachers form a quick understanding of a student's work, how closely such previews align with teacher interpretations of source use, and whether these supports are useful under realistic classroom constraints [5], [13]. This thesis addresses that gap through a case study of the LearnNet environment, where process data, teacher dashboards, and LLM-based support are brought together and evaluated in a concrete educational setting.

3 Case Study: LearnNet

This chapter introduces the case context of the thesis. While the previous chapter focused on generic background concepts, the present chapter describes the educational environment in which the study was conducted and the specific problem that the thesis addresses. The case study is situated in **LearnNet**, an online learning environment developed at the University of Turku for practising science literacy, multiple-source reading, and synthesis writing in school and teacher education contexts. In earlier materials and related project outputs, the same project environment has also been referred to as **Read2Learn**, **KidNet**, or **Neuron**. In this thesis, the name *LearnNet* is used consistently for clarity. The environment is designed as a controlled system in which learners work with a predefined set of texts, evaluate and select relevant information, and produce a synthesis as a task outcome [26], [27].

The purpose of this chapter is to explain the project setting, the student and teacher workflows, and the scope of the present thesis within that setting. This is important because the contribution of the thesis is not the entire LearnNet system, but a specific teacher-facing extension that adds process summaries and evidence-linked AI support to the teacher panel.

3.1 The LearnNet Environment

3.1.1 Project Context and Purpose

LearnNet is a browser-based learning environment created to support the teaching and study of science literacy through controlled online reading tasks. The environment is built around the idea that students should practise not only reading individual texts, but also searching, evaluating, understanding, and synthesising information from multiple sources in ways that resemble contemporary online knowledge work. In the project's research and development work, such tasks are treated as part of a broader effort to strengthen science literacy and related transversal competences in school education [28].

A central design feature of LearnNet is that it limits and structures the information environment. Instead of allowing unrestricted internet search, the system presents a predefined collection of texts related to a given task. This makes it possible to study how learners search, select, and synthesise information under controlled conditions, while still preserving many of the complexities of

multi-source reading. In one recent study conducted in the same environment, student teachers were asked to write a 200–350-word synthesis based on ten online texts, some of which were relevant, some irrelevant, and some deliberately fake. The purpose was to examine how they constructed a coherent understanding of biodiversity loss from multiple sources [26].

The project also has a practical pedagogical purpose. Earlier development work in the same project area has examined the challenges teachers face when teaching science literacy in primary school and has used those findings to motivate the development of digital tools and learning materials. These studies show that teachers encounter difficulties related to students' information retrieval, reading evaluation, understanding of text, and direct copying of source texts, as well as challenges caused by the demanding nature of internet-based information environments [27]. Such findings provide the broader motivation for LearnNet as a teaching support environment and for the present thesis as a teacher-panel extension within it.

In this thesis, LearnNet serves as the case study setting in which teacher-facing process summaries and AI-supported evidence previews are designed and evaluated. The next subsections describe the student workflow, the teacher workflow, and the specific scope of the thesis within this broader environment.

3.1.2 Student Workflow in the Environment

The student workflow, as shown in Figure 3.1, in LearnNet is organised as a guided multi-source reading task in which the learner works within a predefined online environment rather than the open internet. Students are given a task prompt, a controlled set of source texts, and a writing goal that requires them to construct a synthesis from those materials. In the biodiversity-loss study conducted in the same environment, student teachers were asked to prepare for a school lesson by writing a 200–350-word synthesis based on ten online texts. The source set was intentionally mixed: some texts were relevant, some irrelevant, and some deliberately fake, which made source evaluation part of the task rather than a separate exercise [26]. This design reflects the broader purpose of the environment, which is to let students practise not only reading, but also searching, evaluating, understanding, and using online information in a bounded and researchable setting. [28]

In practical terms, the workflow moves from orientation to source use and finally to synthesis writing. Students first read the task prompt and inspect the available texts. They then decide which

texts are relevant for the task, compare ideas across those texts, and construct a written response that integrates multiple perspectives. The environment, therefore, supports a process in which students must recognise relevant ideas, distinguish reliable from unreliable or irrelevant information, and transform selected ideas into a coherent written product [26]. Earlier project materials and theses also frame the environment as one for practising scientific literacy and online information use, which means that the student workflow is not limited to reading comprehension alone. It encompasses broader skills in information retrieval, evaluation, interpretation, and synthesis. [27], [28], [30]

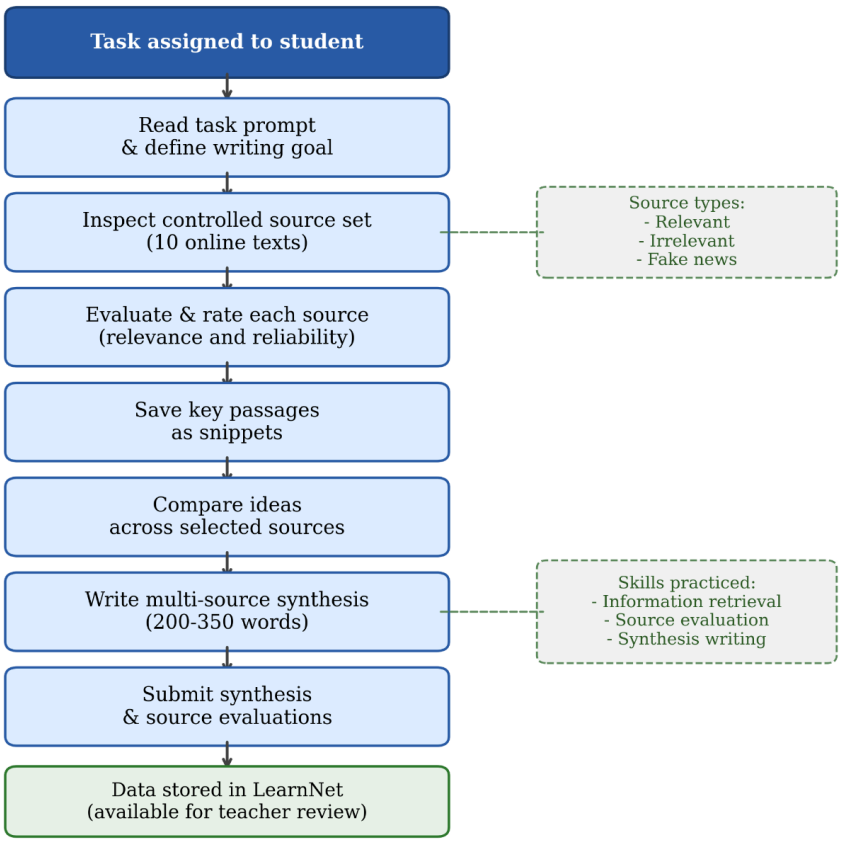

Figure 3.1. Student workflow in the LearnNet environment.

This workflow is pedagogically important because it exposes the difficulties that students face when working with several sources. Prior work in the same project area has shown that learners and student teachers may identify some key facts but still struggle to integrate them into a coherent whole, especially when irrelevant or misleading texts are present [26]. Teachers in related project

studies have also reported that practising online information use is important but difficult to support consistently in ordinary classroom work, especially when the number of sources is large or when the materials do not naturally guide students to evaluate and connect information [27]. In this sense, the student workflow in LearnNet is both a learning design and a research instrument: it creates conditions in which multiple-source reading and synthesis can be practised, observed, and later reviewed through the teacher panel.

3.1.3 Teacher Workflow and Existing Teacher Panel

The teacher workflow in LearnNet is organised around reviewing student task performance after or during the completion of a multi-source reading assignment. In practical terms, teachers need to see which students have completed the task, inspect the students' written synthesis, and review the materials and actions that contributed to that outcome. Because the task involves several texts and multiple process steps, the teacher's work is not limited to reading the final answer. It also includes checking which sources the student used, how the student evaluated those sources, and whether the student appears to have selected and combined relevant ideas appropriately. [26]

In earlier versions of the environment, the teacher panel primarily supported this work by presenting task and student data retrieved from the system database. The interface listed the participating students and allowed the teacher to open an individual student view containing the available task-related information, such as the written synthesis, selected sources, and related records generated during task completion [29], [30]. This made the underlying data accessible, but it still left much of the interpretation work to the teacher. In other words, the panel functioned mainly as a data access layer rather than as an interpretive support tool. Teachers could inspect what had been stored, but the system did not yet provide compact process summaries, aggregated indicators, or evidence-linked AI support that would help reduce the effort of reviewing multiple students in succession. [29], [30]

This limitation is important in the context of the present thesis. Prior work in the LearnNet project has shown that teaching science literacy and online information use is already demanding in classroom practice, and that teachers face difficulties in supporting students consistently when tasks require retrieval, evaluation, and synthesis from several sources [26]. If the teacher panel presents only raw or loosely structured information, the burden of reconstructing the student's process remains high. This becomes even more challenging when a teacher needs to review several students

or identify class-level patterns within a short time. For that reason, the teacher workflow in LearnNet provides a relevant starting point for the present study: the panel already contains the necessary educational data, but there is a clear need for more concise, teacher-facing support that makes the review process faster, more transparent, and easier to interpret. [29]

The present thesis builds on this existing teacher workflow rather than replacing it. The goal is to extend the panel with process summaries and evidence-linked support that help teachers move more efficiently from stored data to instructional insight. The details of that extension are introduced in the following sections.

3.2 The Synthesis Tasks used in this Study

The evaluation conducted for this thesis draws on tasks and materials developed within two interconnected research initiatives at the University of Turku. The first is the FINSCI project (Finnish Science Literacy), which investigates how scientific literacy can be taught, practised, and assessed across educational levels. The second is the Read2Learn project, which focuses on the research-based development of an online reading and synthesis environment in which learners practise critical, exploratory reading of multiple online sources. Critical and exploratory online reading is recognised as one of the most important citizen skills [32], yet it is not systematically taught, studied, or assessed even within Finland's formal education system [27]. Both teachers and learners need tools and support to engage with critical online reading and with the growing role of artificial intelligence in information production, since the internet, and in particular social media and AI-generated content, produces inaccurate and synthesised information that can be difficult to distinguish from verified knowledge [32].

The LearnNet environment, described in Section 3.1, serves as the shared platform for both projects. The synthesis tasks, the source materials, and the analytical coding scheme described in this section were designed within these projects and have been used in published studies on student teachers' conceptual understanding and scientific literacy in multi-source settings [26], [31]. The present thesis reuses the same task design, source set, and coding scheme, but applies them in a new context: the teacher review panel that is the focus of this work.

3.2.1 Multi-source Tasks and Source Materials

The synthesis tasks used in this study were designed as controlled multi-source reading assignments in which students worked with a fixed set of online texts inside the LearnNet environment. Rather than searching the open internet, learners completed the task in a bounded information space that had been prepared in advance for the pedagogical and research purposes of the environment [26], [28]. This design made it possible to observe how students selected, evaluated, and combined information when several sources were available on the same topic, while still keeping the task close to the kinds of online reading situations that learners increasingly face in contemporary education [32].

A central feature of these tasks is that the source set is intentionally varied. The materials are not limited to uniformly relevant scientific texts. Instead, the set includes texts that are directly relevant to the task, texts that are only loosely related or irrelevant, and texts that imitate credible information while containing misleading or false claims [26], [31]. This design choice serves two purposes. Pedagogically, it requires students to practise source evaluation rather than only information extraction. Analytically, it allows researchers and teachers to observe not only whether a student can find information, but also whether the student can distinguish useful information from distracting or misleading material. [26]

In the specific task used for this thesis, the topic was biodiversity loss. Students received a set of ten source texts, each approximately 300 to 400 words long and each presented with a title, an attributed author or organisation, a date, and a reference list meant to resemble a credible online publication [26], [31]. Of the ten texts, four were relevant to the task. These relevant sources were based on scientific knowledge and were shortened from materials provided by a domain expert in biodiversity research [26]. Each source illuminated a different dimension of the phenomenon. The first focused on insect decline as a visible part of the biodiversity crisis, covering habitat destruction and the use of pesticides as key causal factors. The second explained the broader processes of nature's impoverishment, examining how habitat loss, population growth, and the overconsumption of natural resources drive biodiversity decline. The third text framed biodiversity loss as a global societal challenge, linking it to the collapse of ecosystems, the loss of hereditary genetic diversity, and its direct impacts on human health, the economy, etc. The fourth source concentrated on the Amazon rainforest, describing its critical importance for global climate regulation through the carbon stored in its vegetation and soil, its function as the world's largest freshwater system, and the

range of pressures it faces from agriculture, livestock farming, logging, and other land-use activities [26], [31]. Importantly, no single relevant text was sufficient on its own. To produce a strong synthesis, students needed to draw on and integrate information from multiple relevant sources, each of which contributed a distinct angle on the topic.

The remaining six texts were not relevant to the task in the scientific sense. Four of these were irrelevant: they touched on topics related to nature or the environment but did not provide scientifically grounded content on biodiversity loss. [26], [31]. The final two texts were pseudoscientific. They were designed to mimic the appearance and language of credible scientific publications, but their content contained claims that contradict established scientific knowledge [31]. The inclusion of these fake sources made it possible to observe whether students were able to identify misleading material or whether they incorporated it into their syntheses uncritically.

The task output was a short written synthesis of 200 to 350 words. Students were instructed to use the LearnNet search function, evaluate all available sources and justify their evaluations, and then draw on four reliable sources to prepare a concise explanation of biodiversity loss from multiple research-based perspectives [26], [31]. In addition to the synthesis, the environment recorded which sources each student bookmarked, which passages they selected as snippets, and how they evaluated the relevance of each source. These three outputs, the synthesis text, the snippets, and the source evaluations, together constitute the student work that the teacher panel is designed to help a teacher review.

In earlier studies conducted in the same environment with first-semester primary student teachers, the task has shown that students generally succeed in defining biodiversity loss and identifying some of its causes and implications, but that they struggle to produce a coherent synthesis that integrates information from several sources, particularly when fake or irrelevant material enters the text [26], [31]. In one published study, nearly half of the participants used fake content in their synthesis, and many of the resulting texts were incoherent because they blended accurate scientific claims with pseudoscientific ones [31]. These findings confirm that the task generates meaningful variation in student performance, which in turn creates a realistic review challenge for a teacher. This is precisely what makes the biodiversity loss task well-suited for evaluating the teacher panel: the difficulty it produces, distinguishing strong from weak syntheses, identifying which sources a student relied on, and deciding what instructional follow-up is appropriate, is the kind of challenge the augmented panel is intended to support.

3.2.2 Main Ideas and Coding Scheme used in the Project

Because the source set for the biodiversity loss task was designed in advance, it was possible to identify before data collection the specific ideas that a well-constructed synthesis should contain. In the LearnNet project, these target ideas are referred to as main ideas and are given compact codes called P-codes. Each P-code represents a single propositional claim that appears in one of the relevant source texts and that contributes meaningfully to the phenomenon under study [26], [31]. The main ideas are not grades or evaluative labels. They are analytical anchors: a structured description of what the source material makes available for the student to find, select, and use.

The coding scheme for the biodiversity loss task assigns three main ideas to each of the four relevant source texts, yielding twelve main ideas in total. Each code follows the format P[source].[task].[idea], where the first number identifies the source text, the second identifies the task instance, and the third distinguishes the ideas within that source. This structure makes it possible to trace, for any given student, not only how many target ideas appeared in their synthesis or snippets, but also which sources those ideas came from and whether the student drew on multiple sources or relied heavily on one [26], [31].

To illustrate how the scheme works in practice, consider two examples. One of the relevant texts addresses insect decline as part of the biodiversity crisis. Its first main idea concerns insect loss as a visible and accelerating component of the broader biodiversity crisis. Its second main idea states that different insect species are essential to both natural ecosystems and human survival. Its third connects this to food production directly, framing the role of insects in pollination and food security as both a health and an economic concern. A student whose synthesis mentions insect decline in the context of the biodiversity crisis but says nothing about pollinators or food production would receive credit for the first main idea but not the others, even if the synthesis text reads well overall. This illustrates how the scheme can reveal gaps in coverage that are not visible from a general impression of the text.

A second example comes from the source on the Amazon rainforest. One of its main ideas states that the vegetation and soil of the Amazon region form a vast carbon store whose preservation is critically important for slowing global climate change. A student who references Amazon's importance for climate without connecting it specifically to carbon storage, or who discusses Amazon deforestation without mentioning the climate regulation dimension, would not fully satisfy

this main idea. This kind of precision is important for the teacher panel because the purpose of the P-code coverage display is not to score the student but to give the teacher a traceable picture of which ideas were engaged with and which were not, so that the teacher can decide what instructional follow-up is appropriate.

The coding scheme thus serves a dual function in this thesis. It supports the AI in generating structured, evidence-linked feedback by providing a defined set of reference claims against which student text can be compared. It also supports the teacher in verifying that feedback, because each identified main idea is linked to the specific passage in the source text from which it originates, allowing the teacher to check the AI's judgment against the original material.

In earlier published analyses using the same scheme, the main ideas have been used to assess conceptual understanding across four dimensions: definition of the phenomenon, causes and implications, different perspectives, and overall coherence [26]. The distribution of scores showed that while most students could define biodiversity loss at a basic level, fewer were able to present multiple perspectives or produce a coherent synthesis that drew on several relevant sources in an integrated manner [26]. The main ideas, in other words, reveal the structure of the student's reading and thinking, not only what was written, and this is precisely the kind of information the teacher panel is designed to make visible.

3.2.3 Individual and Group Review Needs in Teacher Work

When a teacher reviews the outcomes of a multi-source synthesis task, the review serves two analytically distinct purposes. At the individual level, the teacher needs to understand what a particular student did: which sources they drew on, which ideas they included or omitted, how they organised their synthesis, and what kind of feedback or instructional action would be most appropriate. At the group level, the teacher needs to identify patterns across the whole class: which ideas were broadly understood or consistently missing, which sources were over-relied upon or avoided, and whether any systematic misconceptions appear. These two levels of review call for different kinds of information and different kinds of support. [13]

Research on teacher-facing learning analytics dashboards has shown that teachers generally need information that is directly actionable and interpretable without extensive technical preparation [13]. When dashboards surface raw interaction logs or aggregate statistics without contextualisation, teachers often struggle to translate the data into concrete decisions about their

students [16]. When dashboards are designed around teachers' actual workflow needs, by contrast, they can meaningfully support awareness of student progress, help teachers identify students who require additional attention, and reduce the cognitive burden of handling large volumes of student data. [14]

The individual review challenge is particularly acute in tasks that produce written outputs linked to multiple source materials. To form a reliable picture of what a student did, a teacher must examine not only the synthesis text itself but also the sources the student evaluated, the passages they selected, and whether the connections between those choices and the final text are coherent [26]. In a digital multi-source environment with a large cohort, this kind of manual reconstruction becomes time-consuming, and the risk of missing important details is real: a student may have incorporated a pseudoscientific source convincingly, or may have ignored several relevant perspectives without this being apparent from the synthesis alone. [31]

Process data adds a further dimension. Research on learning process indicators has shown that the sequence and timing of a student's actions during a task can reveal aspects of their strategy and understanding that the final product alone does not capture [9]. A student who produces a modest synthesis after spending considerable time on the most relevant source texts is in a different instructional situation from a student who produced a similar-looking synthesis by briefly visiting every source and committing to none. However, process traces in their raw form are not easily read by a teacher during a practical review session and need to be condensed into summary information that is interpretable in the context of the task. [10]

The group review challenge has its own characteristics. When a teacher wishes to assess the class as a whole, they need to aggregate individual data in a way that highlights the distribution of understanding rather than the idiosyncrasies of any single case. Human-centred approaches to learning analytics dashboard design emphasise that class-level indicators should be aligned with pedagogically meaningful units, such as the specific ideas or skills a task was designed to develop, rather than generic engagement metrics alone [13]. In the context of a multi-source synthesis task with a predefined set of target ideas, class-level coverage of those ideas is a natural and pedagogically relevant indicator: it shows not only how well the class performed on average but also which conceptual gaps are widespread enough to warrant a whole-class instructional response. [13]

These considerations motivate the scope of the teacher panel examined in this thesis. The panel is intended to address both levels of review, using AI-generated summaries and evidence previews to reduce manual reconstruction work at the individual level and an aggregated group view to support class-level planning. The relationship between these components and the specific transparency problem they address is described in the following section.

3.3 Scope of this Thesis within LearnNet

The LearnNet environment is designed to support students in practising critical online reading and multi-source synthesis, and records a substantial amount of data about how each student engages with the task. What it had not provided, before the work described in this thesis, was a layer of support for the teacher who needs to review that engagement efficiently and reliably. The focus of this thesis is on designing, implementing, and evaluating exactly that layer: an augmented teacher panel that draws on process traces, LLM-generated evidence previews, and the structured main-idea coding scheme to give teachers transparent and actionable insight into student synthesis work.

3.3.1 Problem Addressed in the Teacher Panel

The starting point for the teacher panel is a transparency gap. LearnNet already collects, for each student, a complete record of their reading session: which sources they visited, how long they spent on task, which passages they saved as snippets, how they evaluated each source, and when they wrote their synthesis. All of this data exists in the system. What the standard teacher view (Figure 3.2) did not provide was a way for a teacher to see how the synthesis relates to all of this underlying activity without manually opening and cross-referencing multiple tabs and records for each student individually.

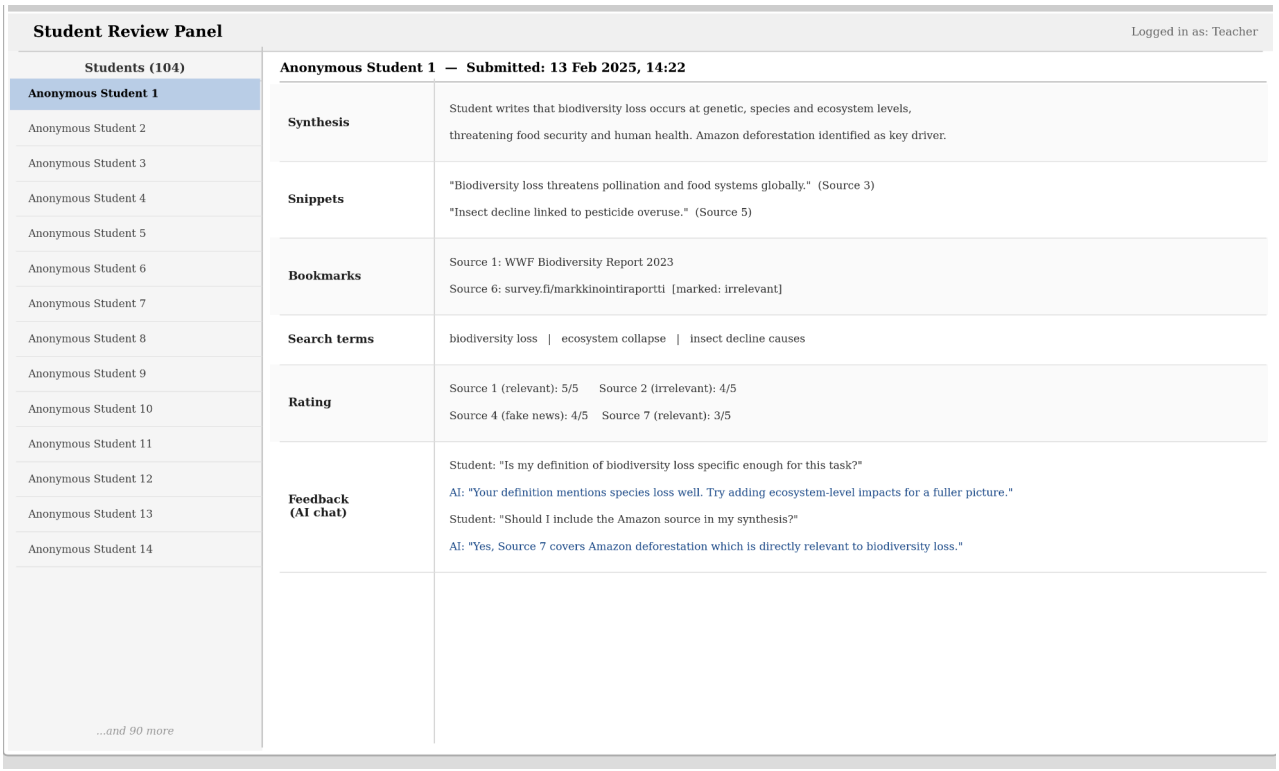

Figure 3.2. Prototype view of the LearnNet teacher view prior to augmentation, showing the student list (left) and the plain-text detail table populated for a selected student (right).

This gap matters because the synthesis text and the process data are not interchangeable sources of information about student understanding. The synthesis shows what a student chose to write. It does not show which sources informed that writing, whether the student engaged with the relevant texts or primarily with the irrelevant and pseudoscientific ones, how many of the predefined main ideas the synthesis actually addresses, or whether the reading strategy was systematic and deliberate or fragmented and superficial. A well-constructed synthesis paragraph could, for example, be based on a single relevant source, leaving three others entirely untouched. Conversely, a thin or disjointed synthesis may reflect a student who consulted multiple relevant sources but struggled to integrate them in writing. Neither situation is visible from the synthesis text alone. [31]

The transparency gap carries a direct practical consequence when the number of students is large. With a cohort of a hundred or more students completing the same multi-source task, a teacher who wants to understand each student's source use and idea coverage in depth would need to open every case individually, read the synthesis, scroll through the source evaluation records, check which snippets were saved from which texts, and form a holistic judgement from these separate pieces. This is not only time-consuming but also cognitively demanding, because the teacher must hold the

structure of the source set and the target ideas in mind while reading each case. In practice, the depth of review individual students receive will vary with available time, and some students may receive only a cursory assessment as a result [13].

At the group level, the problem is compounded. Even after reviewing all individual cases, a teacher synthesising those impressions into a coherent picture of which main ideas were broadly understood and which were consistently missed must perform yet another layer of aggregation without any formal support. Without aggregated information, decisions about follow-up teaching are likely to be driven by the most salient or extreme individual cases rather than by the actual distribution of understanding across the class [13].

The teacher panel described in this thesis is designed to address this transparency gap directly. It does not replace the teacher's judgement; rather, it surfaces the information the teacher needs to make that judgement efficiently and with greater confidence in its accuracy. The specific components of the proposed support are described in the next section.

3.3.2 Process Summaries and Evidence Previews as the Proposed Support

The augmented teacher panel introduces four components to the existing LearnNet interface, each targeting a specific dimension of the transparency gap described above. Together, they are intended to give a teacher both a readable first impression of an individual student's work and the evidence needed to verify or question that impression.

Process summary. The process summary condenses a student's reading timeline, source-by-source time data, and milestone events into a short, readable overview. Rather than presenting raw event logs, the summary translates process traces into a description of the student's approach: how they distributed their time across the process, when they began writing, and whether their interaction pattern reflects systematic or scattered engagement with the material [9]. This addresses the process visibility dimension of the transparency gap by making reading behaviour legible without requiring the teacher to interpret event-level data directly.

Evidence preview. The evidence preview is built around the P-code coverage system described in Section 3.2.2. An LLM analyses the student's synthesis text and saved snippets and returns a judgment of which of the predefined main ideas appear to be present. For each identified idea, the preview links back to the specific passage in the relevant source text from which that idea

originates, giving the teacher a traceable connection between the student's writing and the source material [22]. This addresses the idea-coverage dimension of the transparency gap: instead of reading the synthesis and source texts side by side, the teacher can see at a glance which main ideas were covered and which were not, with the option to inspect the underlying evidence for any specific judgement.

Individual AI report. The individual AI report brings together the process summary, the P-code coverage display, and additional AI-generated observations about the synthesis into a single view of the student's case. This component supports the individual review workflow: a teacher can open one student, read the report, form a first impression, check the evidence preview for specific ideas of interest, and decide on a follow-up action, all within a single panel and without navigating between multiple separate views [12].

Group-level view. The group-level view aggregates P-code coverage across all students in the class and presents the distribution as a set of class-level indicators. A teacher can see which main ideas were addressed by most students, which were consistently absent, and how the class distributed its time and source selections overall. This supports the group review need identified in Section 3.2.3: a teacher planning a follow-up lesson or identifying students who need additional support can use the group view to prioritise without having to mentally aggregate individual impressions. [13]

These four components are evaluated in this thesis through a controlled think-aloud study in which teachers reviewed student cases first without and then with the augmented panel, completing a micro-survey after each case and a short semi-structured interview at the end of the session. The research questions that frame this evaluation are presented in the next section.

3.3.3 Research Questions in the Case-study Context

The four research questions introduced in Chapter 1 are grounded in the specific design context described in this chapter: a multi-source synthesis task on biodiversity loss, completed by 104 students in the LearnNet environment, and reviewed by teachers using an augmented panel with process summaries and LLM evidence previews. Each question addresses a distinct aspect of whether and how well the proposed support serves its intended purpose.

RQ1 – Transparency. What do process summaries and LLM evidence previews help teachers understand quickly about a student's synthesis process? In the LearnNet context, this concerns

whether a teacher who opens a student case in the augmented panel is able to orient faster and more accurately than in the manual condition, and which specific components of the panel most directly contribute to that orientation. This question is motivated by the transparency gap described in Section 3.3.1: if the panel successfully surfaces the connection between synthesis quality and source use, teachers should be able to form a reliable first impression more quickly than through manual review alone [13].

RQ2 – Alignment. To what extent do LLM evidence previews overlap with teacher-identified passages or ideas in the provided sources? This question examines whether the main ideas the AI identifies as present in a student's synthesis are the same ones a teacher would identify independently, and whether the source passages linked by the evidence preview match the passages a teacher would consider most relevant. Alignment is not assumed to be perfect: the AI operates from a predefined coding scheme and a static model, while teachers bring domain knowledge and contextual judgement. Understanding the nature and extent of any gap between AI and teacher assessment is necessary for calibrating how much trust teachers can reasonably place in the preview and where manual verification remains important [22].

RQ3 – Usefulness and efficiency. Does the augmented panel reduce time to first insight and navigational effort compared to manual review, and how do teachers rate its clarity and perceived cognitive load? Efficiency matters for two interconnected reasons. A teacher reviewing a hundred or more student cases cannot afford a review process that is only marginally faster than manual inspection. And if the panel itself is cognitively demanding to interpret, any gains at the information level may be offset by the effort required to use the interface. This question, therefore, addresses speed and usability as joint conditions for practical adoption [15].

RQ4 – Technical feasibility and integrity. Can the LLM evidence service meet the latency and reliability constraints of a real classroom review workflow, and does it avoid the specific failure mode of citing passages or ideas that are not actually present or that misrepresent the student's work? This question concerns the technical performance of the service under realistic conditions, the rate at which teachers notice incorrect AI outputs, and whether the interface gives teachers the means to check the AI's judgments when needed. A system that is fast and usable but factually unreliable would undermine teacher trust and potentially lead to incorrect feedback decisions, making reliability a foundational condition for all other aspects of the panel's value. [22], [25]

Together, these four questions form a progression from the most immediate, perceptual dimension of the support, whether a teacher can quickly understand what a student did, to the most foundational technical condition, whether the system can be trusted to produce accurate outputs consistently. Answers to all four are necessary to assess whether the augmented panel represents a viable and responsible addition to the LearnNet environment.

4 Research Design and Methodology

This chapter describes how the research problem identified in Chapter 3 was investigated. The study combines the design and implementation of a software artifact with an empirical evaluation of that artifact in use. The chapter explains the overall research strategy, the methodological framework that organises the build-and-evaluate cycle, and the role of qualitative and quantitative data in producing the findings reported in Chapter 6. Methodologically, the study operates at two levels: Design Science Research (DSR) [34][35] provides the outer framework governing the build-and-evaluate logic of the work as a whole, while Yin's [33] case study strategy provides the inner evaluation logic through which the deployed artifact is examined within its real-world context.

4.1 Research Approach

4.1.1 Case Study as the Overall Research Strategy

The overall research strategy adopted in this thesis is that of a case study. A case study is an empirical inquiry that investigates a contemporary phenomenon in depth within its real-world context, particularly when the boundaries between the phenomenon and the context are not clearly evident and when the inquiry depends on how and why questions rather than on questions of frequency or distribution [33]. These conditions apply directly to the present work. The phenomenon under investigation, the use of an augmented AI-supported teacher panel to review multi-source synthesis work, is inseparable from the specific educational context in which it occurs: the LearnNet environment, the biodiversity loss task, the structure of the source set, and the review needs of the teachers involved.

The case is defined as the design and deployment of the augmented teacher review panel within a bounded LearnNet instance used for reviewing student synthesis work on biodiversity loss. The boundaries of the case are set by four constraints: a single task instance (the biodiversity loss assignment completed by 104 students in the LearnNet environment), a single teacher-facing interface extension (the augmented panel with process summaries and evidence previews), a controlled evaluation setting (three think-aloud sessions with teacher participants), and a fixed evaluation period during the spring of 2026. This bounded character is central to the case study strategy: it makes it possible to examine the panel and its use in their full contextual complexity rather than abstracting them into variables measured across a broad population [33].

In Yin's typology [33], this constitutes a single-case, embedded design: the case is bounded to a single LearnNet deployment and evaluation period. Within this single case, individual think-aloud sessions serve as embedded units of analysis. Each session involved one participant reviewing a set of student cases under two conditions, manual review and augmented review, and produced multiple types of data: concurrent verbal protocols, micro-survey responses, and interview accounts. Treating each session as an embedded unit makes it possible to examine variation across participants while preserving the integrity of the overall case [33].

A case study does not aim for statistical generalisation in the sense of inferring population-level patterns from a representative sample. Instead, it supports analytical generalisation: the findings can be related to theoretical propositions about how AI-generated summaries and evidence previews might function in teacher review workflows more broadly, while acknowledging that the particular configuration of task, tool, and participants studied here shapes those findings [33]. This is appropriate for a design and evaluation study at the stage of development reported in this thesis, where the primary goal is to understand whether and how the proposed support works in a realistic context, rather than to establish effect sizes that are stable across varied settings.

Case study methodology also accommodates the use of multiple data sources and the triangulation of evidence across them, which is an explicit goal of the present evaluation. The think-aloud transcripts, micro-survey responses, and system log data each illuminate a different facet of the same case, and their convergence or divergence provides a richer and more trustworthy picture than any single source could offer alone. The triangulation of these sources supports construct validity, since findings do not rest on any single data channel alone. Reliability is strengthened by the consistent session protocol applied across all three sessions, which ensured that data collection conditions were comparable across participants. External validity takes the form of analytical generalisation, as the goal is to extend findings to theoretical propositions about AI-supported teacher review rather than to claim representativeness across a broader population [33].

4.1.2 Build-and-evaluate Approach

Within the overall case study strategy, the specific research approach adopted in this thesis is oriented towards design science. Design science research (DSR) generates knowledge through the creation and evaluation of an artifact: the act of building something useful and examining how it performs in its intended use context contributes simultaneously to practical problem-solving and to

theoretical understanding of the problem and its solution [34]. In the present thesis, the artifact is the augmented teacher panel, comprising process summaries, LLM-generated evidence previews, an individual AI report, and a group-level view, implemented as an extension to the existing LearnNet teacher interface.

The methodological structure of the study follows the design science research methodology (DSRM) proposed by Peffers and colleagues [35], which organises design science work into five phases. The first is problem identification and motivation: establishing that a transparency gap exists in the current LearnNet teacher view and that this gap has practical consequences for the quality and efficiency of teacher review. This is developed in Chapter 3. The second phase is the definition of objectives for a solution: specifying what a successful teacher panel would need to achieve in terms of transparency, alignment, usefulness, and technical reliability, operationalised as the four research questions introduced in Chapter 1. The third phase is design and development: implementing the augmented panel as a functional software system, the choices and architecture of which are described in Chapter 5. The fourth is demonstration: deploying the panel in a real task context with actual student data and having teachers interact with it under realistic review conditions. The fifth is evaluation: assessing how well the artifact addresses the defined objectives, using think-aloud sessions, micro-surveys, and system logs as the primary data sources. The present chapter describes the evaluation methodology. Chapter 6 reports the findings. A sixth phase, communication, is represented by the present thesis, which documents the problem, artifact, and evaluation for the research community and thereby completes the full DSRM cycle. [35]

It is important to note that the build-and-evaluate cycle in this thesis is not iterative in the sense of multiple rounds of redesign and re-evaluation. The evaluation reported here represents a single cycle of demonstration and assessment. The findings from this evaluation are intended to inform future iterations of the panel rather than to validate a final, optimised design. This is consistent with what Hevner et al. [34] and Peffers et al. [35] describe as the evaluative logic at early stages of DSR artifact development, where the goal is to establish proof of concept and identify the most important directions for refinement rather than to deliver a product ready for wide deployment. [34], [35]

The build-and-evaluate framing also has implications for what counts as a contribution. In design science research, the contribution is not only the artifact itself but also the knowledge gained about the problem and the design space through building and evaluating it [34]. For this thesis, that means the contribution includes the specific transparency gap identified and addressed, the design choices

made in implementing the panel and the reasoning behind them, the empirical findings about how teacher participants experienced the panel in use, and the design implications those findings produce. Together, these constitute a form of design knowledge that is transferable to other contexts in which AI-assisted review of student text work is being developed. This approach to evaluation corresponds to Hevner et al.'s [34] Guideline 3, which calls for rigorous assessment of a DSR artifact's utility and quality within its intended context of use.

The implementation of the augmented panel is illustrated with anonymised screenshots of the deployed interface at relevant points in Chapter 5, so that the design choices described in text can be connected directly to what participants saw and interacted with during the evaluation sessions.

4.1.3 Role of Qualitative and Quantitative data

The evaluation in this thesis draws on both qualitative and quantitative data, with each type serving a distinct purpose relative to the four research questions. The use of multiple data types is not simply a procedural choice but reflects a substantive methodological rationale: the questions asked in this thesis concern both the measurable outcomes of using the panel (efficiency gains, alignment scores, latency figures) and the situated, interpretive experience of teachers as they work with it (what they notice, what confuses them, what they trust, and why). Neither type of data alone is sufficient to address this combination of concerns, following the multi-source evidence logic of Yin's case study approach [33] and the qualitative evaluation guidance of Seaman [36].

Qualitative data. The primary qualitative sources are the think-aloud transcripts and the semi-structured interview responses collected at the end of each session. These capture teachers' moment-to-moment reasoning during the review process, their reactions to specific interface elements, their assessments of how well the AI's judgements corresponded to their own, and their broader views on the usefulness and trustworthiness of the panel. Qualitative data of this kind is particularly well suited to the early stages of artifact evaluation, where the goal is to understand how and why a system works or fails in context, rather than to establish the magnitude of an effect across a population [36]. Accordingly, the qualitative data in this study are analysed through systematic coding using established qualitative analysis techniques, specifically thematic and content analysis, as described in Section 4.4.2.

Quantitative data. The quantitative sources are the Likert-scale responses from the micro-survey and the technical performance metrics from the LLM service logs. The survey ratings provide

comparable numerical data across participants and cases for constructs such as speed to first insight, perceived effort, clarity of information, and overall usefulness, each corresponding directly to a dimension of RQ3. The technical metrics, specifically LLM response times and token counts recorded during the evaluation sessions, provide objective data for RQ4 on whether the service meets the latency and resource constraints of a realistic classroom review context. These numerical data points are descriptively summarised rather than subjected to inferential statistical analysis, given the small number of participants and the exploratory character of the study.

Triangulation. Across both RQs and data types, the analysis proceeds by triangulation: qualitative observations are connected to survey ratings, and survey ratings are contextualised by qualitative accounts. For example, a participant who gave a high rating for "augmented panel helped me reach first insight faster" but verbally described confusion with the group view interface illustrates a pattern that neither data source captures fully on its own. The mapping of specific data sources to specific research questions is described in Section 4.4.3.

4.2 Data Collection

Data were collected through three main channels: system-generated logs from the LearnNet platform and the LLM service, concurrent think-aloud protocols during teacher review sessions, and micro-surveys and semi-structured interviews administered within those sessions. Each channel was designed to address a different subset of the research questions, and together they provide the triangulated evidence base described in the preceding section. The following subsections describe each channel in turn, including its scope, the conditions under which data were collected, and any limitations relevant to its use in the analysis.

4.2.1 System Logs and Process Data

Two separate logging systems capture interaction data within the LearnNet ecosystem. Student reading sessions are tracked through the platform's own custom database, which records the sequence and timing of source visits, time spent on task, bookmarking and snippet-saving events, source relevance evaluations, and synthesis writing milestones. This student-process data is the input to the process summaries generated by the augmented panel; it is not re-analysed in this thesis but forms the foundation of the artifact being evaluated.

Teacher-session interaction during the evaluation was tracked separately via a Firebase real-time database. This telemetry layer was designed to capture interface-level events, including tab open and close actions, evidence preview expansions, panel section toggles, and timestamps for key review actions, providing a navigational trace of how each teacher moved through the panel during their session.

In addition to interface telemetry, the LLM service records a performance log for each inference call, capturing request timestamp, response time in milliseconds, input token count, and output token count. These LLM service logs are the primary technical data source for RQ4, providing objective evidence on whether the service meets the latency and resource constraints of a realistic classroom review context.

The analysis of teacher interaction sequences and the extent to which telemetry data contributes to the findings is described in Section 4.4.1. Where interaction details require verification beyond what the logs provide, session screen recordings are available to supplement the analysis. LLM performance metrics from the service logs are reported as part of the technical feasibility findings in Chapter 6.

4.2.2 Teacher Think-aloud Sessions

The primary qualitative data in this thesis were generated through three think-aloud sessions conducted with teacher participants between February and March 2026. Think-aloud methodology asks participants to verbalise their thoughts continuously while performing a task, producing a concurrent verbal protocol that reflects moment-to-moment cognitive processing rather than retrospective reconstruction [36]. This makes it particularly appropriate for evaluating a novel interface, since it reveals not only what participants do but also what they notice, how they interpret what they see, and where they encounter confusion or surprise.

Each session was conducted remotely with screen sharing enabled, allowing the researcher to observe the participant's navigation of the teacher panel in real time. Sessions were audio- and screen-recorded with participant consent. The target duration was 75 to 90 minutes, and all three sessions fell within this range.

Session structure. Each session followed the same protocol. After a brief introduction and consent confirmation, the participant was asked to open the micro-survey form alongside the teacher panel

so that both could be used side by side. The session then proceeded in three phases. In the first phase, the manual review condition, the participant reviewed two student cases using the overview tab only, without access to AI features or the process tab. Seven minutes were allocated per case, followed by two minutes to complete the corresponding section of the micro-survey. In the second phase, the augmented review condition, the participant reviewed student cases with full access to the augmented panel, including the process tab, the evidence preview, and the AI synthesis report. The same timing structure was applied. In the third phase, the participant reviewed the group-level view for the entire class and completed the group section of the survey. A short semi-structured interview of approximately ten minutes followed the survey, conducted after the participant had stopped screen sharing.

Think-aloud instructions. Participants were asked to speak aloud continuously throughout the review phases, describing what they were looking at, what they were trying to determine, what was helping or confusing them, and how they were forming their judgment of each student. They were told that there were no correct answers and that the evaluation was of the panel, not of the students' performance. If a participant fell silent for more than a few seconds, the researcher gently prompted them to continue thinking aloud. The researcher did not otherwise intervene or answer questions about the interface during the review phases, in order to preserve the authenticity of the think-aloud protocol. Questions raised during the review were noted and addressed after the session, where appropriate.

Language and transcription. Sessions were conducted bilingually. Participants were free to speak in either Finnish or English, and only one made complete use of Finnish, whereas two did the think-aloud in English. Sessions were transcribed in full. Portions spoken in Finnish have been translated and paraphrased for presentation in Chapter 6, with care taken to preserve the meaning and emphasis of the original utterances. Translated quotations are marked as such in the analysis.

Participants. Three participants took part in the sessions, each identified by a session code (T1, T2, T3). Their backgrounds are described in Section 6.1. Participants were recruited from an educational science context at the University of Turku. All had experience with student teaching tasks and with the subject matter of biodiversity loss to varying degrees, but none had previously used the augmented teacher panel.

4.2.3 Interviews and Retrospective Discussion

Each think-aloud session concluded with a semi-structured interview of approximately ten minutes. The interview was conducted after the participant had stopped sharing their screen, shifting the setting from task performance to reflective discussion. This transition was deliberate: the retrospective format allowed participants to step back from the moment-to-moment experience of using the panel and offer more considered judgments about its overall value, its limitations, and the conditions under which they would or would not use it in a real classroom.

The interview guide comprised eight to ten questions derived directly from the four research questions. The questions covered: what on the screen most quickly communicated what the student had done (RQ1); whether there were any moments in which the evidence preview appeared to mislead or be inaccurate, and how the participant handled those moments (RQ2 and RQ4); how the augmented review compared to the manual condition in terms of steps and effort (RQ3); whether the interface language and framing felt supportive rather than judgmental (RQ1/RQ3); how well the group report matched the participant's own sense of the class (RQ2); any circumstances in which the participant would choose not to use the AI previews (RQ4); and whether they would use the panel in a real lesson setting and what changes would be needed before they did so (RQ3). The full interview guide is included in Appendix A.

Participants were explicitly invited to answer in Finnish or English as preferred. The interviews were recorded as part of the session recording and transcribed in full alongside the think-aloud transcript, so that interview responses could be read in the context of the participant's earlier verbalisations during the review tasks. This continuity between the concurrent and retrospective data is analytically important: it allows interview statements to be connected to specific moments in the review, and it makes it possible to identify cases where a participant's reflective account aligns with or departs from what they said in the moment.

The interview component addresses a limitation of the concurrent think-aloud protocol alone: participants under time pressure during a seven-minute review task may not fully articulate preferences or concerns that they are only partly aware of while working. The retrospective interview provides a space to surface these, making the combined data richer than either source would be independently. [36]

4.2.4 Micro-surveys and Questionnaires

Alongside each review task, participants completed a micro-survey delivered as a Google Form that remained open side by side with the teacher panel throughout the session. The survey was structured to follow the session phases, with separate sections for the manual review cases, the augmented review cases, the block-level comparison, the group view, and an open-feedback field. Completing each case section took approximately two minutes and was integrated into the session timing rather than added at the end, so that responses would reflect the participant's immediate impressions of each case rather than a composite recollection across the whole session.

Manual review section. For each manually reviewed case, participants answered four questions: how quickly they felt they understood what the student had done (five-point Likert scale), which elements of the panel helped most (multi-select from a list of panel components), which main ideas they believed the student had used (multi-select from the twelve P-codes), and what they would do next for that student (multi-select from a set of pedagogical actions). These questions establish a baseline assessment of the manual review experience against which the augmented condition can be compared.

Augmented review section. The augmented case section contained additional items beyond the baseline four. Participants were asked whether the evidence preview helped them judge idea coverage (five-point Likert scale), how much the previewed quotes and ideas overlapped with what they themselves had identified (None / Some / Most / All), how in control they felt while using the preview (five-point scale), whether they noticed any incorrect quotes or mismatched source labels (yes/no with optional description), and whether response time disrupted their flow (five-point scale). They also answered the same main-idea and next-action questions as in the manual section. An optional open comment field was included at the end of each augmented case.

Block comparison section. After completing both the manual and augmented cases, participants answered five Likert items comparing the two conditions directly: whether the augmented panel helped them reach a first insight faster, whether they needed fewer clicks to reach a decision, their rating of the clarity of information in the augmented panel, their rating of the perceived effort involved, and their overall assessment of the augmented panel's usefulness for real classroom review. Two open-text questions followed: which single new element they would keep if forced to choose one, and what they would change or remove. These items map directly to RQ3.

Group view section. The group view section asked participants whether the group report helped them see class-wide strengths and weaknesses quickly (five-point scale), which group-level elements they used (multi-select), which single group element they would retain, how well the group P-code coverage matched their own sense of the class (None / Some / Most / All), what class-level action they would take next based on the group view (multi-select), and whether they encountered any correctness or performance issues in the group view (open text). A final open-feedback field invited any additional comments on the panel as a whole.

Three complete survey response sets were collected, one per participant and session. Each response set covers multiple cases, giving a total of six manual-condition case assessments and nine augmented-condition case assessments across the three participants. The survey responses are analysed both as individual case-level data, to examine within-session patterns, and as aggregated participant-level data, to support the comparison between conditions and the mapping to research questions described in Section 4.4.3.

4.3 Evaluation Design

The evaluation followed a within-subjects design in which the same three teacher participants completed both a manual review condition and an augmented review condition during a single session. This structure allowed direct comparison of review behaviour and perceived usefulness across conditions without the confound of between-participant variation. The ordering of conditions was fixed: each participant began with the manual condition before moving to the augmented condition. This sequencing was chosen to prevent exposure to AI-generated outputs from influencing the independent judgments formed during manual review. Each session was conducted individually with one teacher at a time, lasted approximately seventy to ninety minutes in total, and was divided into three functional phases: a manual review block, an augmented review block, and a retrospective interview. The following subsections describe each phase and the tasks performed within it.

4.3.1 Manual Review and Augmented Review Conditions

In the manual review condition, participants accessed the standard LearnNet teacher view, which displays the full text of each student's synthesis alongside basic submission metadata such as word count and submission timestamp. No AI-generated content, source coverage indicators, or process

summaries were visible during this phase. Participants were asked to read and evaluate each assigned student synthesis as they would in a normal review situation, forming their own judgments about source use, argument quality, and the overall appropriateness of the synthesis. Following each individual case, participants completed the manual-review section of the micro-survey, which captured their evaluation judgment, estimated time, and confidence level. The manual block typically lasted between fifteen and twenty minutes.

In the augmented review condition, participants used the extended teacher panel developed as the artifact of this study. The augmented panel was accessible as an overlay on the same LearnNet environment and presented the following features for each student case: an AI-generated process summary of the student's reading behaviour derived from interaction logs, a P-code coverage indicator showing which main ideas from the relevant source texts appeared in the synthesis, a colour-coded alignment estimate reflecting the degree to which the synthesis engaged with relevant versus irrelevant or pseudoscientific sources, and a brief narrative AI report interpreting the evidence for that case. Participants were free to consult these features in any order before arriving at their review judgment. Following each augmented case, participants completed the augmented-review section of the micro-survey, which included additional questions on perceived alignment between the AI assessment and their own, the helpfulness of specific panel features, and estimated time saving compared with manual review. The augmented block typically lasted between twenty and twenty-five minutes.

A brief transition was provided between the two conditions. The researcher explained the augmented panel features before the second block began, using a short orientation in which the participant could ask clarifying questions about the interface. This orientation was not counted as part of the review time, and the relevant survey questions on time estimation referred only to the review activity itself.

4.3.2 Individual Review Tasks

Each participant was asked to select a set of individual student synthesis cases to review in both conditions. The cases were drawn from the pool of 104 student submissions produced during the biodiversity loss task and selected to provide variation in synthesis quality, source use pattern, and P-code coverage depth. Across the three sessions, participants collectively reviewed six cases in the

manual condition and nine cases in the augmented condition, yielding fifteen case-level evaluation records in total.

The individual review task in both conditions followed the same procedural structure. The participant opened a student case, read the synthesis text, and formed a holistic evaluation judgment before completing the per-case micro-survey. In the augmented condition, the participant was also expected to consult the AI panel features before finalising their judgment. The think-aloud protocol was active throughout both blocks, with participants asked to verbalise their reasoning continuously as they read and evaluated each synthesis.

The number of cases per condition was calibrated to keep each block within a manageable time window while still generating sufficient data per research question. Given the exploratory and qualitative emphasis of the study, the priority was on depth of verbalisation and survey response quality rather than on maximising the number of cases reviewed.

4.3.3 Group-level Review Tasks

Following the individual case reviews in the augmented condition, each participant was shown the group-level view of the augmented panel. The group view aggregates AI-generated assessments across all student submissions for the task and presents a class-wide summary of source engagement patterns, P-code coverage distribution, and the proportion of cases flagged by the AI as primarily engaging with relevant, irrelevant, or pseudoscientific sources. The view was designed to support the kind of formative feedback that a teacher might direct at the class as a whole rather than at individual students.

Participants were invited to explore the group view freely and to verbalise their impressions of the information presented. The think-aloud protocol remained active during this phase as well. After exploring the group view, participants completed the group-view section of the micro-survey, which asked five Likert-scale questions covering the clarity of the class-level summary, the perceived usefulness of the group view for planning instructional responses, and whether the pattern of AI classifications across the class seemed consistent with the participant's prior experience of the student cohort. Two open-text fields in the same survey section invited participants to describe what additional information would improve the group view and to note any aspects they found misleading or incomplete.

The group-level review task was introduced only in the augmented condition because the standard LearnNet interface does not provide an equivalent aggregated view. No corresponding group task was therefore conducted during the manual block. This asymmetry means that responses to group-view questions reflect reactions to the AI-supported interface exclusively and cannot be compared with a baseline group review under manual conditions. This limitation is acknowledged in the discussion of findings in Section 6.

4.3.4 Alignment and Integrity Checks

The evaluation design included two categories of alignment check, each addressing a distinct research question. The first category concerned the alignment between AI-generated assessments and the independent judgments formed by teachers during manual review (RQ2). To operationalise this check, the micro-survey included a direct comparison question in the augmented review block asking whether the AI assessment of the current case agreed with, partially agreed with, or disagreed with the judgment the participant had formed during their earlier manual review. Participants were also asked to note, in an open-text field, any specific aspect of the AI output that contradicted their own reading of the synthesis. These responses provide a case-level record of perceived alignment from the teacher's perspective and are analysed alongside the qualitative coding of concurrent verbalisations in Section 6.3.

The second category concerned technical integrity, defined as the consistency and completeness of AI-generated outputs across cases (RQ4). Technical integrity was assessed using the LLM service logs through Langfuse, which record the input token count, output token count, and response time in milliseconds for each generation request. Consistent output lengths relative to input size indicate that the prompt structure functioned as designed across the range of student synthesis texts encountered during the evaluation. Unusually short outputs may indicate truncation or generation failures; abnormally long response times may indicate service latency that would affect usability in deployment. These metrics are reported descriptively in Section 6.4.

An additional integrity consideration concerns the traceability of AI claims to the underlying evidence. Each AI-generated report produced by the augmented panel was designed to reference specific P-codes as the basis for its assessment of source engagement. Where participants commented during think-aloud or interview that an AI claim appeared unsupported or difficult to verify, these observations were coded as instances of transparency concern and contributed to the

findings reported under RQ1. The triangulation of survey-based alignment scores, qualitative verbalisation data, and log-based technical metrics thus provides a multi-source basis for evaluating both the functional and perceived integrity of the artifact.

4.4 Data Analysis

The study generates data from four distinct sources: interaction logs and LLM service records, concurrent think-aloud verbalisations, retrospective semi-structured interviews, and micro-survey responses. Because these sources differ in format, granularity, and evidential function, separate analytical procedures were applied to each before the findings were triangulated. The overall analytical approach is consistent with a qualitative case study in which quantitative descriptive data from logs and surveys serve to contextualise and corroborate interpretations derived from qualitative verbal data, following the multi-method logic described by Yin [33] and the qualitative analysis guidelines of Seaman [36]. The following subsections describe the specific procedures applied to each data type.

4.4.1 Analysis of Telemetry and Dashboard Usage

Two categories of log data were collected during the study. The first category comprises teacher-session telemetry captured through a Firebase real-time database, recording tab-switching events and associated timestamps as each participant navigated between panel sections during the augmented review block. The second category comprises LLM service logs generated each time the augmented panel requested an AI-generated output, recording the input token count, output token count, and server response time in milliseconds for each call.

Telemetry data from teacher sessions were analysed descriptively to characterise navigation patterns across the augmented panel. Where the log records were sufficiently complete, the sequence and duration of tab visits per case were used to infer which panel features attracted the most attention and in what order participants typically consulted them. Where interaction details require verification beyond what the Firebase logs provide, session screen recordings are available to supplement the analysis, as noted in Section 4.2.1.

LLM service log data were analysed to address RQ4 regarding the technical feasibility and integrity of AI-generated outputs. The analysis focused on three descriptive metrics: the distribution of response times across all generation calls made during the three sessions, the ratio of output tokens

to input tokens as an indicator of generation consistency, and whether any calls produced anomalously short outputs suggestive of truncation or generation failure. These metrics are reported as descriptive summaries in Section 6.5 and are interpreted in relation to the usability threshold discussed in Section 5. No inferential statistical testing was applied to the log data, as the sample of generation calls, while larger than the participant sample, reflects a single task context and cannot support generalisable performance claims.

4.4.2 Analysis of Teacher Interviews and Think-aloud Data

The think-aloud transcripts and interview transcripts were analysed using a combined deductive and inductive thematic approach, following Seaman's [36] guidance on qualitative analysis of verbal data in software and systems evaluation studies. NVivo software was used to support systematic coding across the three sessions.

In the first phase, a deductive coding scheme was derived directly from the four research questions. Each research question was operationalised as a primary theme: transparency of AI-generated evidence (RQ1), alignment between AI assessments and teacher judgments (RQ2), perceived usefulness and efficiency of the panel (RQ3), and technical performance and reliability (RQ4). All utterances in the think-aloud and interview transcripts were read and assigned to one or more of these primary themes where applicable. Utterances that could not be assigned to any primary theme were retained for the second phase of analysis.

In the second phase, an inductive pass was conducted over the coded data to identify subthemes that emerged from participants' own language and concerns, regardless of whether they mapped directly to a research question. This process yielded additional interpretive categories, including, for example, references to cognitive load during review, comparisons between the augmented panel and prior review practices, and expressions of uncertainty about the basis of AI claims. In total, ten themes were identified and frequency-counted across all three transcripts, with each participant's contribution recorded separately to preserve session-level variation. The resulting theme-by-participant frequency matrix is presented in Section 6.6 alongside the chart produced from the frequency data. Together, the deductive and inductive phases constitute an explanation-building approach in the sense described by Yin [33]: findings are progressively refined across the three embedded cases until a coherent account of teacher experience with the panel is reached.

Think-aloud data and interview data were coded and analysed as separate layers before being interpreted together. This distinction was maintained because concurrent verbalisations reflect real-time reasoning during task performance, while retrospective interview responses reflect participants' reflective accounts after the fact. Where the two layers converged on the same evaluation judgment, this was treated as stronger evidence than either source alone. Where they diverged, the discrepancy was noted and discussed in the relevant findings section, as divergence may itself be analytically informative about the role of reflection in shaping perceived usability.

Given that one of the three participants conducted parts of their session in Finnish, Finnish-language utterances were interpreted with reference to context, task materials, and the bilingual competence of the researcher. Passages whose meaning was uncertain in the original Finnish were flagged during coding and treated conservatively, contributing to frequency counts only where their thematic assignment was clear.

4.4.3 Mapping Data Sources to Research Questions

Each of the four research questions is informed by a specific subset of the data collected during the study. The mapping below makes explicit which data sources contribute evidence to each question and how those sources relate to one another analytically.

RQ1 asks whether the augmented panel presents AI-generated assessments in a way that teachers find transparent and interpretable. The primary evidence for RQ1 comes from think-aloud verbalisations in which participants commented on the clarity, comprehensibility, or traceability of AI outputs while using the panel, and from interview responses to questions explicitly addressing transparency. Micro-survey items asking participants to rate the clarity of AI explanations and the legibility of P-code coverage indicators provide supplementary quantitative evidence. Together, these sources allow the transparency finding to be grounded in both spontaneous in-task reactions and considered post-task reflections.

RQ2 asks whether AI-generated assessments align with the evaluation judgments that teachers form independently through manual review. The primary evidence for RQ2 comes from the per-case alignment question in the micro-survey, which asked participants to characterise the degree of agreement between the AI assessment and their own prior judgment. Open-text responses in the same section, in which participants described specific points of disagreement, provide qualitative depth. Think-aloud verbalisations recorded during the augmented review block, particularly those in

which participants compared the AI output with their own reading, contribute additional evidence. Across these sources, RQ2 findings are triangulated at the case level.

RQ3 asks whether the augmented panel improves the usefulness and efficiency of the review process from the teacher's perspective. The primary evidence for RQ3 comes from the block comparison section of the micro-survey, which asked participants to compare the two conditions across five Likert-scale dimensions, including perceived time saving, confidence in judgment, and likelihood of use in practice. Interview responses to questions about overall helpfulness and workflow integration provide qualitative context. Think-aloud references to task difficulty, time pressure, or the effort required to interpret AI outputs are coded under the usefulness theme and contribute to this question as well.

RQ4 asks whether the AI components of the augmented panel operate with sufficient technical reliability and integrity for use in an authentic teacher review context. The primary evidence for RQ4 comes from the LLM service logs, which provide objective metrics of generation performance independent of participant perception. Secondary evidence comes from think-aloud and interview instances in which participants encountered or commented on apparent errors, inconsistencies, or unexpected outputs in the AI-generated content. Micro-survey items asking whether participants detected any factual errors in the AI report for a given case provide a third source. The convergence of log-based performance data with qualitative accounts of reliability concerns allows RQ4 to be addressed at both the technical and experiential levels.

4.5 Ethical Considerations

The study involved human participants in their professional capacity as teacher educators and engaged with student-produced data collected under a prior course activity. Both dimensions of the study required careful attention to consent, data protection, and the boundary between the current evaluation study and the broader Read2Learn and FINSCI research projects from which the task context and student data were drawn. The subsections below describe the specific arrangements made for consent and anonymisation, data handling and storage, and the ethical limitations that bear on the evaluation design.

4.5.1 Consent and Anonymisation

All three teacher participants were recruited voluntarily and provided informed consent before taking part in the study. Participants were informed of the purpose of the research, the data that would be collected during their session (screen recording, think-aloud audio, survey responses, and interaction logs), how the data would be stored and used, and their right to withdraw at any point without consequence. Consent was documented in writing prior to each session.

Participants are identified in all data records and in the thesis using the pseudonyms T1, T2, and T3, assigned in session order. No identifying information linking these codes to individual names or institutional affiliations is retained in any file associated with this study. Think-aloud and interview transcripts were prepared with names and any incidentally disclosed personal identifiers removed before analysis.

The student synthesis texts processed by the augmented panel during the study were produced as part of a course activity conducted under the Biodiversity Loss task, which is itself part of the Read2Learn research project at the University of Turku. Student participants in that course activity provided consent for their data to be used in research associated with the project. The current evaluation study operates within the scope of that consent. Student authors are not research subjects of the present study and are not discussed individually in the findings; their syntheses appear only as the input material processed by the augmented panel and reviewed by teacher participants.

4.5.2 Data Handling and Storage

All data collected during the study are stored on the University of Turku infrastructure in accordance with the data management policies of the university and the requirements of the General Data Protection Regulation (GDPR) [37]. Session screen recordings and think-aloud audio files are stored in access-controlled project storage accessible only to the researcher and supervisor. Micro-survey responses were collected through a Google account, which operates under the data processing agreement with Google, and were exported and stored locally in the project folder before analysis. Firebase telemetry data were generated and stored within the LearnNet project infrastructure managed by the research group.

One data handling consideration specific to this study concerns the transmission of student synthesis texts to an external LLM service for the generation of AI reports and process summaries.

Before each API call, synthesis texts were stripped of any remaining student identifiers so that the content sent to the external service contained no personal data attributable to an individual student. This arrangement was reviewed as part of the broader data management planning for the Read2Learn project.

Data collected during this study will be retained for the period required by the University of Turku research data policy and will not be shared with parties outside the project without participant consent. Anonymised excerpts from think-aloud and interview transcripts are included in the thesis solely in the form of illustrative quotations and thematic summaries.

4.6 Limitations of the Evaluation Design

Several design decisions and practical constraints introduce limitations that affect the scope and generalisability of the findings. The most significant is the small participant sample [n=3]. With three teacher participants, the study cannot support statistically generalisable claims about the usability or effectiveness of the augmented panel across the broader population of teachers or teacher educators. The findings are best understood as forming an interpretive basis for further evaluation and as informing design iterations rather than as definitive evidence of efficacy, consistent with the evaluative logic of Design Science Research. [34], [35]

The within-subjects design, while enabling direct comparison of conditions, introduces a potential ordering effect. Because all participants completed the manual condition before the augmented condition, any improvement in review efficiency or judgment confidence observed in the augmented block might partly reflect familiarity with the task and the student cases rather than the effect of the AI features alone. Counterbalancing across participants was not possible given the constraint that exposure to AI-generated outputs must not precede the formation of independent manual judgments; this constraint was judged to outweigh the ordering-effect risk.

Think-aloud protocol introduces reactivity: the act of verbalising reasoning may slow participants down, cause them to reflect more explicitly than they would in normal practice, or produce self-presentational effects in which participants frame their evaluations for the researcher's benefit. These effects cannot be eliminated entirely, though the concurrent and unrehearsed nature of the think-aloud reduces the scope for post-hoc rationalisation compared with purely retrospective methods.

The evaluation was conducted in a single task context, the biodiversity loss synthesis task, and with a specific cohort of university students. The augmented panel's performance, particularly the quality and relevance of AI-generated content, may differ with other task topics, different source configurations, or different student populations. Findings relating to P-code coverage and source alignment should therefore be interpreted with reference to the specific source structure of the task rather than as claims about the system's behaviour in general.

Finally, the group-level review task was conducted only in the augmented condition because no equivalent class-wide view exists in the manual LearnNet interface. This means that findings regarding the group view cannot be compared against a manual baseline, limiting the conclusions that can be drawn about the added value of that specific feature relative to what teachers could achieve through unaided review of individual cases.

5 System Design and Implementation

This chapter describes the design and implementation of the augmented teacher panel, the artifact produced through the Design Science Research process documented in Chapter 4. The panel extends the existing LearnNet teacher view with AI-supported features for reviewing student multi-source synthesis work. The chapter is organised around four concerns: the design goals and overall system architecture (Section 5.1), the process summaries, evidence previews, and report generation features (Section 5.2), the guardrail, verification, and logging mechanisms that support reliability (Section 5.3), and the key implementation decisions and challenges encountered during development (Section 5.4).

5.1 Design Goals and Overall Architecture

The augmented teacher panel was developed as an extension to the existing LearnNet platform, designed to surface AI-supported insights alongside the deterministic learning artifacts already available to teachers. Three interrelated design goals shaped the system from the outset: providing teachers with rapid, actionable insight into student synthesis quality; making AI-generated assessments traceable to observable evidence in student work; and preserving teacher control over interpretation and instructional decision-making. These goals collectively define the utility dimensions against which the artifact is evaluated in Chapter 6, following Hevner et al.'s [34] Guideline 3. The architecture that realises these goals is a staged, multi-layer pipeline that transforms raw student learning traces into a unified teacher-facing view without exposing internal model reasoning directly. Each design goal is discussed in the following subsections before the full architecture is presented.

5.1.1 Support for Quick Teacher Insight

A central design concern was reducing the time and cognitive effort required for a teacher to form an informed initial assessment of a student synthesis case. In the standard LearnNet teacher view, a thorough review of a single student case requires reading the full synthesis text and cross-referencing source engagement from raw logs, a process that is comprehensive but time-intensive when applied across a large cohort. The augmented panel aimed to reduce this time-to-first-insight without replacing the teacher's own judgment.

No strict time threshold was formalised as a system requirement during design. The practical reference point came from the evaluation protocol: participants in the think-aloud study were asked to complete each individual case review within approximately seven minutes, with a further two minutes allocated for the per-case micro-survey. This framing guided decisions about information density and the ordering of panel elements, since the most interpretable outputs needed to be immediately visible on opening a student case, with supporting detail accessible on demand rather than presented all at once.

Two architectural choices directly support rapid insight. First, deterministic summary statistics – source engagement counts, snippet activity, timing data – are computed by the LearnNet backend independently of any LLM call and are delivered as the first data stream to the teacher panel. Baseline indicators, therefore, appear without LLM latency, allowing teachers to begin forming a picture of the case while AI-generated content continues to load. Second, LLM-generated outputs for individual student cases are cached after the first generation and returned from cache on subsequent views. Teachers reviewing a case for a second time or reopening the panel encounter no additional wait time. The combination of a fast deterministic stream and a cached AI stream was designed to make the panel feel responsive under realistic classroom review conditions.

5.1.2 Traceability and teacher control

The design goals for traceability and teacher control respond to a well-documented tension in AI-supported educational systems: that opacity in AI-generated outputs limits teacher trust and constrains meaningful use of those outputs [22]. The augmented panel addresses this tension through two related design decisions.

Traceability is implemented primarily through source-linked evidence previews and idea-level coverage indicators rather than through exposure of the model's internal reasoning. For each student case, the panel presents an AI-generated narrative summary alongside indicators showing which predefined main ideas (P-codes) from the relevant source texts appear to be covered in the student's synthesis and supporting snippets. Where the AI report makes a specific claim about source engagement, it is anchored to an evidence preview drawn from the student's own text. This design gives teachers a path from the AI's assessment back to observable student work without requiring them to interpret raw model outputs or probability scores. The goal was interpretability for the

teacher rather than transparency of the model's internal mechanism, a distinction consistent with the explainable AI principles discussed in the literature [22].

Teacher control in the current implementation is preserved through the read-only nature of the AI outputs. The panel presents AI assessments as review support material: teachers can inspect, accept, discount, or mentally override any AI output, but the panel does not function as a workflow in which teachers are required to formally confirm or correct AI labels. This design choice reflects the formative and exploratory character of the evaluation study, in which the priority was to observe how teachers engaged with AI support rather than to embed AI outputs into an official assessment record. The implications of this design boundary - particularly whether future versions should include a structured teacher feedback mechanism - are discussed in Section 7.

5.1.3 Overall Architecture and Data Flow

The augmented panel is implemented as a React/TypeScript frontend extension to the LearnNet teacher view, communicating with two backend services: the existing LearnNet Node.js backend, which handles deterministic aggregation of student activity data, and a dedicated Python-based LLM service referred to as FeedbackAPI, which manages AI input assembly, model inference, and post-processing. The full system operates as a seven-layer pipeline, illustrated in Figure 5.1, that transforms raw student learning traces into the integrated teacher-facing view.

Data flows through the system via two parallel streams that converge in the teacher panel frontend. Stream A is initiated when the teacher panel loads and requests cohort analytics from the LearnNet backend. The groupSummary route computes participation metrics, search and bookmark counts, snippet activity, synthesis timings, and process highlights directly from SQL queries against the LearnNet relational database. This stream produces the baseline analytical view with no dependency on the LLM service and is therefore available to the teacher with minimal latency.

Stream B is initiated by requests to the FeedbackAPI endpoints and is responsible for all AI-generated content. Before any LLM call is made, the FeedbackAPI assembles a structured input from the database: it fetches instance and exercise context, retrieves the reference P-codes and their descriptions, collects the relevant student synthesis and snippet texts, and constructs a normalised metadata object. P-code coverage indicators are computed at this stage through one of two paths: if a sufficient number of cached individual evaluations exist, coverage is aggregated from the cache; if the cache falls below the configured threshold, coverage is computed directly from student content

using a fallback algorithm. Only after this feature engineering step does the system construct a prompt and submit it to the GPT-5 API.

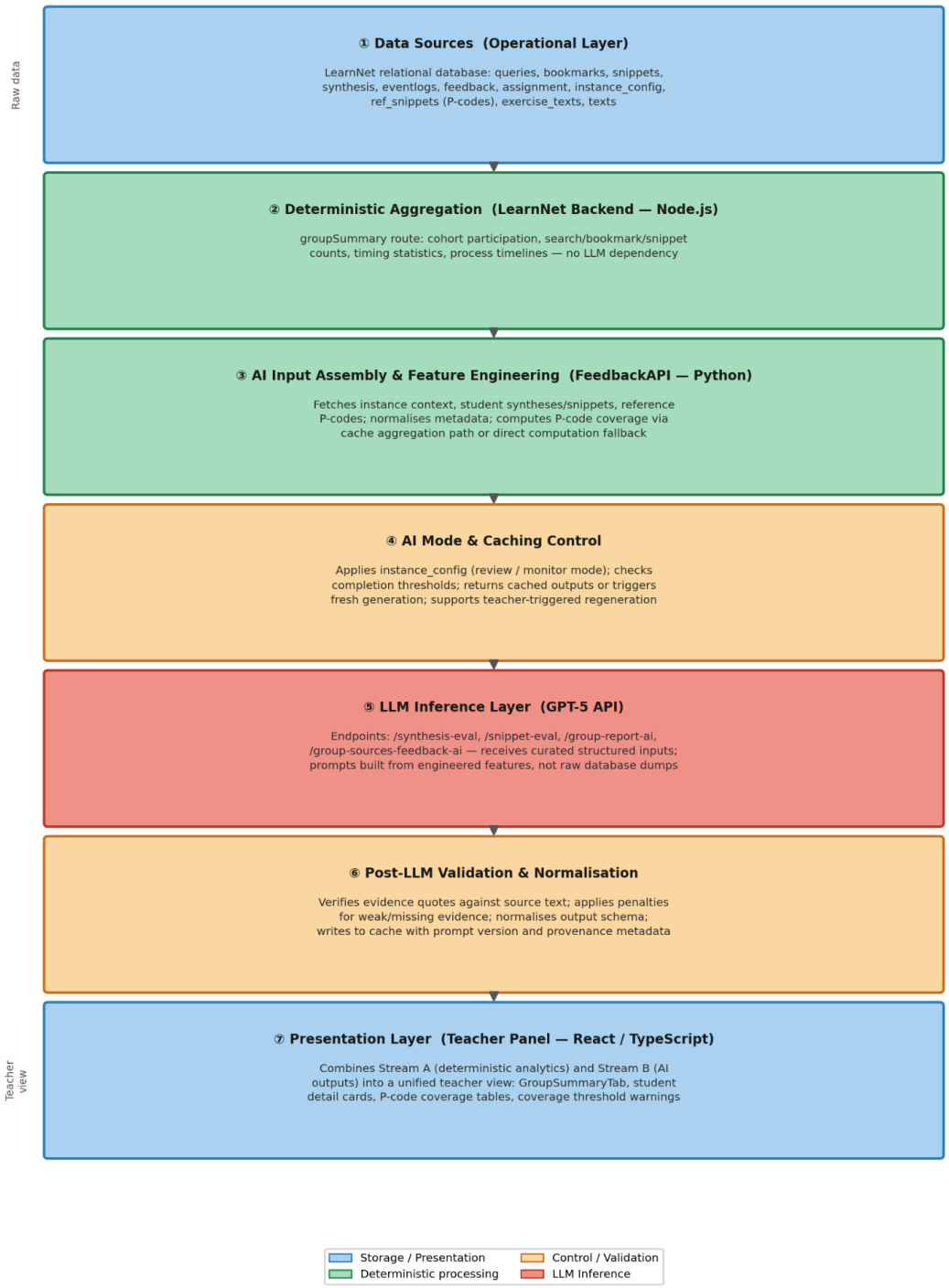

Figure 5.1. End-to-end data pipeline of the augmented teacher panel, from raw student learning traces in the LearnNet database through deterministic aggregation, AI input assembly, caching control, LLM inference, post-processing validation, and final rendering in the teacher panel frontend.

Stream B is initiated by requests to the FeedbackAPI endpoints and is responsible for all AI-generated content. Before any LLM call is made, the FeedbackAPI assembles a structured input from the database: it fetches instance and exercise context, retrieves the reference P-codes and their descriptions, collects the relevant student synthesis and snippet texts, and constructs a normalised metadata object. P-code coverage indicators are computed at this stage through one of two paths: if a sufficient number of cached individual evaluations exist, coverage is aggregated from the cache; if the cache falls below the configured threshold, coverage is computed directly from student content using a fallback algorithm. Only after this feature engineering step does the system construct a prompt and submit it to the GPT-5 API.

After the model response is received, a post-processing stage verifies evidence quotes against the original source texts, applies penalties where evidence is weak or unverified, normalises the output to a defined schema, and writes the result to cache with associated provenance metadata, including the prompt version used. The teacher panel frontend, illustrated in Figure 5.2, then combines the Stream A deterministic output and the Stream B processed AI output into a single integrated view. Coverage warning indicators are displayed when cached data falls below the minimum threshold. From the teacher's perspective, the panel presents one coherent view; the multi-stage provenance of each element is not surfaced to the interface.

The separation of the system into two streams, one deterministic and one AI-driven, was a deliberate architectural decision with both technical and pedagogical motivations. Technically, it insulates the baseline analytics from LLM latency and from any variability in model output. Pedagogically, it ensures that teachers always have access to the factual, verifiable activity summary regardless of whether an AI-generated report is available for a given case. Consistent with the emphasis in teacher-facing dashboard research on keeping human-interpretable data visible alongside any generated content [15], [16], AI insights in the panel are designed to augment rather than replace the deterministic summaries.

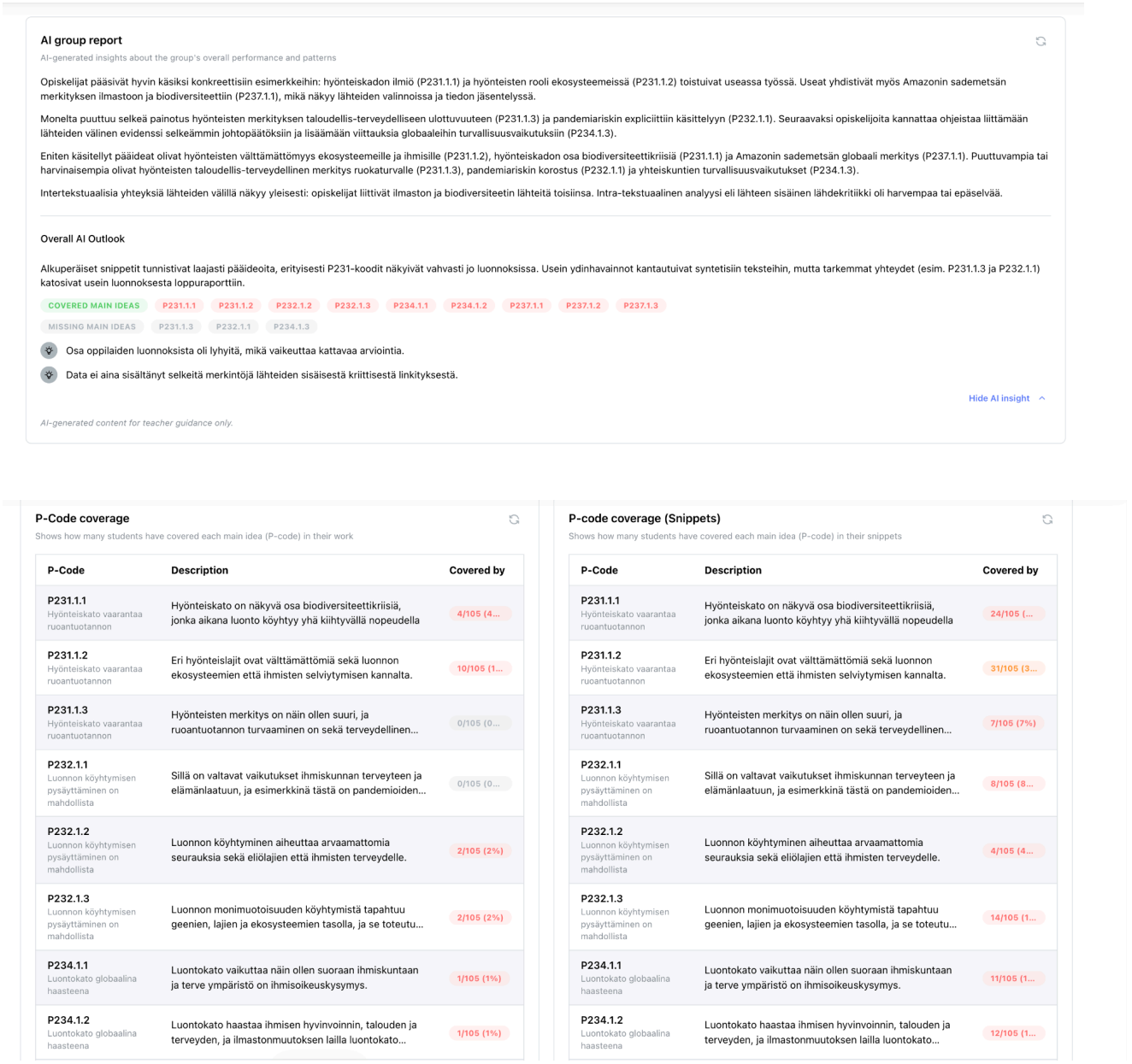

Figure 5.2. Screenshot of the augmented teacher panel showing the integrated view with GroupSummaryTab, P-code coverage table, and AI report card.

5.2 Process Summaries and Evidence Previews

The augmented teacher panel presents information at two levels of granularity: individual student cases and the class as a whole. Within each level, the interface combines deterministic process summaries derived directly from student activity logs with AI-generated evidence previews and narrative reports. This section describes how process summaries and derived indicators are constructed (Section 5.2.1), how LLM-based evidence previews are generated and verified (Section

5.2.2), how individual and group reports are produced (Section 5.2.3), and how all of these components are integrated into the teacher panel interface (Section 5.2.4).

5.2.1 Process Summaries and Derived Indicators

Process summaries are produced entirely by the deterministic aggregation layer of the pipeline and require no LLM involvement. They are derived from the raw student activity tables in the LearnNet database, which includes searches, bookmarks, snippets, synthesis edits, and event logs, and are presented to the teacher as a compact behavioural profile of each student's engagement with the task. The design of these indicators draws on the principle that meaningful process feedback must be grounded in observable learning traces rather than in model-generated inferences [9], [10].

At the individual level, the process summary for each student is organised around four at-a-glance statistics: total time on task, number of search queries issued, number of source texts bookmarked, and number of snippets captured from sources. These four values give the teacher an immediate sense of the student's level of engagement before any detailed reading is required. Below the summary statistics, a milestones bar presents the chronological sequence of key process events as a colour-coded timeline: the first search, first bookmark, first snippet, synthesis start, and last synthesis edit are each marked with their timestamp. From these events, three derived time indicators are computed: task understanding time, reading time, and writing or synthesis time, alongside the total elapsed time. Together, the at-a-glance statistics and milestone timeline allow a teacher to assess at a glance whether a student spent meaningful time engaging with sources before beginning to write. (Figure 5.3).

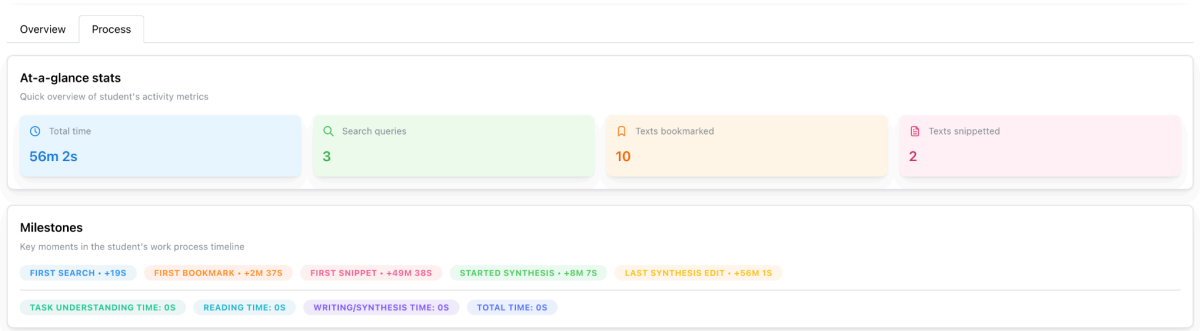

Figure 5.3. At-a-glance statistics and Milestones for a single student

A filterable chronological timeline (Figure 5.4) beneath the milestones bar presents the full sequence of recorded student actions across six categories: all events, searches, bookmarks,

snippets, synthesis, and feedback interactions. This view enables teachers to inspect the detailed temporal structure of a student's work process when the summary indicators suggest patterns worth investigating further.

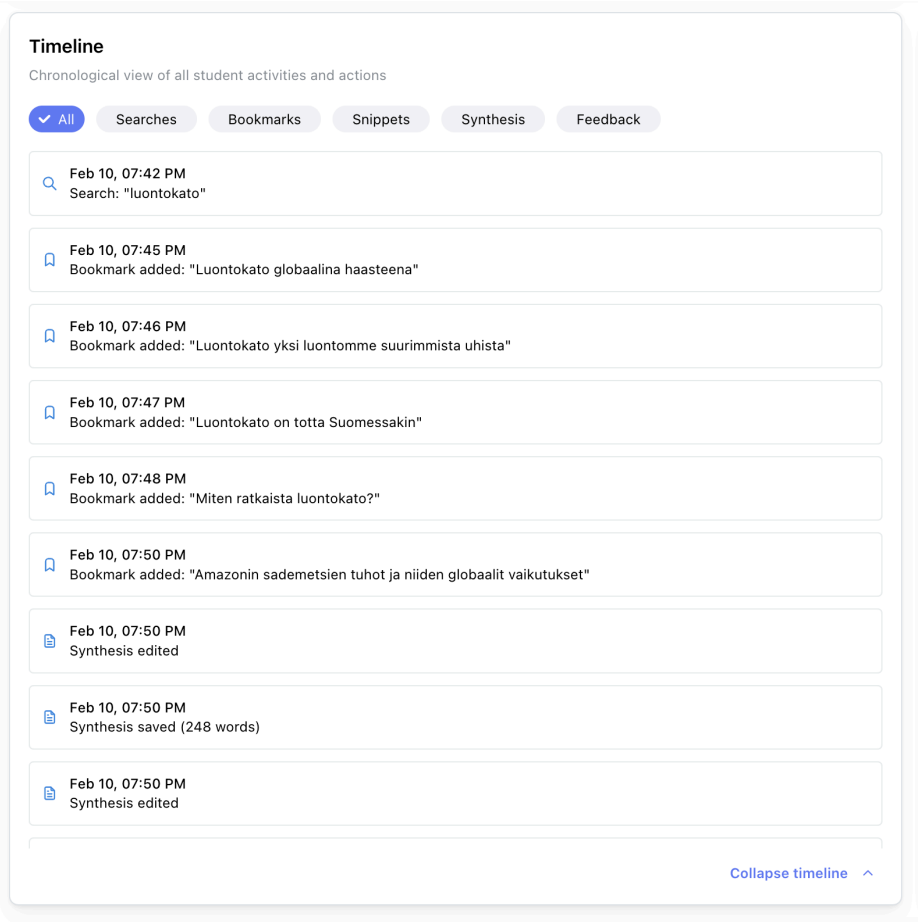

Figure 5.4. Filterable chronological timeline of the milestones recorded from the student while they worked on the task

At the group level, two complementary indicator views are provided (Figure 5.5). The first is a bookmark evaluation coverage display in which each student's source evaluations are colour-coded: green cells represent how students evaluated the relevant source texts - whether they correctly or incorrectly marked them as relevant or left them unanswered - and red cells represent how students evaluated the irrelevant source texts. This visualisation allows teachers to identify at the class level which sources were most consistently misclassified and whether patterns of source confusion cluster around particular students.

The second group-level indicator is a pair of P-code coverage tables, one for synthesis coverage and one for snippet coverage, each listing all predefined main ideas with the count and percentage of

students whose work was assessed as covering that idea. Low-coverage P-codes, displayed in orange or red, signal main ideas that the majority of students failed to address, providing a direct basis for class-level instructional decisions.

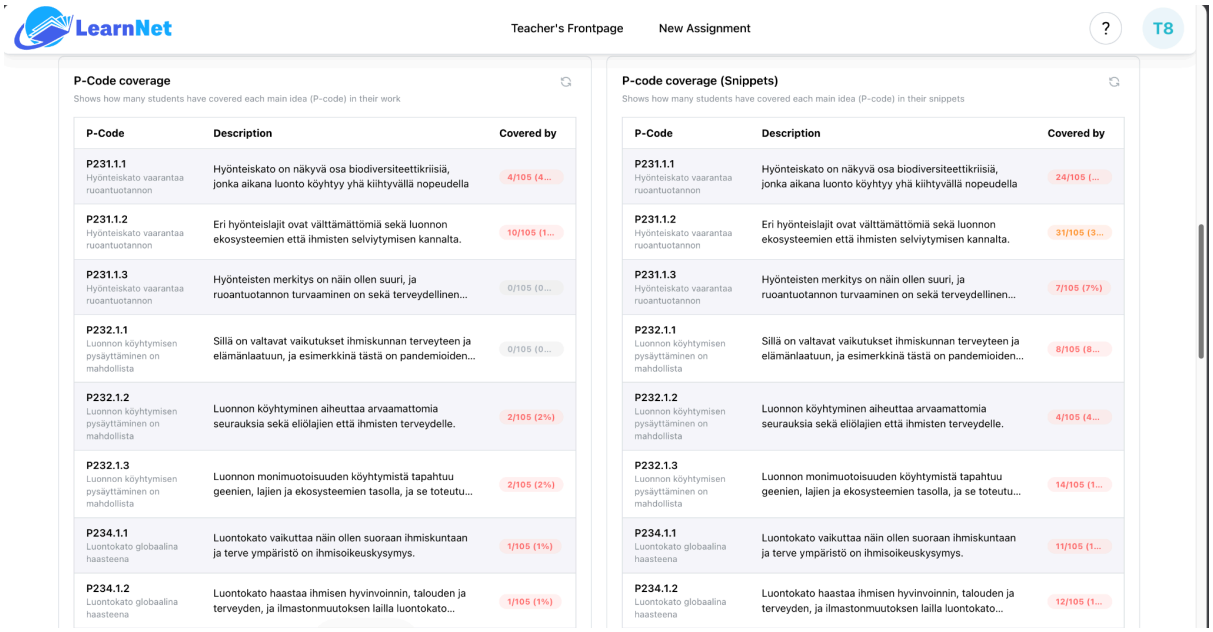

Figure 5.5. Group-level P-code coverage tables showing the number and percentage of students who addressed each predefined main idea in their synthesis text (left) and in their captured snippets (right). Low-coverage ideas are highlighted in orange.

5.2.2 LLM-based Evidence Previews

Evidence previews are the primary mechanism through which the augmented panel makes AI-generated assessments traceable to observable student work. Rather than presenting a single summary judgment, the panel exposes the reasoning behind each P-code score through a structured breakdown that combines an LLM-generated rationale with a quoted evidence excerpt drawn from the student's own text. This design reflects the explainable AI principle that educational AI tools should allow users to understand the basis of a system's output rather than requiring them to accept it on trust [22].

Evidence previews are generated by two FeedbackAPI endpoints operating in parallel: one evaluating the student's synthesis text and one evaluating the student's captured snippets. Both endpoints receive the same structured input: the selected P-codes and their descriptions, the relevant student text, and the supporting source content. The LLM is prompted to assess the degree to which

each P-code is covered and to identify a specific passage from the student's text that supports its judgment. For each P-code, the system produces a score, a brief rationale explaining the score, and an evidence quote attributed to either the synthesis or the snippets as appropriate.

After the LLM response is received, a post-processing step verifies each evidence quote by checking whether the quoted passage appears verbatim or near-verbatim in the source text. Where a quote cannot be verified against the student's actual text, a penalty is applied to the score, and the evidence item is flagged accordingly. This verification step guards against a known failure mode of language models in which the model generates plausible-sounding but fabricated quotations. The resulting output for each P-code is displayed in the teacher panel as an expandable card showing the score badge, the rationale paragraph, and the evidence excerpt tagged with its source type - Synthesis or Snippets - and a Verified or Not Verified indicator in green or red.

Above the per-P-code breakdown, the panel displays an overall feedback narrative summarising the student's synthesis at a holistic level, together with two sets of P-code chips: covered main ideas in green and missing main ideas in a muted colour. This two-level presentation - a high-level overview followed by drill-down detail - is designed to support teachers who need a quick overall impression as well as those who want to examine the evidence for individual ideas. All of this is shown in Figure 5.6.

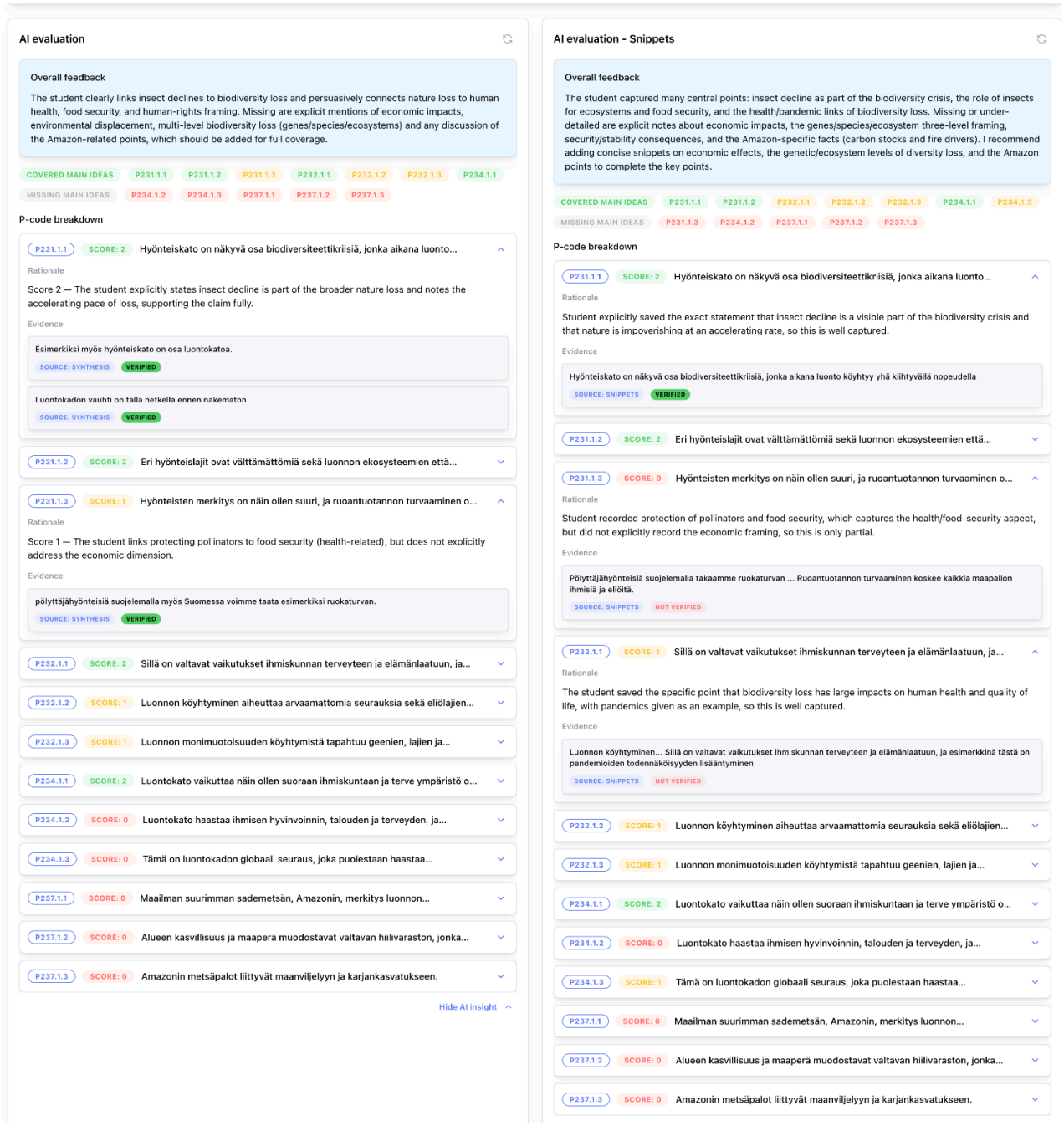

Figure 5.6. AI evaluation panel for an individual student case, showing the overall feedback narrative, covered and missing main idea indicators (P-code chips), and an expanded P-code breakdown card with LLM-generated rationale and verified evidence excerpt. The parallel panel on the right shows the equivalent evaluation for the student's captured snippets.

5.2.3 Individual and Group Report Generation

The augmented panel generates AI reports at two levels: individual student cases and the class as a whole - through separate FeedbackAPI endpoints with distinct input assembly procedures and output structures.

Individual reports are produced by the synthesis evaluation and snippet evaluation endpoints described in Section 5.2.2. Each individual report is specific to one student case and consists of the overall feedback narrative, the covered and missing P-code indicators, and the full per-P-code breakdown with evidence previews. Individual reports are generated on demand when a teacher opens a student case in the augmented panel and are cached after the first generation. In **review** mode, the cached report is returned directly on subsequent views, ensuring consistent output and avoiding redundant API calls. In **monitor** mode, fresh generation is permitted to reflect updated student data. A teacher-triggered regeneration endpoint allows the cache to be manually refreshed if required.

Group reports are produced by two dedicated endpoints: the group report endpoint, which generates a cohort-level narrative, and the group sources feedback endpoint, which analyses patterns of source engagement across all students. Before the LLM call is made, the Feedback API assembles group-level inputs using the cache aggregation or fallback computation paths described in Section 5.1.3, ensuring that P-code coverage summaries reflect the full cohort rather than a subset of evaluated cases. The group report is only generated when the proportion of individually evaluated cases meets a configured minimum completion threshold; below this threshold, a coverage warning is displayed to the teacher to indicate that the group summary may not yet be representative of the whole class.

The group AI report presented to teachers consists of two parts (Figure 5.7). The first is a multi-paragraph narrative describing patterns of main idea coverage across the cohort, identifying which ideas were well addressed by most students and which were systematically absent, and offering a brief instructional orientation based on the observed patterns. The second part, labelled Overall AI Outlook, provides a structured summary: a set of covered main idea chips, a set of missing main idea chips, and a short list of data quality or limitation notes – such as observations about synthesis length affecting assessment reliability or about the absence of explicit source-critical framing in student texts. These limitation notes, displayed with a lightbulb icon, are

intended to help teachers calibrate their confidence in the AI report. A disclaimer reading "AI-generated content for teacher guidance only" is displayed at the foot of the report. This is followed by a Group-level bookmark evaluation coverage view shown in Figure 5.8. The green cells indicate how students evaluated the relevant source texts (correctly, incorrectly, or without response), and red cells indicate how students evaluated the irrelevant source texts. The view enables teachers to identify class-wide patterns of source confusion.

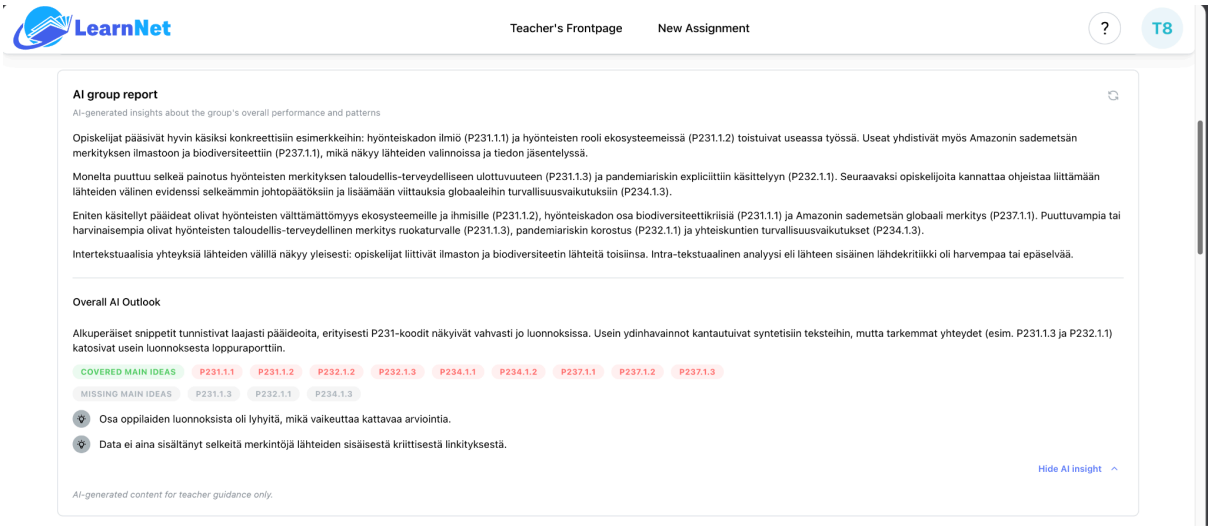

Figure 5.7. AI group report for the biodiversity loss task showing the cohort-level narrative, covered main ideas (green chips) and missing main ideas (grey chips) at the class level, and data quality limitation notes. The disclaimer at the foot of the panel indicates the report is intended as teacher guidance only.

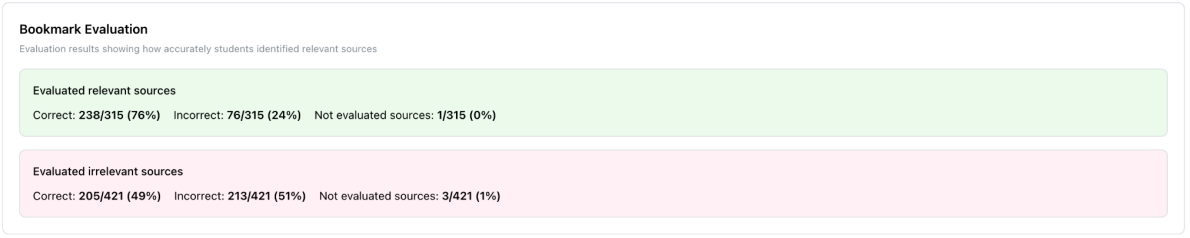

Figure 5.8. Group-level bookmark evaluation coverage view.

5.2.4 UI Integration in the Teacher Panel

The augmented panel integrates all of the components described in the preceding subsections into a unified teacher interface that is designed to minimise navigation overhead while supporting both

quick overview and detailed inspection. The interface is organised around two primary levels of view: the individual student view and the group view.

The individual student's view (Figure 5.9) is structured around two tabs: Overview and Process, which are accessible at the top of the student page. The Overview tab presents the content artefacts produced by the student: the full synthesis text with word count, the search terms used during source exploration displayed as chips, a grid of bookmarked source texts each labelled with a Relevant or Irrelevant classification badge and rated on Used, Relevance, and Reliability dimensions, and the list of snippets captured from sources This tab gives the teacher a full picture of what the student produced and which sources they engaged with, without AI involvement.

The Process tab integrates both the deterministic process summary and the AI-generated evaluation content. On opening the tab, the teacher sees the four at-a-glance statistics and the milestones timeline at the top of the page. Below, the AI evaluation panels for synthesis and snippets are displayed side by side. At the bottom of the page, the filterable chronological timeline, the bookmarks-by-text view, and the snippets-by-text view allow the teacher to trace specific events or source interactions in more detail. The vertical organisation of the Process tab reflects a deliberate information hierarchy: the fastest indicators appear at the top, and detailed evidence is available further down on demand.

The group view presents the class-level information in a scrollable page containing three main sections. The bookmark evaluation coverage display appears first, providing the colour-coded overview of how students across the cohort engaged with relevant and irrelevant source texts. The P-code coverage tables follow, showing synthesis and snippet coverage counts for each main idea across the full cohort. The AI group report is displayed below the coverage tables, presenting the narrative assessment and the overall AI Outlook.

Across both the individual and group views, the interface follows a consistent visual language: green indicates positive or relevant coverage, red or orange indicates low count or irrelevant content, and muted grey is used for incomplete items. AI-generated content is consistently labelled to distinguish it from deterministic data, and the disclaimer on the group report reinforces that AI outputs are intended as review support rather than authoritative assessments. This visual consistency was a deliberate design choice to reduce the cognitive effort required to interpret the

panel and to maintain a clear boundary between factual activity data and AI-generated inference [13], [16].

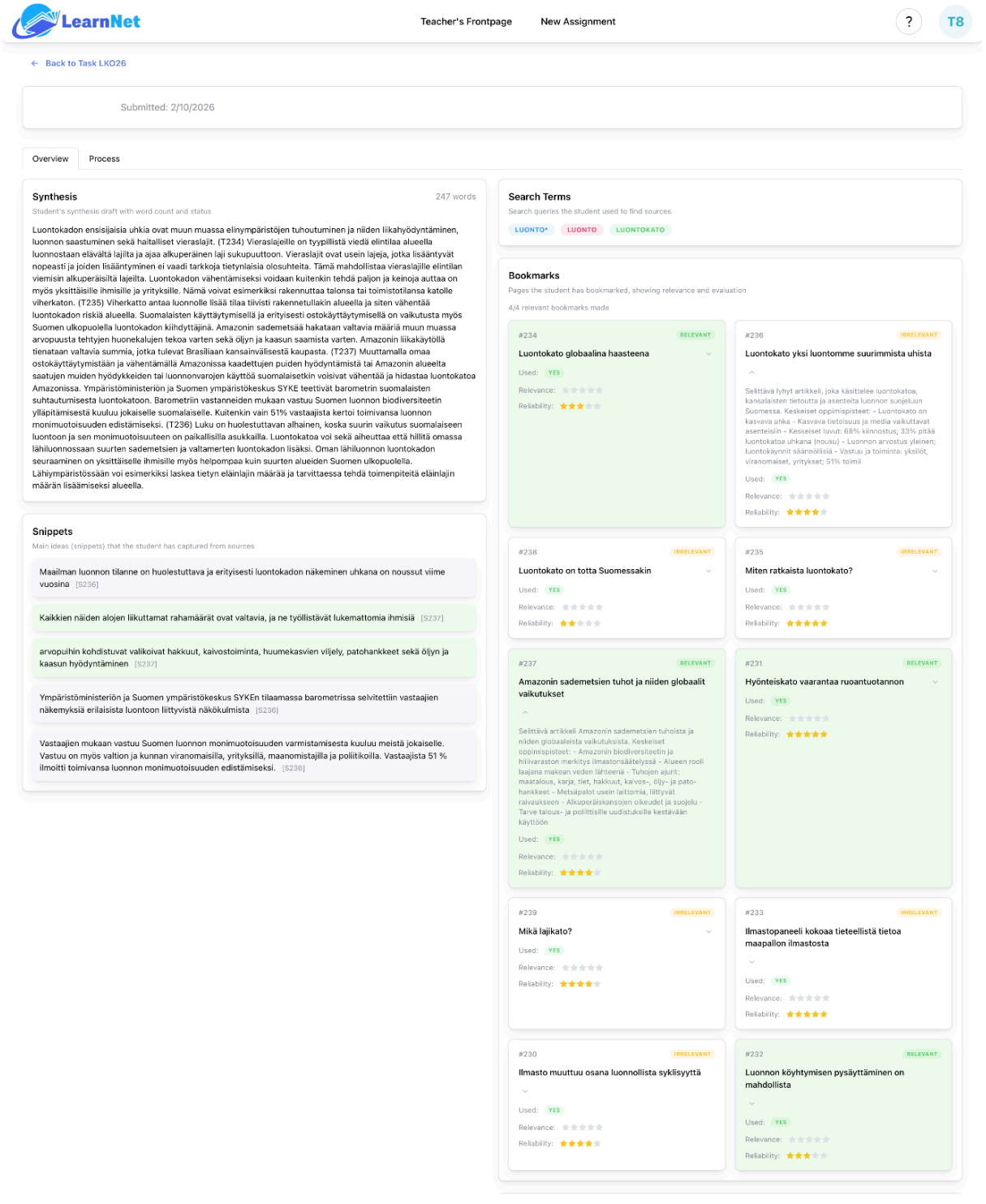

Figure 5.9. Individual student Overview tab showing the synthesis draft, search terms used, and the bookmarked source grid with relevance and reliability ratings. Relevant and Irrelevant labels reflect the student's own source evaluation judgments.

5.3 Guardrails, Verification, and Logging

Deploying a language model within an educational tool intended for teacher use introduces risks that go beyond those typical of general-purpose LLM applications. Inaccurate evidence quotes, malformed outputs, or inconsistent responses across similar inputs could undermine teacher trust, produce misleading assessments of student work, and compromise the validity of findings derived from AI-generated content during the evaluation study. The augmented panel, therefore, incorporates a set of guardrails and verification mechanisms at the post-generation stage, a logging infrastructure that captures both teacher interaction data and LLM performance data, and a defined set of metrics used to assess reliability and integrity across all generation calls. Each of these components is described in the subsections below.

5.3.1 Schema Validation and Source Verification

Each LLM endpoint in the FeedbackAPI is designed to produce output in a predefined structured schema. The prompt instructs the model to return a JSON object conforming to the expected field structure for that endpoint, including fields for the overall feedback narrative, the list of covered and missing P-codes, and for each P-code a score, a rationale string, and an evidence excerpt. Schema validation is applied immediately after the model response is received and before any downstream processing takes place. If the response does not conform to the expected schema, the output is treated as a generation failure and is not written to cache or presented to the teacher interface. This validation step prevents malformed or partially generated responses from propagating into the panel.

Beyond structural validation, the system applies a source verification step to the evidence excerpts included in each P-code breakdown. Each excerpt provided by the model is checked against the student's actual synthesis text and snippet content to confirm that the quoted passage exists in the source material. This check addresses a well-documented failure mode of large language models in which the model generates text that resembles a quotation but does not correspond to anything the student actually wrote [20]. Where an excerpt cannot be verified, a penalty is applied to the associated P-code score, and the evidence item is marked accordingly in the output payload. The teacher-facing display reflects this through the Verified and Not Verified indicators on each evidence card, so that a teacher who notices a low score can also see whether it reflects a genuine gap in the student's synthesis or a verification failure.

Verified outputs are written to cache with provenance metadata that records the prompt version used, the generation timestamp, and whether the output was produced fresh or aggregated from earlier cached evaluations. This provenance record allows any cached result to be traced back to the conditions under which it was generated, which is relevant both for study reproducibility and for future debugging or auditing of the system. All this is shown as a flow chart in Figure 5.10.

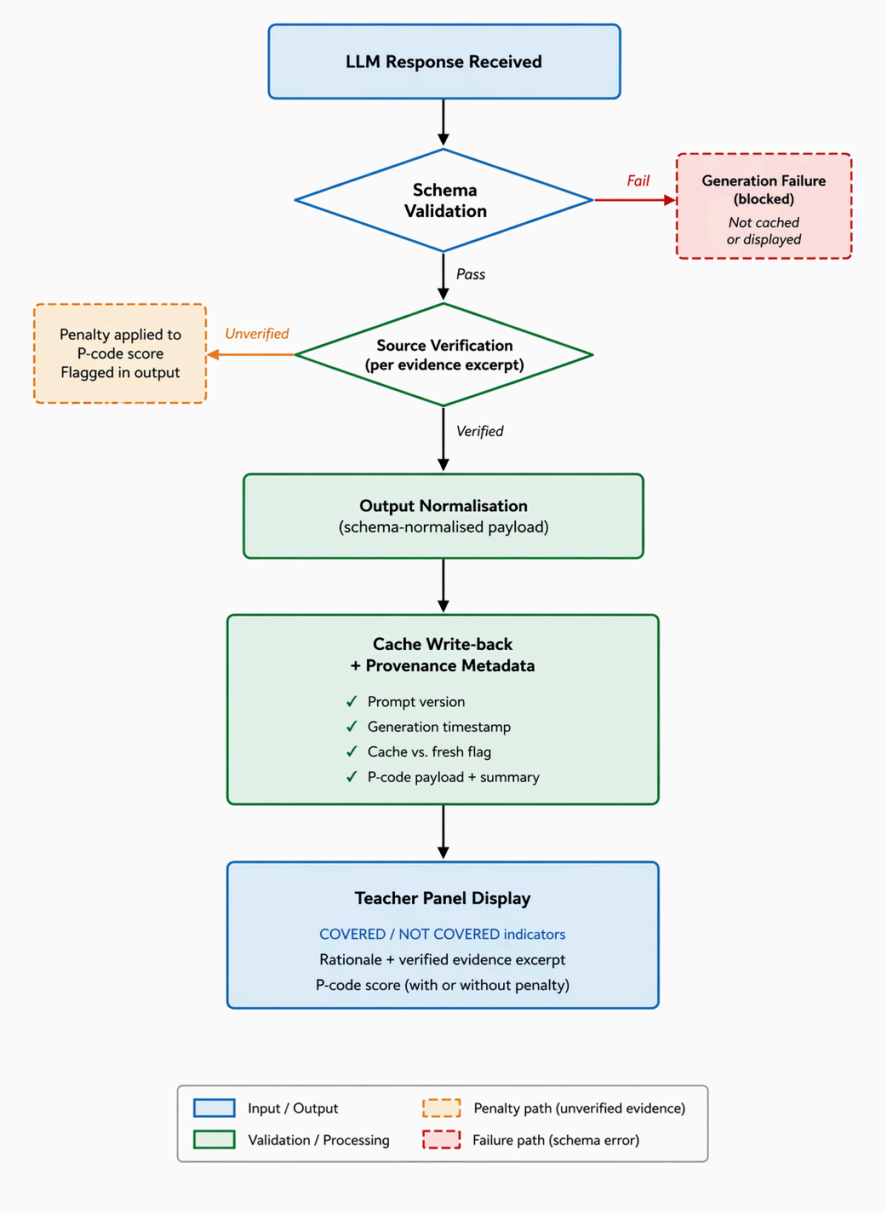

Figure 5.10. The post-generation validation pipeline flowchart: schema check, source verification, penalty application, and cache write-back with provenance metadata.

5.3.2 Logging and Study Instrumentation

Two distinct logging systems were operated during the evaluation study, each serving a different analytical purpose and capturing data at a different level of the system stack.

The first system is teacher-session telemetry recorded through a Firebase real-time database. Firebase was used specifically for this layer because it allows lightweight event capture without modifying the core LearnNet backend. During each augmented review session, the panel emitted timestamped events each time a teacher interacted with the components or navigated between sections of the panel. These events were intended to capture the sequence and duration of teacher interactions with the different panel components, providing a behavioural record of how teachers moved through the interface during the augmented review block. The telemetry data are used in the analysis of RQ3, specifically in relation to which panel features attracted teacher attention and in what order.

The second system is the LLM service log maintained by the FeedbackAPI for every call made to the GPT-5 API during the study sessions. For each generation request, the log records the following fields: a request identifier, the endpoint called, the input token count, the output token count, the server response time in milliseconds, the cache status (whether the result was served from cache or generated fresh), and a success or failure flag indicating whether the response passed schema validation. These records are the primary data source for addressing RQ4 regarding the technical feasibility and integrity of the AI components under realistic study conditions.

In addition to these two active logging systems, the FeedbackAPI cache stores a provenance record alongside each validated output, as noted in Section 5.3.1. While the cache itself is not a study instrument in the strict sense, the provenance metadata it contains enables post-hoc verification that the AI outputs reviewed by teacher participants during the augmented block correspond to specific generation events in the LLM service log, which is important for triangulating survey and think-aloud findings against the technical performance record.

It should be noted that the completeness of the Firebase telemetry across the three evaluation sessions was not uniform. Where telemetry records are incoherent for a given session, session screen recordings are available to supplement the interaction analysis, as described in Section 4.2.1. The LLM service logs, by contrast, were recorded server-side and are complete for all calls made during the study.

5.3.3 Metrics for Reliability and Integrity

A defined set of technical metrics is used to assess the reliability and integrity of AI-generated outputs across the evaluation study. These metrics are derived from the LLM service logs described in Section 5.3.2 and are reported descriptively in Section 6.4. The choice of metrics reflects the practical requirements of deploying an LLM-based tool in a teacher-facing context, where response latency, output consistency, and generation completeness are all relevant to usability.

The first metric is **response time**, defined as the elapsed time in milliseconds between the FeedbackAPI submitting a generation request to the GPT-5 API and receiving the full response. Response time is relevant to the usability of the panel because a teacher who opens a student case and waits an extended period for AI content to appear may disengage from the panel or form an impression of the case before the AI output has loaded. For cached responses, the effective response time experienced by the teacher is negligible; the metric, therefore, applies specifically to fresh generation calls. The distribution of response times across all fresh generation calls during the study is examined to identify whether any calls exceeded a threshold that would represent a meaningful disruption to the review workflow.

The second metric is the **output-to-input token ratio**, computed as the number of output tokens divided by the number of input tokens for each generation call. Because the prompt structure is held constant across all calls of the same endpoint type, substantial variation in this ratio may indicate that the model truncated its response before completing the required schema, or conversely that it generated excessive content beyond the expected output length. Both of these conditions are operationally undesirable: truncation produces incomplete or invalid schema output, while excessive generation inflates cost and may indicate prompt drift. A stable token ratio across calls of the same endpoint type is therefore treated as an indicator of generation consistency.

The third metric is the **schema validation failure rate**, defined as the proportion of generation calls whose output did not pass the schema validation step described in Section 5.3.1. A low failure rate indicates that the prompting strategy reliably elicits structurally conformant output from the model. A high failure rate would suggest that the prompt or the model configuration requires revision before the system could be considered ready for deployment beyond the evaluation context.

The fourth metric is the **source verification pass rate,** defined as the proportion of evidence excerpts across all evaluated cases that were successfully verified against the student's source text

without penalty. A high pass rate indicates that the model reliably grounds its evidence claims in the student's actual writing. A low pass rate would raise concerns about the trustworthiness of the evidence previews presented to teachers and would suggest that the post-processing penalty mechanism is compensating for systematic model behaviour rather than isolated failures.

Together, these four metrics provide a multi-faceted technical picture of system performance that complements the perceptual reliability judgments gathered through the think-aloud and micro-survey data. The relationship between objective technical performance and teacher-perceived reliability is discussed in Section 6.5 and in the broader evaluation discussion in Section 7

5.4 Key Implementation Decisions and Challenges

The development of the augmented panel involved a series of technical and design decisions that shaped the final system in ways not fully visible from the architecture description alone. Three areas of challenge were particularly significant: managing the practical constraints of latency, token limits, and API cost; navigating trade-offs in the interface between depth of information and usability; and refining the system across multiple development iterations in response to output quality problems and evolving understanding of teacher needs. Each of these is discussed in the subsections below.

5.4.1 Latency, Token limits, and Cost

Token limit constraints were encountered during development and required substantive changes to the prompting strategy. The most direct manifestation was a `max_output_tokens` error that occurred even when the configured output limit appeared moderate. This indicated that the combined size of the prompt structure and the expected response was sufficient to exhaust the available token budget for certain requests, particularly for richer report types where the input included multiple student syntheses or extended source material alongside the full set of reference P-codes and their descriptions. The risk was most acute at the group reporting level, where the number of student syntheses passed into a single prompt could grow large enough to destabilise generation.

The practical response to this constraint was to treat the LLM as a processor of structured, pre-filtered inputs rather than as an end-to-end interpreter of raw task content. Instead of passing full source texts, the system was redesigned to pass only the content directly relevant to the

evaluation task: selected P-codes with their descriptions, the student synthesis or snippet text, and precomputed source summaries. Verification and aggregation logic that had initially been handled within the prompt was moved into the backend preprocessing layer, reducing prompt size and improving generation stability. This shift is consistent with the broader pipeline architecture described in Section 5.1.3, where the separation of deterministic preprocessing from LLM inference is a deliberate structural choice rather than an incidental one.

Cost was a parallel concern throughout development. Repeated on-demand generation for every teacher's view of the panel would have made the system expensive to operate at scale and introduced unnecessary API call volume during the evaluation study itself. The primary architectural response was to shift AI evaluation from a pull model, where generation happens when a teacher opens a case, to a push model, where evaluation is triggered when the student submits the task, and the result is stored in cache. Teachers subsequently retrieve the cached result rather than triggering a fresh generation call. This design stabilised group-level statistics that had previously been recomputed through repeated AI calls and reduced the overall number of generation requests needed per study session. A teacher-triggered regeneration option was retained to allow cache refresh when required, but this was positioned as an exceptional action rather than the default interaction.

Latency was a usability concern from early in development. Initial tests showed that synchronous LLM generation on demand produced waiting times that were long enough to disrupt the review flow, particularly when a teacher opened a student case and expected the panel to load promptly. The caching architecture described above addressed this directly: once a case had been evaluated and cached, subsequent views were served with negligible latency. The micro-survey used in the evaluation study included explicit questions about whether response time disrupted review flow, reflecting the recognition that perceived latency remained a usability risk even after caching was introduced, since not all cases would necessarily be in cache at the time a teacher first accessed them.

5.4.2 Interface Trade-offs and Feature Prioritisation

A central tension in the interface design was between providing comprehensive information and maintaining the readability needed for fast teacher review. The panel needed to present process data, source evidence, and AI-generated assessments within a single coherent view, and decisions about

what to include, what to surface at the top level, and what to make available on demand reflected ongoing judgments about which information teachers would find most useful under time pressure.

One clear prioritisation that emerged from development feedback was the relative weight given to synthesis-level assessment versus snippet-level detail. Feedback indicated that overall synthesis-level support was more useful to teachers than highly granular snippet-level analysis, which led to the synthesis evaluation view being treated as the primary teacher-facing output. Snippet evaluation was retained later as a parallel panel because it captures a distinct phase of the student's work process: snippets represent the student's idea extraction from sources, and comparing snippet coverage with synthesis coverage allows teachers to identify where ideas present in a student's notes were lost before reaching the final text. However, the snippet panel was positioned as supplementary detail rather than a primary review instrument.

Several features that would be natural extensions of the teacher-support concept were not included in the final implemented panel. An explicit override or correction mechanism, through which a teacher could formally mark an AI judgment as incorrect or record their own assessment alongside the AI output, was not part of the implemented interface. The panel was designed as a **read-only** review support tool in which the teacher retains full interpretive authority, but the system does not provide a structured channel for the teacher's response to feed back into the AI output. Similarly, more advanced filtering and group-level manipulation options, such as sorting students by AI assessment category or filtering the group view by P-code coverage pattern, were considered but remained outside the scope of the current implementation. These omissions reflect a deliberate decision to prioritise a functional and interpretable first version over a feature-complete system, consistent with the iterative evaluation logic of Design Science Research [34], [35].

The overall layout of the individual student view, organised into an Overview tab for content artefacts and a Process tab for behavioural and AI-generated data, also represents a trade-off. Separating the two views keeps each tab manageable in length and allows teachers to choose their starting point depending on their review strategy. The cost is that switching between a student's synthesis text and the AI evaluation of that synthesis requires a tab change, which adds a small navigation step. An alternative single-page layout was not pursued because it would have produced a page too long and dense for efficient first-pass review.

5.4.3 Iterations During Development

The development of the augmented panel involved multiple rounds of revision across both the prompting strategy and the scope and emphasis of the reports presented to teachers. These iterations were driven by output quality problems, multilingual requirements, and evolving understanding of what teachers found useful, consistent with the build-evaluate-refine cycle that characterises Design Science Research [34], [35].

Prompt engineering was one of the most persistent iterative challenges. Early prompt versions produced outputs that were vague, awkwardly phrased, or inconsistent across similar inputs. A particularly demanding aspect was the multilingual requirement: the panel needed to generate teacher-readable summaries and reports in Finnish, English, and Swedish, reflecting the linguistic context of Finnish university teacher education. Achieving clear and natural output in Finnish proved more challenging than in English, and prompt design was repeatedly revised to improve the quality, specificity, and register of Finnish-language outputs. The need to support multiple languages also influenced decisions about output structure, since a more tightly constrained output schema reduced the variability introduced by language-dependent generation patterns.

Beyond language quality, prompt iterations addressed the problem of output size and instability. Early prompts that passed too much raw content to the model produced responses that were either truncated, structurally incomplete, or internally inconsistent across P-code entries. As described in Section 5.4.1, the response to this was to move more preprocessing logic outside the prompt and to constrain the expected output more tightly. Over successive iterations, this produced more stable and predictable responses, at the cost of some reduction in the richness of content the model could draw on within a single generation call.

The scope and emphasis of the reports also evolved across development. Early feedback indicated that teachers found the higher-level synthesis summary and the consolidated list of covered and missing main ideas immediately useful. Subsequent iterations shifted the report design toward cleaner, more concise summary outputs, with detailed evidence available on demand through expandable P-code cards rather than presented in full by default. The addition of snippet coverage as a distinct evaluation layer also came later in the development process. Its purpose was to allow teachers to assess idea loss across the task: by comparing which P-codes were present in a student's snippets against which appeared in the final synthesis, teachers could see where ideas captured

during reading did not survive into the written text. This comparison was not part of the original panel concept but emerged as a meaningful analytical dimension once the synthesis evaluation was functional.

A further area of iteration concerned the location of verification logic. Initial designs attempted to handle evidence verification within the prompt itself, asking the model to confirm that its quoted evidence appeared in the student's text. This approach was found to be unreliable: the model would sometimes confirm quotes that did not exist verbatim, reflecting the general tendency of large language models to produce confident outputs even when the underlying content does not support them [20]. Moving quote verification into a deterministic post-processing step, as described in Section 5.3.1, produced more consistent results and ensured that the penalty mechanism operated on objectively verifiable criteria rather than on the model's own self-assessment.

6 Results

This chapter reports the findings of the evaluation study directly against the four research questions. The presentation follows a consistent structure across research questions: qualitative evidence from think-aloud verbalisations and retrospective interviews is reported first, followed by corroborating or qualifying evidence from the micro-survey responses. Where quantitative interaction log data or LLM service metrics are available, these are integrated into the relevant subsections. Section 6.1 first describes the participants and the study material used in the sessions.

6.1 Participants and Study Material

6.1.1 Teacher/Researcher Participants

Three participants took part in the evaluation study, each completing a single session covering both the manual review condition and the augmented review condition. The participants are referred to throughout this chapter as T1, T2, and T3, assigned in session order. All three were affiliated with the University of Turku and had experience working with LearnNet as a platform, either as teacher educators, researchers, or both. Two of the three participants held researcher-practitioner roles at the time of the study; the third, T3, was also a school teacher with direct classroom experience, which gave her a practitioner perspective on the kinds of judgments a teacher conducting a synthesis task review would be expected to make.

The three participants differed notably in their degree of familiarity with the content domain of the student synthesis task. T1 had moderate familiarity with the biodiversity loss topic and engaged with the AI-generated assessments partly as a source of domain orientation, using the panel to identify ideas that the student appeared to have addressed or missed. T2 had high domain expertise and was consequently able to identify both accurate and incomplete AI assessments based on their own knowledge of the subject matter. T3 had lower prior familiarity with the specific task content and adapted to the augmented panel interface most quickly among the three, showing a strong disposition to accept AI-generated coverage indicators as a reliable starting point for review. Her experience as a practising teacher gave her a distinct perspective on the practical relevance of process data for classroom use, as reflected in her comment during the think-aloud:

> *"Tämä prosessipuoli on kyllä kiinnostava, koska opiskelijoiden ja oppilaiden prosessista ei helposti saa tietoa, varsinkin kun he tekevät itsenäistesti näitä, niin tyypillisesti*

koulutehtäviä ainakin osittain tehdään ihan itsenäistesti, niin se on kiinnostavaa, että voi katsoa, mitä tapahtuu." (T3, originally in Finnish)

["This process panel is quite interesting, because information about the process of students and pupils is not easy to obtain, especially when they complete tasks independently. School tasks are typically done at least partly on one's own, so it is interesting to be able to see what happens."]

These differences in domain familiarity and professional background are relevant to interpreting variation in responses across participants, particularly for RQ2 and RQ3, and are discussed further in the relevant sections.

Sessions were conducted individually with one participant at a time. The think-aloud protocol was active throughout both the manual and augmented review blocks. Each session concluded with a retrospective semi-structured interview of approximately ten minutes. One of the three participants conducted parts of their session in Finnish; passages originally in Finnish are quoted in the original alongside an English translation, and the translation is noted where it introduces interpretive uncertainty.

6.1.2 Student Cases and Tasks used in the Sessions

The student material reviewed during the sessions was drawn from the biodiversity loss synthesis task conducted as part of a university course. A total of 104 students completed the task, producing a short written synthesis of 200 to 350 words based on a set of ten source texts provided within the LearnNet environment. The source set comprised four relevant texts addressing different dimensions of biodiversity loss, four irrelevant texts on related but off-topic subjects, and two pseudoscientific or fake texts designed to test students' source evaluation skills. Twelve predefined main ideas, three per relevant source text, served as the analytical anchors for both the AI-generated assessments and the teacher review judgments in this study.

Across the three evaluation sessions, participants collectively reviewed six student cases in the manual condition and nine cases in the augmented condition, yielding fifteen case-level evaluation records in total. Each participant selected their own cases independently from the full pool of 104 student submissions; cases were not pre-assigned or curated by the researcher. The two manual cases reviewed by each participant were revisited in the first two augmented cases of the same

session, allowing each participant to compare their manual judgment with the AI-generated assessment for the same student. The third augmented case per participant was a new student selected during the extended review block. The distribution of cases across participants and conditions is summarised in Table 6.1.

Table 6.1. Distribution of reviewed cases across participants and conditions.

Participant	**Manual cases (7 min each)**	**Augmented cases(7 min each)**	**Extended augmented case (13 min, both tabs)**	**Total cases per participant**
T1	2	2 (same students)	1 (new student)	5
T2	2	2 (same students)	1 (new student)	5
T3	2	2 (same students)	1 (new student)	5
Total	**6**	**6**	**3**	**15**

6.2 RQ1: Transparency

RQ1 asks what process summaries and LLM evidence previews helped teachers form a quick, confident understanding of a student's synthesis process. Findings are reported across three subsections covering the specific panel elements that supported the first insight (6.2.1), teacher perceptions of clarity and sense of control during the augmented review (6.2.2), and observed usage patterns in the panel (6.2.3).

6.2.1 Process Summaries and Evidence Previews that Supported the First Insight

Across all three participants, two elements of the augmented panel were consistently identified as the most useful for forming a rapid first understanding of a student case: the overall AI synthesis evaluation and the main idea coverage indicator. Both were mentioned spontaneously during think-aloud and confirmed in the retrospective interview, and both were rated as the features participants would most want to retain if only a subset of features could be kept.

The overall AI synthesis evaluation, which presents a short narrative summary of the student's key content choices and apparent gaps, reduced the need for participants to read the full synthesis text before forming an initial impression of the case. T1 described this directly during think-aloud:

> *"The overall feedback was maybe most helpful... the overall feedback is it, yeah." (T1)*

The main idea coverage indicator, displayed as a colour-coded table showing which predefined ideas appear to be present or absent in the student's synthesis and snippets, was cited by all three participants as the element that most efficiently communicated the scope of a student's content coverage. T1 noted that it communicated the key judgment directly:

> *"It tells in like quite straightforwardly which ideas were covered and which were weakly or not at all covered." (T1)*

T3, speaking from her experience as a practising teacher, highlighted the process view specifically as the element that gave her the most immediate orientation to a student case:

> *"That process view where I could see like evaluations or what snippets were wrong or right, or that kind of things. Those gave me quickly a view of students, like how the student managed in task." (T3)*

The evidence preview within the main idea breakdown, which links each coverage assessment to a specific passage drawn from the student's text, was identified by T1 and T2 as an important transparency mechanism. It allowed participants to verify the basis of an AI coverage claim without needing to search the synthesis manually. T1 described the value of this link as follows:

> *"Because there is the evidence, what the student actually wrote, and what the AI has written." (T1)*

T2 similarly noted that the source evaluation view made it possible to assess a student's source selection decisions at a glance:

> *"There I can see pretty easily which, whether they have found the relevant texts." (T2)*

The bookmarks-by-text and snippets-by-text views were cited as useful by T2 and T3 for obtaining a structured overview of which sources a student had engaged with and what preliminary ideas they had extracted. T2 described this as providing background orientation before reading the synthesis:

"It gives some like preliminary information about their ideas." (T2)

Micro-survey ratings for the question of how quickly participants felt they understood what a student had done in the augmented condition are shown in Table 6.2. Ratings were provided on a five-point scale (1 = very slowly, 5 = very quickly).

Table 6.2. Quick understanding rating on a five-point Likert scale

Participant	**Quick understanding rating (augmented condition, 1-5)**
T1	3-4
T2	4
T3	4-5

The pattern of ratings indicates that all three participants found the augmented panel supported faster case understanding than manual review, though the degree of benefit varied. The higher ratings from T3 are consistent with that participant's faster adaptation to the panel interface and lower prior domain familiarity, which meant the AI summary provided more orienting information relative to what T3 could derive independently from reading the synthesis.

6.2.2 Teacher Perceptions of Clarity and Control

Perceptions of clarity and control diverged markedly between the individual student's view and the group-level view of the augmented panel. The individual view was described by all three participants as clear and interpretable, with the colour-coded coverage indicators and the evidence-linked breakdown providing a legible structure that required little prior experience with the panel to navigate. In contrast, the group-level aggregation view was found confusing by all three participants, with the numerical percentages and cohort-level summaries described as unintuitive and difficult to verify.

On the group-level view, T1 noted uncertainty about the basis of the figures presented:

"I wasn't sure of the numbers and percentages." (T1)

T3 echoed this in her own interview, expressing difficulty understanding the origin of the figures displayed:

> *"Niin, siellä oli joitain semmosia lukuja, joista mä en niinku ymmärtänyt ihan, että mistä ne tulee. Ja sitten, vaikka mä tykkäsin siitä AI-yhteenvedosta, niin, tai siitä sais niinku nopeasti sen kuvan." (T3, originally in Finnish)*
>
> *["Yes, there were some figures there that I did not quite understand where they came from. And then, even though I liked the AI summary, from that you can get a quick picture."]*

T2 expressed a similar reaction, stating that the percentages were not intuitive. The common difficulty appeared to be that the group-level aggregation logic was not transparent to teachers: it was not clear to participants how a cohort percentage had been computed or what threshold or underlying data it represented, making it difficult to calibrate confidence in the number. This stands in contrast to the individual coverage indicators, where the link between the AI claim and the student's own text was directly visible through the evidence preview.

Perceptions of control during individual review were generally positive for T2 and T3, both of whom rated their sense of control with the evidence preview as 4 out of 5. T1 rated control lower, at 2 out of 5, reflecting a greater reliance on the AI panel to orient within the content domain and a corresponding sense that the panel was leading the review more than the participant was directing it. This finding is consistent with the domain familiarity pattern described in Section 6.1.1: participants with higher prior knowledge of the task content felt more in control of the review process, while participants with lower domain familiarity were more dependent on the AI summary and consequently felt the panel was setting the agenda.

The read-only nature of the panel, in which teachers can inspect and use AI outputs but cannot formally record a correction or override, was not raised as a source of frustration by any participant. T2 described the panel as functioning appropriately as a review support tool rather than as a decision-making system, and noted that the ability to verify claims against the visible student text was sufficient to maintain a sense of informed judgment.

6.2.3 Observed Usage Patterns in the Panel

Firebase and Video interaction logs from the evaluation sessions show that participants followed broadly similar navigation patterns across the augmented condition, while differing in the depth to which they engaged with individual panel components.

In terms of tab navigation, six of the nine augmented cases, two per participant, began with the Process tab, where the deterministic process statistics, milestone timeline, and reading event log are displayed. The remaining three cases, all the third case reviewed by each participant, began with the Overview tab. This pattern suggests that participants who had become more familiar with the panel structure by their third case tended to move toward the synthesis-level Overview tab first to read about the synthesis, while defaulting to the more granular process view during earlier cases when they were still orienting to the student's reading behaviour.

All participants accessed the AI synthesis evaluation panel in every augmented case reviewed, confirming that this component was central to the review workflow across all three participants. The main idea coverage table was similarly accessed in all cases, consistent with participants' self-reports that it was the element they relied upon most for a quick initial understanding.

The depth of engagement with individual main idea evidence cards varied across participants. T2 expanded the largest number of evidence cards per case, with four cards in Case 1 and five in Case 3, indicating close engagement with the evidence excerpts underlying each coverage assessment. T1 expanded fewer cards overall and did not expand any evidence cards in Case 1, relying more on the summary-level coverage indicators. T3 expanded between zero and three cards per case, with the third case showing no card expansion, which may reflect a higher level of trust in the summary indicators as the session progressed. Evidence quote inspection was observed for T2 in all three cases and for T1 and T3 in selected cases.

Bookmark and snippet views were accessed by T1 in Case 2 and by T2 and T3 across multiple cases, but were not universally consulted, confirming their secondary status relative to the AI synthesis evaluation and main idea coverage table. The timeline feature was accessed by T1 in Case 3 and by both T2 and T3 across cases, while the process summary view was opened by all three participants in at least one case. Taken together, the interaction patterns reflect a hierarchy of panel engagement in which the AI synthesis evaluation and main idea coverage table functioned as the

primary review entry points, with process-level and snippet-level components serving as secondary reference points consulted when participants wanted additional detail.

6.3 RQ2: Alignment

RQ2 asks to what extent the LLM evidence previews overlap with teacher-identified passages or ideas. Findings address the degree of overlap reported by participants (6.3.1), the patterns of mismatch that were observed (6.3.2), and participant views on the overall adequacy of the previews as a basis for review (6.3.3).

6.3.1 Overlap between Evidence Previews and Teacher-identified Passages

The micro-survey asked participants to characterise the degree of overlap between the AI assessment of each case and the judgment they had formed during their prior manual review. Responses across the augmented cases are summarised in Table 6.3.

Table 6.3. Responses to the degree of overlap of ideas between AI and human assessment

Participant	**Predominant overlap characterisation across reviewed cases**	**Range across cases**
T1	Some overlap	Partial to none on specific ideas
T2	Most overlap	Mostly consistent with occasional gap
T3	Full overlap	Consistently full alignment

The pattern shows a gradient from T1, who reported only partial overlap on a number of cases, through T2, who found alignment in most cases while noting specific gaps, to T3, who reported full alignment across all augmented cases reviewed. Taken together, these responses indicate that the AI evidence previews were at a minimum, partially aligned with teacher-identified content in the majority of cases, with no participant reporting systematic disagreement.

On the question of citation accuracy, no participant reported encountering a fabricated or hallucinated quote during the evaluation sessions. When participants checked the evidence excerpts presented in the panel against the student's actual synthesis text, they found the quoted passages to be present. T1 explicitly noted having verified the Amazon-related evidence claims manually and confirmed they matched the student's text. T2 acknowledged the theoretical risk of fabrication while noting it had not been observed in practice:

> *"Even though there might be that risk that it has some mistakes in it, which might be hard for me to see." (T2)*

The absence of detected fabrication across the sessions is consistent with the post-processing source verification mechanism described in Section 5.3.1, which checks evidence excerpts against the student's actual text and applies penalties where a match cannot be confirmed. It should be noted, however, that participants could only verify evidence claims against the portions of the student text visible in the panel; they were not able to audit the full verification logic of the system, and T2's caveat reflects a reasonable epistemic caution about claims the participant could not directly check.

6.3.2 Common Mismatch Patterns

While the overall level of alignment was positive, all three participants identified specific dimensions on which the AI assessments diverged from their own reading of the student work. Three patterns of mismatch emerged across the sessions.

The first pattern concerns coverage breadth versus content depth. The AI evidence previews assessed whether a predefined main idea appeared to be present or absent in the synthesis, but did not assess the depth or accuracy of the student's engagement with that idea. T1 identified several cases where the AI indicated coverage of an idea that the participant judged to be only superficially addressed, including missing framings around economic impacts and human rights dimensions of biodiversity loss. T2, drawing on higher domain expertise, similarly noted cases where the AI had characterised a student as having covered an idea when the relevant passage addressed only one aspect of a multi-part concept.

The second pattern concerns the distinction between irrelevant and pseudoscientific sources. The source set used in the biodiversity loss task included both sources that were irrelevant to the task

and sources that were deliberately fake or pseudoscientific. Both T2 and T3 observed that the panel did not clearly surface this distinction. T3 articulated the underlying conceptual problem directly:

> *"Tietysti se voi olla luotettava, mutta se ei ole välttämättä hyödyllinen tämän tehtävän näkökulmasta. Siinä on vähän se haaste just, että onko se tehtävän näkökulmasta hyödyllinen, niin se ei tule välttämättä ilmi tuossa, että onko se luotettava." (T3, originally in Finnish)*
>
> *["Of course, a source can be reliable, but it is not necessarily useful from the perspective of this task. That is somewhat of a challenge: whether it is useful from the task's perspective does not necessarily follow from whether it is reliable."]*

T2 raised a closely related observation, noting that some irrelevant sources were produced by credible institutions and were therefore reliably written, even though they did not address the task:

> *"Irrelevant texts can be reliable but not so useful for the task." (T2)*

Together, these observations highlight a gap in the current panel display: a student who engaged with an irrelevant-but-credible source and a student who engaged with a fake source present different pedagogical concerns, and the panel did not surface this difference clearly enough for participants to act on it directly.

The third pattern concerns the relative weight given to snippet-level versus synthesis-level evidence. T1 observed during think-aloud that the snippets a student captured from sources are an intermediate artefact rather than the final product of the synthesis process, and questioned whether the snippet evaluation panel was as informative as the synthesis evaluation for the purpose of teacher review:

> *"The snippets are a tool for the student to do the synthesis, so I think the synthesis is the main thing here." (T1)*

This observation is consistent with the design iteration described in Section 5.4.3, where synthesis-level assessment was prioritised in the panel layout over snippet-level detail. It also suggests that while the snippet evaluation provides useful diagnostic information for identifying idea loss between the reading and writing phases, its value for teacher review is secondary to the synthesis evaluation for at least some use cases.

6.3.3 Teacher Views on Adequacy of the Previews

All three participants judged the AI evidence previews as adequate for supporting an initial overview of a student case, while consistently noting that the previews did not substitute for manual verification in cases where the teacher needed to make a substantive pedagogical judgment about a student's understanding. This position was articulated most explicitly by T2, who framed the panel as a useful first-pass tool that required human confirmation before its outputs could be acted upon:

> *"We have to check manually things." (T2)*

T1 expressed a similar view, noting that the panel supported identification of potential issues but that the teacher retained the interpretive responsibility for determining whether a flagged gap reflected a genuine student misunderstanding or a limitation of the AI assessment. T3, who showed the highest overall alignment ratings, nonetheless noted the fake-versus-irrelevant distinction as a gap in the preview design that reduced its adequacy for source evaluation judgments.

The micro-survey item asking whether the evidence preview helped participants judge main idea coverage produced ratings of 3 out of 5 for T1, 4 out of 5 for T2, and 5 out of 5 for T3. The pattern mirrors the overall alignment ratings and is consistent with the domain familiarity explanation: participants who could independently verify coverage claims found the previews more credible and therefore more adequate as a basis for judgment, while participants who depended more on the AI output for domain orientation were more aware of its limitations.

A recurring theme across all three participants was that the adequacy of the previews was contingent on the ability to verify their basis in the student's actual text. Where the evidence link was clear, and the quoted passage was directly inspectable, participants felt the preview was adequate. Where the coverage indicator appeared without a verifiable evidence anchor, or where the aggregation logic at the group level was opaque, adequacy judgments were lower. This finding points to traceability as the key determinant of perceived adequacy rather than accuracy alone.

6.4 RQ3: Usefulness and Efficiency

RQ3 asks whether the augmented panel reduced time-to-first-insight and effort compared with manual review, and how teachers rated clarity and usefulness. Findings are reported across three subsections: time-to-first-insight and interaction patterns (6.4.1), perceived effort and usefulness (6.4.2), and use of the group-level view for class-level decisions (6.4.3).

6.4.1 Time-to-first-insight and Interaction Patterns

The evaluation session followed a prescribed timing protocol across both conditions. In the manual condition, each participant reviewed two student cases with seven minutes allocated per case. In the augmented condition, the same two students were revisited with the AI panel visible, again with seven minutes per case, followed by a third case reviewed over a thirteen-minute extended block in which both tabs were available. Review durations were therefore determined by the session design rather than by participant choice, and direct comparison of time-on-task across conditions reflects the protocol rather than an emergent efficiency difference.

Within these time allocations, video observations allow the structure of the review workflow to be described. In the manual condition, participants used the seven-minute window to read the student synthesis in the Overview tab and form an initial judgment, reporting a confidence level of 4 out of 5 across all manual cases. In the augmented condition, participants navigated between the Overview tab, where the student's written synthesis is displayed, and the Process tab, where the AI evaluation panels and main idea coverage table are located. The panel thus added a navigation layer that was absent from the manual condition, but also provided structured support that reduced the interpretive work required within each review window.

A proxy for time-to-first-insight is the navigation sequence that preceded each participant's first explicit evaluative comment during the augmented review. In all nine augmented cases, this first comment followed within the first one to two minutes of opening the panel and occurred after participants had consulted either the AI synthesis evaluation or the main idea coverage table, whichever they opened first. This pattern is consistent with participants' self-reports that these two features provided an immediate initial orientation to the case.

Block-comparison survey items asked participants directly whether the augmented panel had helped them reach a first understanding faster than in the manual condition. T1 rated this at 4 out of 5, T2 at 5 out of 5, and T3 at 5 out of 5. On the item asking whether the augmented condition involved fewer clicks, T1 and T2 each rated this at 3 out of 5, reflecting the additional navigation steps required to move between panel tabs and verify individual evidence cards; T3 rated it at 5 out of 5. These ratings suggest that the efficiency benefit of the augmented panel was clearest for T3, consistent with the domain familiarity pattern observed throughout the study, and somewhat more

qualified for T1 and T2, who navigated the panel alongside their own domain-grounded review reasoning.

6.4.2 Perceived Effort and Usefulness

Overall usefulness ratings for the augmented panel were high across all three participants. On a five-point scale, T1 and T2 each rated the panel at 4 out of 5, and T3 at 5 out of 5. The qualitative data contextualise these ratings and reveal that the nature of the efficiency benefit was more nuanced than the summary scores suggest. The results are summarised for perceived effort in Table 6.4.

Participants described a consistent trade-off in the augmented condition: the panel reduced the time and effort required to read and interpret the synthesis text itself, since the AI summary provided an immediate orientation to the student's content choices. However, this saving was partly offset by the effort required to interpret the AI-generated indicators, navigate between panel sections, and verify AI claims against the visible student text. T1 described this interpretive cost directly:

> *"Because this topic is not that familiar to me, it required some effort." (T1)*

T2 noted a specific friction in the session workflow, where navigating between the student's written synthesis in the Overview tab and the AI evaluation panels in the Process tab added to the cognitive load of each case:

> *"I had to read a lot and like go back and forth between the synthesis and the Process tab." (T2)*

T3, who adapted most quickly to the panel interface, reported a clearer efficiency gain, rating both time-to-first-insight and overall usefulness at the maximum of the scale. Her perspective as a practising teacher was reflected in a positive assessment of the process data, specifically:

> *"Tuo prosessitieto on tosi hyodyllinen. Siina on joitakin. Ajattelisin, etta voisi aika nopeasti hyodyntaa tallaisenaankin." (T3, originally in Finnish)*

> *["That process information is really useful. There are some areas for development. I would think it could be used quite quickly, even in its current form."]*

This pattern is consistent with the domain familiarity effect observed across the other RQs: for a participant with lower prior knowledge of the content, the AI summary provided substantial added

value as an orientation resource, whereas participants who could independently evaluate a synthesis had to weigh that independent judgment against the AI output, adding interpretive steps.

Table 6.4 summarises key perceived effort and usefulness ratings from the block comparison section of the micro-survey.

Survey item	**T1**	**T2**	**T3**
Overall usefulness of the augmented panel (1-5)	4	4	5
Panel helped reach the first insight faster (1-5)	4	5	5
Augmented panel required fewer clicks (1-5)	3	3	4
Clarity of augmented panel display (1-5)	3	4	3
Evidence preview helped judge idea coverage (1-5)	3	4	5
Sense of control with evidence preview (1-5)	2	4	4

T1's faster-first-insight rating of 4 out of 5 reflects a positive but qualified efficiency experience: the panel accelerated the initial orientation to each case, though the unfamiliarity with the content domain required additional interpretive effort that partly offset the time saving. T2 and T3 both rated faster first insight at 5 out of 5. The lower click-reduction ratings from T1 and T2 (both 3 out of 5) correspond to the additional navigation steps involved in moving between panel tabs and inspecting evidence cards; T3's rating of 5 out of 5 on this item suggests a more streamlined interaction pathway for that participant. The clarity rating for T3 at 3 out of 5 is notable given that participants' otherwise high ratings and reflects ongoing uncertainty about specific panel elements, particularly the group-level aggregation view.

When asked which features would be most valuable to retain, all three participants identified the overall AI synthesis evaluation and the main idea coverage table as essential. The timeline and process summary were rated as useful by T1 and T2 but were not described as critical, while the snippet evaluation panel was seen as supplementary. These preferences are consistent with the

interface trade-off decisions described in Section 5.4.2, where synthesis-level summary was prioritised over granular snippet-level detail.

6.4.3 Use of the Group-level view for Class-level Decisions

All three participants completed the group-level review segment of the evaluation session, accessing both the AI narrative group report and the main idea coverage aggregation view. The group review was allocated approximately seven minutes in the primary review block, with an additional two-minute period for targeted follow-up questions.

On the overall usefulness of the group report for revealing class-wide patterns, T1 rated the group view at 4 out of 5, and both T2 and T3 rated it at 3 out of 5. When asked specifically whether the group-level information would support a class-level instructional decision, ratings were T1 at 4 out of 5, T2 at 4 out of 5, and T3 at 5 out of 5. The divergence between overall usefulness and decision-support ratings reflects a pattern in the qualitative data where participants found the group-level narrative useful for identifying broad coverage patterns but found the quantitative aggregation indicators difficult to interpret in a way that would directly guide an instructional action.

T1 read the AI group narrative report and described it as informative, but did not engage with the percentage-based coverage table, indicating a preference for the narrative summary over aggregated figures. T2 engaged with both the narrative and the percentage view but described the percentage display as confusing, specifically noting difficulty in interpreting the colour coding used to distinguish coverage levels across the cohort. T3 accessed both components but described a similar difficulty in forming an overall picture:

> *"Silloin en saanut tasta kokonaiskuvaa. Kuitenkin varsinkin hammenta AI-ryhmaportti siita, etta hyonteiskadon ilmio tulisi siella. Se ei nay paaideoiden kattavuudessa oikeastaan." (T3, originally in Finnish)*
>
> *["Then I did not get an overall picture from this. What was particularly confusing was the AI group report: the phenomenon of insect decline should have appeared there, but it does not really appear in the coverage of main ideas."]*

T3 also independently raised the fake-versus-irrelevant source distinction as a gap in the group-level display, a concern that had also been identified at the individual case level in Section 6.3.2.

These findings are consistent with the individual-level clarity data reported in Section 6.2.2. The AI narrative group report functioned as a reasonable starting point for class-level pattern identification, conveying common strengths and weaknesses in accessible language. The quantitative aggregation view, by contrast, was not yet sufficiently transparent in its computation logic to support direct pedagogical decisions without additional explanation or redesign of the display.

6.5 RQ4: Technical Feasibility and Integrity

RQ4 asks whether the LLM suggestion and evidence service can meet classroom constraints while preventing poor citations. Findings address the reliability of outputs and citation verification (6.5.1), performance and latency (6.5.2), and teacher trust and concerns as reported qualitatively (6.5.3).

6.5.1 Reliability of Outputs and Citation Verification

The Langfuse observability logs recorded a combined total of 74 LLM service calls across the three evaluation sessions, covering all four endpoint types used in the augmented panel: synthesis evaluation, snippet evaluation, group report generation, and group sources feedback. Schema validation was applied to all outputs as part of the post-processing pipeline described in Section 5.3.1. No schema validation failures were recorded across any of the 74 calls, indicating that all LLM outputs were delivered in a structurally valid form.

Source verification, which checks evidence excerpts against the student's actual synthesis text before the output is displayed, was embedded in the synthesis evaluation and snippet evaluation pipelines. The verification mechanism assigns a penalty score to excerpts that cannot be matched against the student text and flags these in the output metadata. Across the calls logged during the evaluation sessions, no verified citation failures were recorded that would have resulted in a flagged or suppressed output being presented to a participant. This is consistent with the participant-level finding reported in Section 6.3.1, where no participant detected a fabricated or hallucinated quotation during review.

One exception to the pattern of full output delivery was noted in the micro-survey data from T3, who indicated that one case included a missing AI evaluation for a specific main idea. The accompanying note clarified that this referred to an absence of generated assessment content for that idea within the structured response, rather than a fabricated or incorrect quotation. This type of omission is distinct from a hallucination failure and was not logged as a schema validation error; it likely reflects a case where the LLM output did not produce coverage assessment for that idea within the allocated response structure. The overall reliability of outputs across the evaluation sessions was therefore high, with no citation fabrication failures and one instance of evaluation omission across the full set of cases reviewed.

6.5.2 Performance, Latency, and Usability Implications

LLM service latency was computed from Langfuse trace records for the 74 calls attributed to the three evaluation sessions. Across all sessions, the mean response time was 26.5 seconds, and the median was 20.4 seconds. The 90th percentile response time was 52.4 seconds, and the maximum observed response time was 86.5 seconds. A total of 55 calls, representing 74.3 percent of the session total, were completed within 30 seconds. Five calls, 6.8 percent of the total, exceeded 60 seconds, and no calls exceeded 120 seconds. Figure 6.1 illustrates all these stats visually.

Per-session latency profiles differed notably across participants. T1's session showed the lowest mean latency at 20.3 seconds, with no calls exceeding 60 seconds. T2's session had a mean of 34.4 seconds and included three calls exceeding 60 seconds, with a maximum of 72.2 seconds. T3's session showed a mean of 33.8 seconds, with two calls exceeding 60 seconds and a maximum of 86.5 seconds. The higher latency in the T2 and T3 sessions is consistent with the caching model described in Section 5.2.3: T1's session followed an earlier period of pre-computation, meaning that more calls in that session were likely served from the cache, while some calls in T2 and T3 may have triggered fresh generation where cache entries were not available.

Mean input token count across all 74 calls was 28,477 tokens, and mean output token count was 2,144 tokens, giving a mean output-to-input ratio of 0.517. This ratio reflects the structured and constrained nature of the output format, where each response is expected to deliver a fixed schema of assessments rather than open-ended text.

In terms of user impact, T3 rated response time disruption at 3 out of 5 for the first two augmented cases, indicating that noticeable latency was experienced in those cases. For all other cases across

all three participants, response time disruption was rated at 1 or 2 out of 5. T3's elevated rating in the first two cases corresponds to that session's above-average latency profile and likely reflects cases where generation was triggered fresh rather than served from the cache. The absence of latency complaints in the qualitative data for T1 and T2, and the return to low disruption ratings for T3 in Case 3, suggest that once the cache was populated, response times did not substantially disrupt the review workflow. The operational picture is therefore one of acceptable latency under the caching model, with occasional slower responses when fresh generation is required.

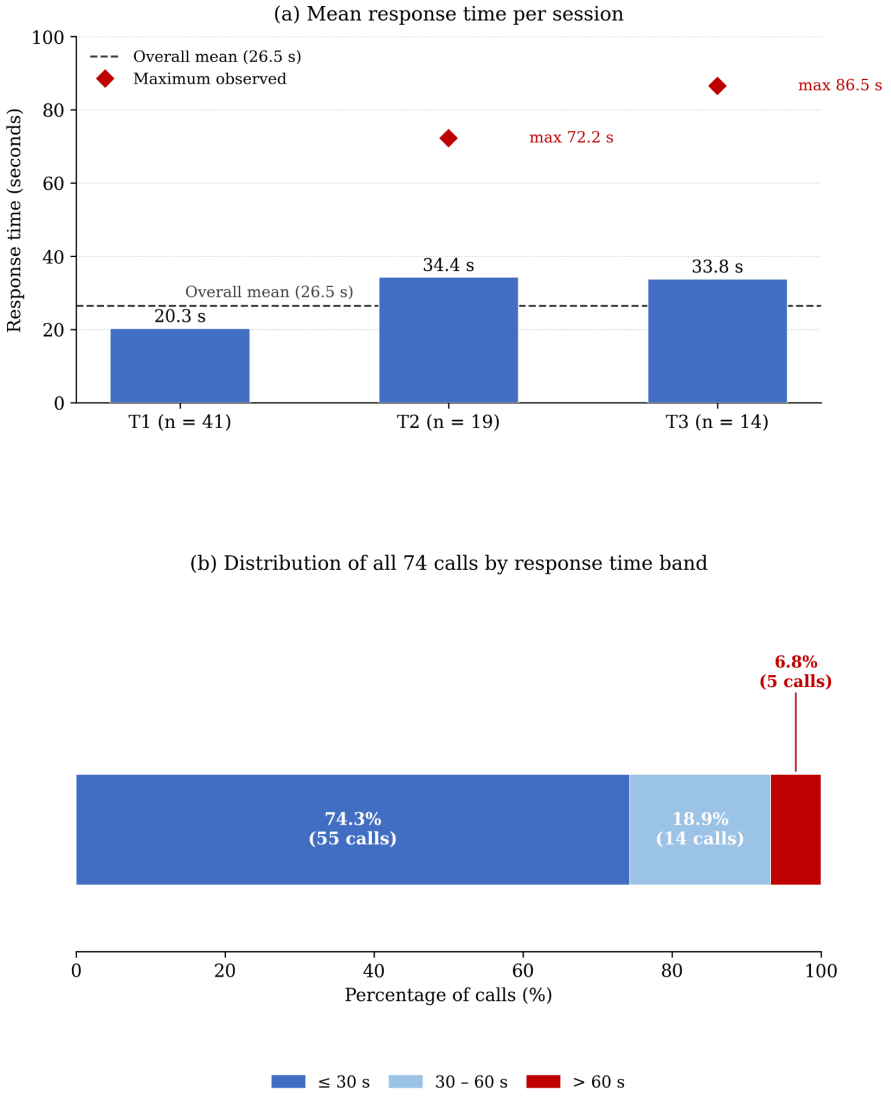

Figure 6.1. LLM service response time across evaluation sessions. (a) Mean response time per session with overall mean reference line and maximum observed values. (b) Distribution of all 74 calls across three latency bands. Higher mean latency in T2 and T3 is consistent with reduced cache availability relative to T1.

6.5.3 Teacher Trust and Concerns

All three participants reached a consistent position on the question of trust in the augmented panel: the system was appropriate and useful as a review support tool, but its outputs should not substitute for independent teacher judgment. This position was arrived at through different routes by each participant, reflecting their different relationships to the content domain and to AI-generated content more generally.

T1 maintained an explicitly verification-oriented stance throughout the session, checking specific AI claims against the student text and noting instances where the AI had not flagged ideas that T1 considered missing. The trust extended to the panel was conditional on the availability of verifiable evidence for each AI claim, and T1 was willing to accept coverage assessments where the evidence link was clear while remaining sceptical of summary-level judgments that lacked visible grounding.

T2 expressed trust in the panel's outputs for the cases reviewed while explicitly noting the theoretical possibility of errors that a domain expert might not easily detect. The trust framework T2 articulated was one of provisional acceptance subject to professional verification rather than either full trust or blanket scepticism: the panel functioned as an aid to judgment rather than as a source of final verdicts, and teacher verification remained an essential companion to any AI-generated assessment.

T3 showed the highest baseline trust in the panel and accepted AI-generated coverage indicators as a reliable starting point across all cases reviewed. However, T3 identified a specific concern about the system's failure to distinguish between irrelevant sources and pseudoscientific or fake sources in the panel display, as discussed in Section 6.3.2. This distinction was described as important for source literacy instruction: a student who used an irrelevant-but-credible source and a student who used a fake source present different pedagogical concerns, and the panel did not surface this difference clearly enough for T3 to act on it directly from the panel alone.

On the specific question of citation integrity, all three participants either explicitly confirmed that quoted passages were accurate or did not encounter any instance of fabrication during the review sessions. Concerns about citation integrity were expressed as anticipatory cautions rather than as reactions to observed errors: participants were aware that language models can produce inaccurate quotations and adjusted their trust accordingly, but did not encounter this problem in practice during the evaluation. This finding is consistent with the source verification mechanism described in

Section 5.3.1, though participants were not informed of that mechanism's existence during the session, and their trust calibration was therefore based on their own direct checking rather than on knowledge of the system's internal safeguards.

Language quality concerns were raised by two participants. T1 observed occasional typographic irregularities and instances of language mixing in the Finnish-language AI outputs. T3 independently noted the same issue during the interview:

> *"Kieli. Paasiassa OK. Ehkä ne olivat eniten siellä AI-yhteenvedon kohdalla sellaista, mitä kyllä kehittäisin. Että siellä oli ehkä vähän joissakin kohdin englannista suoraan suomeksi muokattuja sanoja, jotka eivät sellaisenaan ole suomen kieltä." (T3, originally in Finnish)*

> *["Language -- mainly fine. Perhaps the most [room for improvement] was in the AI summary section, which I would develop further. There were some words adapted directly from English into Finnish, which are not Finnish as such."]*

These observations, consistent across two participants, point to multilingual output quality as a recurring concern for deployment readiness in Finnish-medium instructional settings. This finding is consistent with the multilingual prompt engineering challenges described in Section 5.4.3.

6.6 Thematic Analysis of Think-Aloud Protocols

To complement the survey-based findings reported in Sections 6.2 through 6.5, a thematic analysis was conducted across the three think-aloud transcripts. The transcripts were read in full and inductively coded into recurring discourse topics, defined as coherent episodes in which a participant sustained attention on a specific aspect of the panel or the student work. Ten themes were identified across all three sessions. Figure 6.2 shows the number of coded episodes per theme for each participant.

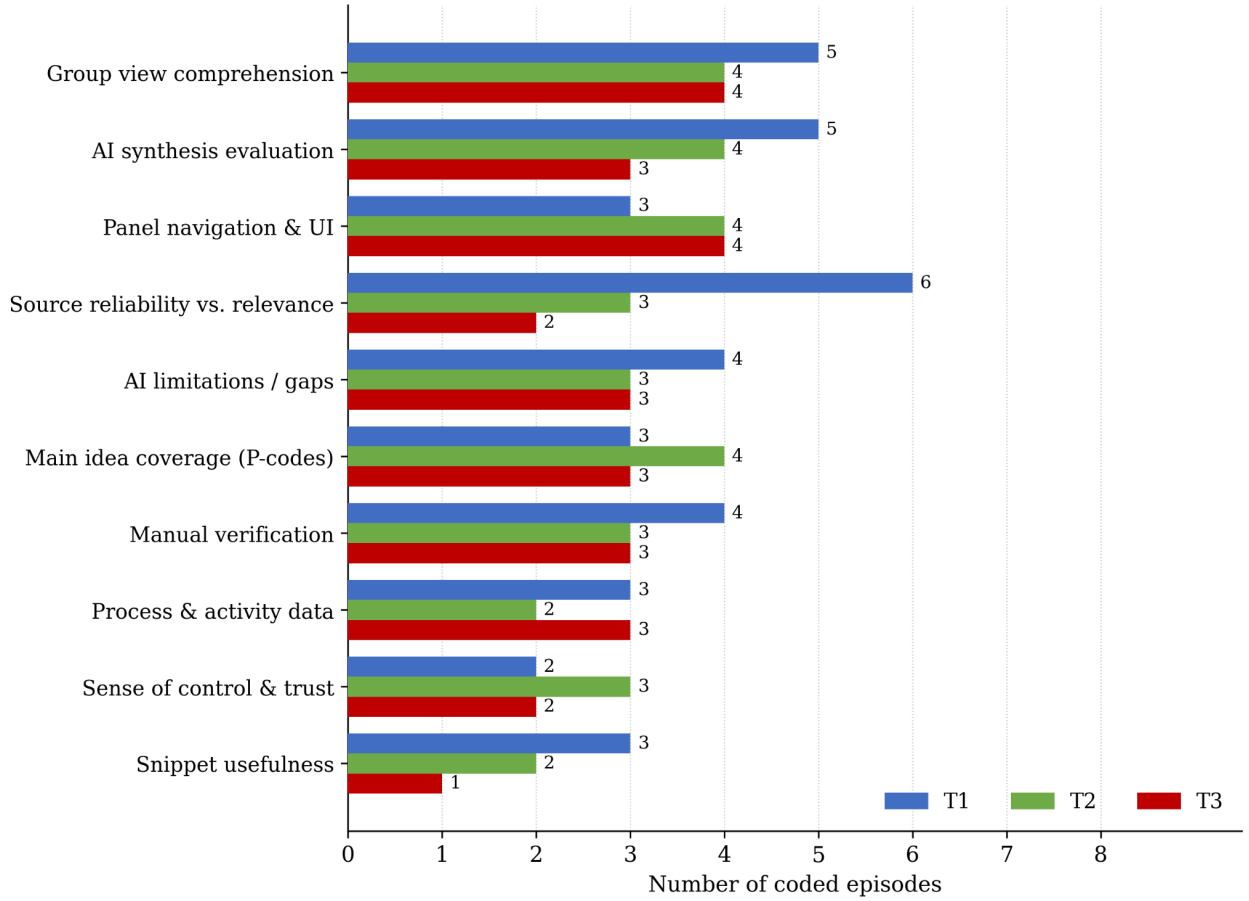

Figure 6.2. Thematic coding of think-aloud protocols by participant. Bars indicate the number of distinct coded episodes per theme. T1 = biodiversity specialist; T2 = educational researcher; T3 = school teacher.

6.6.1 Dominant Themes across Participants

The most frequently coded theme across all three participants was group view comprehension (13 episodes total), reflecting the consistent difficulty teachers experienced when attempting to read the class-level aggregation display. All three participants spent substantial time questioning the colour coding, the percentage figures, and the relationship between the group AI report and the P-code coverage breakdown. T2 articulated the difficulty clearly:

> *"The individual student view worked quite nicely, and it was easy to interpret and understand. But I think there were more issues concerning the group overview. Many things didn't make sense to me, at least based on this very short lens, like the colors and numbers and percentages." (T2)*

AI synthesis evaluation was the second most frequent theme (12 episodes), appearing consistently as teachers read and reacted to the LLM-generated synthesis feedback for each student case. Participants generally accepted the AI summary as a useful first orientation, but cross-checked it against the student text before concluding. T1, drawing on deep domain knowledge, engaged most critically with the synthesis content, identifying gaps the AI had not flagged.

Source reliability versus relevance was the third most coded theme (11 episodes), and was raised most prominently by T1, whose domain expertise allowed immediate detection of cases where students had rated irrelevant or fake sources as highly reliable. T1 noted this as a significant limitation of the current panel:

> *"I would like the AI to point out if there is irrelevant or even fake information included." (T1)*

Panel navigation and UI interactions constituted the fourth most common theme (11 episodes), reflecting the cognitive effort required to move between the Overview and Process tabs, interpret unfamiliar interface elements, and reconcile information presented across different parts of the panel. T2 and T3 raised this equally, with T3 spending notable time working out the colour logic of the P-code coverage display before being able to use it productively.

6.6.2 Themes reflecting the verification-first stance

Three themes collectively reflect the verification-first pattern identified in Section 7.1 as the central interpretive frame of the study. Manual verification (10 episodes) captures the repeated instances in which participants explicitly described checking AI claims against the student text or their own knowledge before accepting them. This was observed across all three participants and was not prompted by the session protocol. T1 summarised the stance succinctly: *"I think the AI report is very good if I can count on it - but we have to check manually things."*

Main idea coverage, or P-code engagement, appeared in 10 episodes, as participants worked through the coverage indicator to understand which ideas a student had addressed and which were missing. T2 found this display the most actionable element of the panel: *"The AI evaluation and the P-code breakdown - maybe those are the most crucial things from my point of view."* T3 initially needed the interface explained before the colour logic became clear, after which engagement increased.

AI limitations and gaps were also coded in 10 episodes, reflecting moments where participants identified something the AI had failed to detect or had misrepresented. The most consistent gap across all three participants was the absence of a distinction between irrelevant and fake sources, which T1 raised across multiple cases. Group-level AI report gaps - particularly the underrepresentation of the insect-decline perspective in the class summary - were noted by both T1 and T3.

6.6.3 Lower-frequency Themes and Participant Variation

Process and activity data (8 episodes) were referenced when participants examined student timelines, search counts, and saving behaviour. T3 expressed the strongest positive response to this information, noting that access to student process data is typically unavailable in ordinary teaching contexts. T2 found the milestone and timeline data less interpretable and ultimately did not use it to revise any evaluation judgments.

Sense of control and trust was coded in 7 episodes, capturing moments where participants reflected on their overall confidence in the panel's outputs. T2 expressed the most explicit trust calibration: *"I felt quite much in control. And I think these were the same that I manually checked."* T1 was more guarded, noting that trust was conditional on domain knowledge sufficient to verify individual AI claims. T3 expressed trust at the level of the overall summary while remaining uncertain about specific coverage scores.

Snippet usefulness was the least frequently coded theme (6 episodes). T1 expressed the most sustained scepticism, arguing that predefined snippets are poorly suited to broad, societal topics where relevant content cannot be captured in three fixed sentences per source: "*I don't think the snippets are that useful because here we can see that the student has captured some snippets that are not in green color, but they are still quite good information."* T2 rated snippets as moderately useful as contextual support. T3 engaged with the theme only briefly.

Taken together, the thematic distribution shown in Figure 6.2 reveals that participants' attention during the think-aloud sessions was concentrated on four areas: understanding the group-level display, evaluating the AI synthesis feedback, verifying source classification decisions, and navigating the interface. The themes with the highest episode counts align closely with the survey findings reported in earlier sections, providing convergent evidence that these dimensions -

particularly group view clarity and AI traceability - represent the most consequential design targets for future iterations of the augmented teacher panel.

6.7 Summary of Results

This section provides a brief synthesis of findings across the four research questions, drawing on the evidence reported in Sections 6.2 through 6.5.

On RQ1, the augmented panel supported quick case understanding primarily through two elements: the overall AI synthesis evaluation and the main idea coverage table. Both were identified by all three participants as the most useful features for forming a rapid first impression of a student case, and both were accessed in every augmented case observed in the video interaction logs. T3, drawing on her experience as a practising school teacher, additionally highlighted the process view as an element that provided rapid orientation to a student's task behaviour. Evidence previews, which link coverage assessments to specific passages in the student's text, functioned as the key transparency mechanism at the individual student level. Clarity at the group level was substantially lower: all three participants found the aggregated percentages and cohort-level figures unintuitive and difficult to interpret without understanding the underlying computation logic, while the narrative AI group report received more positive responses.

On RQ2, the AI evidence previews showed partial to full alignment with teacher-identified content across the three participants, with the degree of alignment correlated with participants' domain familiarity. No fabricated or hallucinated quotations were detected during the study sessions, and the Langfuse logs recorded no schema validation failures across 74 service calls. The primary mismatch patterns concerned the distinction between surface coverage and depth of engagement with an idea, the conflation of irrelevant and pseudoscientific sources in the panel display, and the relative weight given to snippet-level versus synthesis-level evidence. All three participants judged the previews as adequate for an initial review orientation while maintaining that manual verification remained necessary for substantive pedagogical judgments.

On RQ3, overall usefulness ratings were 4 to 5 out of 5 across all participants, and all three rated faster first insight at 4 or 5 out of 5 in the block comparison. The efficiency picture was nonetheless more nuanced: the augmented panel reduced the effort of reading and interpreting the synthesis text, while adding a navigation layer between panel tabs and additional steps for verifying individual evidence cards. The benefit of the panel was most pronounced for T3, whose lower prior domain

familiarity meant the AI summary provided the largest relative gain in orientation, and who described the process data as immediately usable in its current form. Video interaction data confirmed that first evaluative comments in the augmented condition consistently followed within the first one to two minutes of opening the panel.

On RQ4, the LLM service operated without schema validation failures across all 74 session calls, and no citation fabrication was observed. Mean latency across sessions was 26.5 seconds, with 74.3 percent of calls completing within 30 seconds. Latency was perceptibly higher in the T2 and T3 sessions, where fresh generation was required, and T3 rated response time disruption at 3 out of 5 for the first two cases before ratings returned to low levels as the cache was populated. One instance of a missing evaluation for a specific main idea was noted by T3, representing an omission rather than a hallucination failure. Language quality concerns in Finnish-language AI outputs were noted by both T1 and T3. The technical feasibility finding is that the system operated reliably under evaluation conditions, with the caching model providing acceptable latency for most calls, while specific improvements to multilingual output quality, the fake-versus-irrelevant source distinction in the panel display, and the transparency of group-level aggregation logic remain needed before deployment in a regular instructional setting.

7 Discussion

This chapter interprets the findings reported in Chapter 6 in relation to the research questions, prior research, and the broader context of AI-assisted teacher support in educational settings. Section 7.1 presents the main findings through an overarching interpretive frame. Section 7.2 situates these findings within the existing literature. Sections 7.3 and 7.4 draw implications for the design of teacher-facing AI support and educational dashboards, respectively. Section 7.5 discusses limitations and threats to validity, and Section 7.6 proposes directions for future work.

7.1 Main Findings in Relation to the Research Questions

To recap, the study is structured around four questions:

- **RQ1 (Transparency).** What process summaries and evidence previews help teachers form a quick, confident understanding of a student's synthesis process?
- **RQ2 (Alignment).** To what extent do model-produced evidence previews overlap with teacher-identified passages or ideas in the provided sources?
- **RQ3 (Usefulness & Efficiency).** Does the augmented panel reduce time-to-first-insight and the number of clicks required to reach key information, compared to the unaugmented panel, and how do teachers rate the clarity and perceived effort involved?
- **RQ4 (Technical feasibility & integrity).** Whether the suggestion and evidence service can meet the latency and reliability constraints of a classroom setting, and whether it can prevent fabricated or unverifiable citations?

These four research questions in this thesis address distinct aspects of an AI-augmented teacher panel: what supported transparency of understanding, how well AI outputs aligned with teacher judgments, whether the panel was useful and efficient, and whether the underlying service was technically feasible. Taken individually, each question yields findings of practical relevance. Read together, however, they reveal a consistent and interpretively coherent pattern: across all four dimensions, the teachers in this study engaged with the AI support not as a trusted authority to be consulted, but as a provisional source of evidence to be verified. This verification-first stance, observed consistently across participants with different backgrounds and domain expertise, is argued here to be the central finding of the study and the appropriate frame for interpreting what the augmented panel did and did not achieve.

The teacher-in-the-loop principle, as a design intention, asserts that AI tools in educational contexts should support rather than replace professional teacher judgment [22], [24]. What this study contributes is evidence of what that principle looks like in practice: it is not a design choice that teachers passively accept, but a behavioural pattern that teachers actively enact. Participants did not wait to be told that AI outputs required verification; they sought verification unprompted, checking evidence excerpts against the student text, noting missing ideas the AI had not flagged, and rating their sense of control against the availability of inspectable reasoning. The implication is that designing for the teacher-in-the-loop is not a matter of adding a disclaimer or a manual override button; it requires building the entire information architecture of the tool around the assumption that teachers will and should verify.

Interpreting RQ1 through this frame, the preference for the AI synthesis evaluation and the main idea coverage indicator over other panel elements reflects not simply that these features were more informative, but that they were the most efficient starting points for a verification sequence. The synthesis evaluation gave an immediate first impression of the case, and the coverage indicator communicated which ideas to check. Together, they allowed teachers to begin their own assessment quickly, with the AI output functioning as an initial hypothesis rather than a final verdict. The evidence preview, which links each coverage claim to a specific passage in the student text, was the mechanism that made this verification possible. Its value was not that it proved the AI was right; it was that it allowed the teacher to judge for themselves. Where this link was clear, adequacy judgments were higher. Where it was absent or opaque, as in the group-level aggregation view, teachers lost confidence not in the AI but in their own ability to use the AI output responsibly.

The RQ2 finding that alignment was partial-to-full, mediated by domain familiarity, similarly reflects the verification-first stance. T1, with moderate domain knowledge, reported the lowest alignment ratings and also the lowest sense of control. This is not a failure of the AI to be accurate; it reflects that a teacher with limited domain expertise cannot easily verify AI claims and therefore cannot distinguish between a correct AI assessment and an inaccurate one. The adequacy of AI support was bounded not by the AI's performance but by the teacher's capacity to verify it. This is a consequential finding for deployment: an AI tool designed for teacher review may provide the greatest value precisely where the teacher has least domain expertise to perform independent verification, yet this is also where the risk of over-reliance is highest. Designing the verification affordances of such tools with this asymmetry in mind is important.

The RQ3 efficiency findings are consistent with the same frame. The panel reduced the effort of initial case orientation, but it added an interpretive and navigational overhead associated with reading AI outputs, moving between tabs, and checking evidence excerpts. For T3, who had the lowest prior domain familiarity, the AI summary provided the largest relative gain in orientation because it compensated for the absence of independent domain knowledge; the verification overhead was lower because T3 had less independent knowledge to compare against. For T1 and T2, who brought stronger domain knowledge to the task, the AI output was one input among several, and the cost of integrating it into an already-active assessment process was more apparent. The net efficiency experience was therefore shaped by what the teacher brought to the review, not simply by the quality of the tool.

The RQ4 technical findings confirm that the system met the functional preconditions for the verification-first stance to operate: no fabricated citations were encountered, schema validation prevented malformed outputs, and latency was generally acceptable. These findings matter because a verification-first teacher cannot verify an output that contains hallucinated text, and a teacher who encounters fabricated evidence may lose trust in the entire panel rather than only in the specific failed output. The absence of hallucination failures is therefore a meaningful reliability result, even if it does not constitute a strong positive quality claim beyond the scope of the evaluation sessions. The language quality concerns raised by two participants and the single instance of a missing evaluation point to a gap between functional reliability (the system runs without failures) and instructional reliability (every output is of sufficient quality for teacher use). Closing this gap is a prerequisite for any deployment beyond a research evaluation.

7.2 Comparison with Prior Research

The student task studied in this thesis, a controlled multi-source synthesis requiring students to identify relevant content, evaluate sources that include fake or misleading texts, and produce a short written synthesis, is an instance of the multiple-document comprehension tasks that have been extensively studied in reading and literacy research [1], [2], [3], [4]. This literature establishes that multi-source comprehension is cognitively demanding even for skilled adult readers, requiring the construction and integration of multiple document representations alongside source evaluation judgments [1], [3]. The difficulty of the task for student teachers has been documented specifically in the LearnNet context: Heikkilä et al. [26] found that student teachers struggled to integrate different text perspectives into coherent syntheses, and Heikkilä et al. [31] found that approximately

half of the participants incorporated fake source content in their syntheses despite encountering the same relevant texts as their peers. These findings directly motivate the teacher review challenge addressed in this thesis: when student syntheses regularly contain undetected source errors or integration failures, the teacher reviewing a cohort of 100 or more submissions faces a task that is both cognitively demanding per case and practically unmanageable at scale without support.

The work of Vidbäck, Iskala, and Mikkilä-Erdmann [27] provides complementary context from the teacher perspective. Their study found that teachers recognized the importance of information literacy and source evaluation as curricular goals but reported a lack of clear pedagogical strategies and tools to support these skills effectively. The augmented panel studied in this thesis does not teach source literacy directly, but it offers a form of post-hoc diagnostic support: it helps teachers identify where source literacy breakdowns occurred in individual student work, which is a prerequisite for responsive instruction. The finding that the fake-versus-irrelevant distinction was absent from the panel is therefore not merely a display limitation; it is a gap in the diagnostic value of the tool for exactly the pedagogical purpose that teachers in Vidbäck et al.'s study identified as most challenging.

The evidence-linking approach used in the augmented panel, in which each coverage assessment is anchored to a specific passage from the student's text, has a structural parallel in the document mapping scaffolds studied by Barzilai et al. [6], [7]. Document mapping helps students construct integrated mental models of multiple sources by making source relationships visible. The augmented panel applies a structurally analogous approach on the teacher side: the evidence preview makes the AI's reasoning relationship to the student text visible, allowing the teacher to construct a model of what the student engaged with and how. The key difference is that in Barzilai et al.'s work, the mapping is performed by the learner as a learning activity, whereas in this thesis, the mapping is generated by the AI and presented to the teacher as an assessment support tool. The finding that the evidence preview was the critical transparency mechanism for teacher trust is consistent with the general principle that making source relationships explicit supports more reliable reasoning about multi-document content.

The process indicators presented in the augmented panel draw directly on the approach developed by Erdmann, Mikkilä-Erdmann, and Rautio [9], who demonstrated that log data from controlled multi-source learning environments can be used to create meaningful indicators of reading and synthesis behaviours. The thesis system extends this work in a specific direction: rather than using

process indicators for retrospective research analysis, it makes them available to the reviewing teacher in real time. This extension is not trivial, because the audience and purpose of the indicators change when they are presented to a teacher rather than to a researcher. Teachers need indicators that support rapid judgment rather than exhaustive analysis, and the design choices described in Section 5.2, particularly the milestone timeline and the at-a-glance statistics, reflect an attempt to adapt research-oriented process metrics to a professional review context. The finding that the process view was consistently among the first panel elements accessed, particularly by T3, suggests that this adaptation was at least partially successful.

The learning analytics dashboard literature provides a broader frame for interpreting the panel design and its evaluation results. Prior work on teacher-facing dashboards has identified several recurring design tensions: between information richness and interpretability [13], [16], between data granularity and actionability [12], [14], and between automated insight generation and teacher agency [15]. The findings of this thesis engage directly with each of these tensions. The preference for narrative AI reports over quantitative aggregation tables at the group level is a concrete instance of the interpretability-over-richness preference documented in prior dashboard research: teachers did not want more data at the class level; they wanted more comprehensible data. The group-level opacity finding, in which participants could not identify what the displayed percentages measured or how they had been computed, echoes a concern that has been raised repeatedly in learning analytics research: metrics presented without computation transparency do not support professional decision-making even when the underlying data is technically accurate [13], [16].

The traceability finding connects directly to the literature on explainable AI in education. Khosravi et al. [22] argue that explainability is a prerequisite for teacher trust in AI tools, and that systems which present outputs without accessible reasoning chains undermine the conditions for appropriate trust calibration. Darvishi et al. [23] demonstrate, in the context of peer assessment, how AI and analytics mechanisms - including instructor spot-checking tools - can be designed to improve trustworthiness and support teacher oversight of automated assessment processes.. The findings of this thesis provide empirical evidence from a practitioner evaluation context that supports these theoretical arguments: the visible link between an AI claim and its evidentiary basis in the student text was not a supplementary feature but the primary mechanism through which teachers calibrated their trust. This finding also extends the theoretical argument by specifying what explainability

means in a synthesis review context: it is not simply that the AI's algorithm be described, but that each AI claim be traceable to the specific student artefact that grounds it.

The log-data analysis approach used to derive usage patterns in Section 6.2.3 is consistent with the methodology reviewed by Wang [5], who found that log data from open-ended learning environments is increasingly used to study learner behaviours and support adaptive feedback. This thesis extends that analytical approach to the teacher-facing context: rather than using log data to study student behaviour, it uses log data and video interaction records to analyse teacher review behaviour. The insight that all participants accessed the AI synthesis evaluation in every augmented case, while bookmark and snippet views were consulted selectively, is an example of the kind of behavioural pattern that log-based interaction analysis can surface and that self-report data alone cannot.

On the question of AI ethics, Adams [24] identifies transparency, accountability, and pedagogical appropriateness as core principles for AI in education, alongside newer principles specific to institutional contexts such as teacher well-being and human oversight. The EU AI Act [25] places AI systems that influence educational assessment in a high-risk category, requiring meaningful human oversight before consequential decisions are made. The read-only, evidence-linked design of the augmented panel reflects these requirements in practice: the system does not assign grades, does not generate records that affect students without teacher mediation, and is explicitly positioned as a support for teacher judgment rather than a replacement for it. The verification-first adoption pattern observed in the evaluation suggests that teachers naturally maintain the kind of human oversight that regulatory frameworks are beginning to require, provided the tool is designed to make that oversight possible.

7.3 Implications for Teacher-facing AI Support

The findings of this study carry several implications for the design of AI support tools intended for teacher use in educational review contexts. These implications are drawn from specific findings rather than from general principles, and they are offered with the caveat that the evaluation involved three participants in a single task context; larger and more varied studies are needed before these can be treated as general design requirements.

The most direct implication concerns the primacy of verification affordances. The finding that traceability, rather than accuracy alone, determined whether teachers judged AI outputs as adequate

suggests that future teacher-facing AI tools should be designed from the outset around the assumption that teachers will and should verify. This means making every AI claim linkable to the student artefact that supports it, presenting evidence excerpts as a standard feature rather than an optional detail, and organizing the tool's information architecture so that verification is the natural next step after receiving an AI assessment. A tool that presents AI coverage claims without evidence links places the teacher in the position of either accepting the AI on faith or re-reading the entire student text to check it, neither of which supports the efficient and trustworthy review that the tool is designed to enable.

A related implication concerns the design of teacher control mechanisms. None of the three participants expressed frustration with the read-only nature of the panel, and the absence of a formal override or correction mechanism was not identified as a limitation. This suggests that, at least at the stage of initial review support, teachers do not need the ability to formally correct AI outputs; they need the ability to inspect and mentally revise them. The implication is that teacher agency in AI-supported review is better served by strengthening the visibility of AI reasoning than by providing formal control mechanisms. This is not to say that correction mechanisms have no value, but that they address a different stage of the review workflow, one that may matter more if AI outputs are used in formal assessment records.

The domain familiarity finding carries a deployment implication that is often overlooked in educational AI design. If the efficiency benefit of AI-generated summaries is largest for teachers with the least prior knowledge of the content area, then the same interface deployed to teachers with varying domain expertise will produce unequal and possibly unpredictable outcomes. A teacher with high domain expertise may find that the AI summary adds little that they could not derive independently, while introducing an additional interpretation step. A teacher with low domain expertise may rely on the AI output as a primary source of orientation, which raises the risk of over-reliance precisely where independent verification is most difficult. Future systems might address this asymmetry by adapting the depth or framing of AI-generated content based on teacher expertise, though this would require additional design and evaluation work.

The source typology gap identified in this study has a specific implication for multi-source synthesis tasks that include deliberately unreliable sources. The current panel treated irrelevant sources and fake or pseudoscientific sources as equivalent in the source evaluation display, which obscured a pedagogically important distinction: a student who engaged with an

irrelevant-but-credible source made a judgment about task relevance, while a student who engaged with a fake source made a judgment about credibility. These are different competencies to diagnose and address. Future AI support for source-aware synthesis review should encode this distinction explicitly in the output schema, so that the source evaluation display can convey not only whether a source was used, but what category of source evaluation error, if any, its use represents.

At the class level, the finding that narrative group reports were more actionable than quantitative aggregation views suggests a different design model for class-level AI support than the data visualization approach that dominates current learning analytics dashboard design. Rather than investing primarily in aggregated metrics and visual charts, which proved difficult for teachers to interpret without knowing the computation logic behind them, future systems might invest in AI-generated narrative summaries of class-wide patterns. These summaries, if grounded in verifiable data and presented with appropriate epistemic caution, can communicate the kind of pattern information that teachers need for instructional planning in a form that is closer to natural professional reasoning.

7.4 Implications for Dashboard Design in Educational Settings

Beyond implications specific to teacher-facing AI, the findings of this study speak to broader questions of dashboard design for educational professionals. Several specific design lessons emerge from the evaluation evidence.

Information hierarchy should reflect the natural structure of professional inquiry. The video interaction data (logs) showed that all three participants began their augmented review by consulting either the AI synthesis evaluation or the main idea coverage table, treating these as entry points into the case, before moving to more granular information as needed. This pattern of starting at the summary level and drilling down selectively is consistent with how professionals manage complex information tasks in time-constrained settings [12], [16]. Dashboard designs that present all available data at the same level of prominence, without a clear summary layer that enables triage, impose a navigation cost that teachers must absorb before they can begin their assessment. The two-tab structure of the augmented panel, with the Overview tab providing synthesis-level information and the Process tab providing detailed behavioural data, partially reflected this hierarchy, though the interaction data suggests that participants developed their own navigation

strategies rather than following a fixed path. Future designs might make the hierarchical structure more explicit by surfacing a clearly labelled summary layer that is always the first point of contact.

Evidence anchoring should be treated as a core design requirement rather than an enhancement. The distinction between a bare coverage indicator, which communicates that a main idea appears to be present, and an evidence-linked indicator, which additionally shows the specific passage from the student text that supports this assessment, is the distinction between a data point and a reasoned claim. From a teacher's perspective, only the latter can be acted upon with confidence, because only the latter enables independent verification. Dashboard designs that present AI-generated assessments without source anchors, or that treat the source link as optional metadata, undercut the verification affordance that this study identified as the primary driver of teacher trust. The design recommendation is to treat every AI claim as requiring an evidence anchor by specification, not as an additional feature to be added if time permits.

Group-level displays require computation transparency. The consistent confusion about what the group-level percentages measured, which affected all three participants and was articulated most clearly by T3, is not primarily a visual design problem. It is a transparency problem: the displays presented a number without explaining what was counted, over which population, with what denominator, or against which threshold. Prior work in learning analytics has identified computation transparency as a recurring challenge in teacher-facing dashboards [13], [16], and this study confirms that challenge in an AI-augmented context. The design recommendation is to accompany any quantitative group-level indicator with a plain-language description of its basis, embedded in the display rather than placed in separate documentation, so that teachers can calibrate their interpretation of the number without having to consult external material.

The relative effectiveness of narrative and quantitative representations at the group level suggests a design principle that has broader relevance for educational dashboards. The AI narrative group report, despite its own limitations, was rated more useful for class-level instructional planning than the percentage-based coverage table, because it presented coherent pattern information in a form that aligned more directly with how teachers think about their class as a group. This preference is not simply a stylistic one; it reflects a deeper point about the relationship between data representation and professional reasoning. Quantitative representations require the teacher to perform an additional interpretive step to convert a number into an instructional implication. Narrative representations, when well-constructed, can perform that step in the output itself.

Dashboard designers working on class-level tools should consider whether the investment in generating and presenting quantitative aggregations produces more actionable insight than a well-constructed natural language summary of the same underlying data [12], [13].

Finally, the latency and caching findings carry a dashboard design implication that is specific to AI-augmented tools. T3's elevated response-time disruption ratings in the first two augmented cases, before the cache was populated, illustrate that the user experience of an AI-augmented dashboard is partly determined by the infrastructure model that supports it. A push-on-submit caching approach, in which AI outputs are generated when the student submits the task and stored for immediate retrieval during the review session, can substantially reduce the latency experienced by teachers. This model requires that generation happens ahead of the review session, which places constraints on the workflow, but it addresses a known usability problem that would be more difficult to solve through interface design alone. This was implemented partially; however, it is not mentioned as part of this thesis.

7.5 Limitations and Threats to Validity

The evaluation study has several limitations that affect the interpretation and transferability of its findings. These limitations are discussed here in terms of their analytical significance rather than simply as caveats, following the approach recommended for case study research [33].

The most significant limitation is the sample size of three participants [n=3]. The study was designed as a case study in a specific learning environment with a specific task, and the small sample was an explicit methodological choice consistent with that design [33], [36]. Within this design, the three participants provided rich, detailed data that supports confident claims about individual experience and about patterns that emerged consistently across all three. However, the sample does not support claims about frequency, distribution, or generalizability. Findings that emerged from only one or two participants, such as T1's low sense of control or T3's initial latency disruption, are best treated as observations that merit further investigation rather than as established patterns. The study is appropriately understood as generating hypotheses for larger-scale evaluation rather than confirming them.

A related limitation concerns the participant profile. Two of the three participants held researcher-practitioner roles and had regular familiarity with learning analytics contexts and data-rich interfaces. Their comfort with navigating complex digital environments and their prior

understanding of log-data concepts may have produced a more favourable evaluation of the panel than would be obtained from classroom teachers without this background. T3's experience as a practising school teacher provides partial mitigation, as her reactions to the panel were grounded in authentic pedagogical concerns rather than research familiarity. Nevertheless, the sample does not represent the range of teachers who would be the primary users of such a system in practice.

The single-task context is a further limitation. All evaluation sessions used the same biodiversity loss task, the same source set, and the same cohort of student submissions. The transparency, alignment, and efficiency findings may be specific to this task type, this content domain, and this student population. Tasks with different source structures, different levels of content complexity, or different expected synthesis lengths might produce different patterns of AI-teacher alignment or different interaction sequences. The generalizability of the findings to other multi-source synthesis tasks is an open empirical question that future work should address.

The within-subjects ordering of conditions introduces a confound that affects the interpretation of both alignment and efficiency findings. Because the manual review condition always preceded the augmented review condition, and because the first two augmented cases per participant were the same student submissions reviewed in the manual condition, participants had already formed a judgment about each case before engaging with the AI assessment of that same case. This familiarity is likely to have elevated alignment ratings, because teachers who had already identified the key features of a case were well-positioned to find that the AI's assessment corresponded to their prior judgment. It may also have affected the efficiency experience in both directions: familiarity with the case could have reduced the time needed for augmented review (inflating the apparent efficiency gain), or the dual task of reviewing the case and evaluating the panel could have increased cognitive load (deflating it).

The prescribed timing protocol, which allocated fixed durations of seven and thirteen minutes per case, creates a further interpretive constraint. Because review time was set by the session design rather than by participant choice, the duration data does not reflect how long teachers would naturally spend on a case in independent practice. The protocol was necessary for within-session comparability, but it means that the efficiency findings describe teacher experience within a controlled time allocation rather than in an authentic, self-paced review context. Real classroom review is typically asynchronous and interrupted, and it is not clear whether the efficiency and effort perceptions observed in this study would transfer to that context.

Think-aloud verbalization may have introduced a reactivity effect, causing participants to be more deliberate and explicit about their uncertainty and verification reasoning than they would be in silent independent review [36]. This is a known methodological limitation of concurrent think-aloud protocols. The risk is that the verification-first pattern documented in this study is at least partly an artefact of the think-aloud method, which prompts reflection, rather than a spontaneous feature of how teachers engage with AI support in practice. Retrospective interview and micro-survey data provide some mitigation, as these data sources are less directly shaped by the think-aloud instruction, but they cannot fully resolve this concern.

7.6 Directions for Future Work

Several directions for future research follow naturally from the findings and limitations described in this chapter.

The most immediate priority is a larger-scale evaluation with a broader and more representative participant sample. A study involving classroom teachers without learning analytics familiarity, across multiple content areas and educational levels, would establish whether the transparency, alignment, and efficiency findings observed here replicate in a more representative population and whether the domain familiarity effect holds across a wider range of expertise levels. Such a study would also provide the statistical power needed to move from the qualitative patterns identified here to quantitative estimates of effect size.

A second direction concerns the deployment context. The evaluation was conducted in a researcher-facilitated session with prescribed timing and active think-aloud. A study in which teachers use the augmented panel in their regular classroom practice, asynchronously, over multiple review sessions and without researcher presence, would provide evidence about long-term adoption patterns, trust calibration over repeated use, and the effect of AI-generated assessments on actual instructional decisions. The latter is perhaps the most important open question: whether the panel changes what teachers do for students, not only what they know about them.

The group-level display is the component of the current system with the clearest design gap, and it warrants targeted design and evaluation work. Possible directions include experimenting with computation-transparent percentage displays that explain what is being counted and over which population, developing alternative visual encodings for cohort coverage patterns that reduce the interpretive load, and testing whether AI-generated narrative summaries of class patterns, when

given more structural support in the generation process, can reliably produce the kind of actionable insight that teachers described wanting but finding absent from the current quantitative view.

The fake-versus-irrelevant source distinction is a second component that merits targeted design work. Extending the AI pipeline to distinguish between sources that are irrelevant to the task and sources that are unreliable or pseudoscientific, and surfacing this distinction in both individual and group-level displays, would address the most consistently identified gap in the current system and would align the tool more closely with the source literacy diagnostic function that teachers in this study and in prior research [27] identified as important.

The domain familiarity effect raises a longer-term question about whether teacher-facing AI support should be adaptive. If the system's value is systematically higher for teachers with less prior domain knowledge, and if over-reliance is a corresponding risk for the same population, then a version of the system that calibrates the depth of AI-generated explanation based on teacher expertise would address both the opportunity and the risk simultaneously. This would require a mechanism for estimating teacher domain expertise, either through explicit profiling or through behavioural inference from interaction patterns, and would introduce new design and ethical questions about how such adaptation should be disclosed and controlled.

Finally, the Langfuse observability data collected in this study, which documents LLM call patterns, latency distributions, and token usage across the evaluation sessions, represents a methodological resource for future research on AI service behaviour in educational deployments. Combining LLM observability logs with interaction telemetry and qualitative data from think-aloud sessions, as attempted in this thesis, offers a multi-layer approach to understanding how AI-augmented educational tools perform in practice that goes beyond either pure performance benchmarking or pure user experience evaluation. Developing this as a replicable methodological approach for evaluating AI tools in educational settings is itself a worthwhile direction for future work.

8 Conclusion

8.1 Summary of Thesis

This thesis studied the design and evaluation of an AI-augmented teacher panel intended to support the review of multi-source student synthesis tasks in the LearnNet learning environment. The problem motivating the work is that reviewing student syntheses produced in controlled multi-source settings is cognitively demanding: each synthesis reflects a student's choices about which sources to read, which to trust, and which ideas to incorporate, and the teacher reviewing the synthesis must reconstruct this decision trail, often without detailed process data, to make a meaningful pedagogical judgment. When student cohorts are large and syntheses are complex, this challenge is difficult to address through manual review alone.

Prior research in the LearnNet context had established that the multi-source synthesis task is genuinely demanding for student teachers: approximately half of participants in related studies incorporated fake source content in their syntheses [31], and students regularly struggled to integrate multiple text perspectives into coherent wholes [26]. Log data from the learning environment captures detailed traces of student reading and annotation behaviour that are potentially informative for teacher review [9], [28], but this data was not previously made accessible to teachers in a usable form.

The augmented panel developed in this thesis addressed this gap by combining two types of output: deterministic process indicators derived directly from student log data, and LLM-generated evidence assessments that link coverage judgments for predefined main ideas to specific passages in the student's synthesis and reading traces. The system was designed according to a design science research methodology [34], [35], with teacher transparency, evidence traceability, and reliability as explicit design goals. A post-processing pipeline including schema validation and source verification was implemented to prevent fabricated citations from reaching the teacher's view.

The evaluation was conducted as a case study [33] involving three university-affiliated participants, one of whom was a practising school teacher. Each participant completed a session covering a manual review condition and an augmented review condition, with think-aloud protocol, per-case micro-surveys, and a retrospective interview. LLM service performance data was collected through

Langfuse observability logs, and interaction patterns were extracted from video recordings of the sessions.

The main findings were as follows. On transparency, the AI synthesis evaluation and main idea coverage table were the most consistently relied-upon features, with the evidence preview functioning as the critical mechanism for teacher trust calibration. On alignment, AI evidence previews showed partial to full correspondence with teacher-identified content, with the degree of alignment mediated by participant domain familiarity; no fabricated citations were encountered during the evaluation sessions. On usefulness and efficiency, overall usefulness ratings were high (4 to 5 out of 5), while faster first insight was rated at 4 to 5 out of 5 across all participants; the efficiency benefit was most pronounced for the participant with the lowest prior domain familiarity. On technical feasibility, the LLM service operated without schema failures across 74 session calls, with a mean response time of 26.5 seconds and acceptable latency under the caching model, while multilingual output quality and source typology distinctions were identified as areas requiring improvement.

8.2 Contributions of the Work

The thesis contributes at three interrelated levels, which together constitute a combined contribution to the design, empirical study, and theoretical understanding of teacher-facing AI support in multi-source learning environments.

The first level of contribution is a design artifact: the augmented teacher panel itself, as a proof-of-concept implementation of evidence-linked AI support within an existing educational technology context. The panel demonstrates that student log data from a controlled multi-source environment can be combined with LLM-generated coverage assessments into a coherent and technically reliable teacher-facing view, delivered at latency levels that are acceptable for professional review use. The seven-layer processing architecture, the post-generation validation pipeline, and the push-on-submit caching model together constitute a set of implementable design decisions that could inform the development of similar systems in comparable educational contexts. The system is not a finished product, but it demonstrates technical feasibility and provides a concrete starting point for iteration.

The second level of contribution is empirical: three specific findings about teacher-AI interaction in a synthesis review context that are novel and practically significant. First, traceability - the visible

link between an AI assessment and the specific student text passage that grounds it - was found to be a stronger determinant of perceived adequacy than accuracy claims alone. Teachers calibrated their trust against the availability of inspectable evidence, not against the AI's confidence or overall performance. Second, the efficiency benefit of AI-generated summaries was found to be mediated by teacher domain familiarity: the system provided the largest relative gain in case orientation for the participant with the least prior knowledge of the content domain, and more qualified benefits for participants who could independently evaluate the synthesis content. Third, narrative AI group reports were consistently more interpretable for class-level instructional planning than quantitative aggregation displays, suggesting that natural language generation for class-level summaries may be a more effective design approach than visualization-centred aggregation for this use case.

The third level of contribution is design knowledge: the teacher-in-the-loop principle, documented here not as a design aspiration but as an empirically observed behavioural pattern. Across all three participants and all four research questions, teachers engaged with the AI support in a verification-first mode: accepting AI assessments provisionally, seeking evidence to check them, and calibrating their trust against the availability and clarity of that evidence. This pattern suggests that effective teacher-facing AI is not primarily a question of AI accuracy - it is a question of AI transparency and the design of verification affordances. The design knowledge contribution is the articulation of this principle as a transferable design constraint: tools for teacher review should be built around the assumption of conditional trust and designed to make verification easy, because this is how professional teachers engage with AI support in practice.

8.3 Final Remarks

The proliferation of AI-generated content in educational technology is raising a question that is both practical and ethical: how should teachers relate to AI outputs that bear on their professional judgments about students? One answer, reflected in much current product development, is that teachers should accept AI outputs as efficient and reliable approximations that reduce workload. Another answer, reflected in the ethical and regulatory literature on AI in education [24], [25], is that teachers should maintain meaningful oversight and remain the final decision-makers. This study does not adjudicate between these positions abstractly, but it offers evidence about what the second position looks like when instantiated in a concrete tool and evaluated with real practitioners.

What the evidence shows is that teachers do not need to be instructed to maintain oversight; they do it spontaneously, provided the tool is designed to make oversight possible. The participants in this study checked AI claims against student text, identified gaps that the AI had missed, and calibrated their trust against the clarity of the AI's reasoning chain, without being prompted to do so. This is a practically hopeful finding: it suggests that the teacher-in-the-loop principle is not merely a design aspiration or a regulatory requirement to be engineered around, but a disposition that professional teachers bring to AI-supported review when the tool is designed in a way that makes their professional judgment relevant.

The limits of the current system are equally instructive. The group-level aggregation view, the missing fake-versus-irrelevant distinction, and the multilingual output quality concerns are not peripheral issues; they are places where the tool did not yet support the teacher's professional judgment adequately. They are also places where future design work has clear targets. The contribution of a study like this is not only to demonstrate what is possible but to identify precisely where the gap between what is possible and what is needed remains, so that future work can close it.

Multi-source synthesis tasks ask students to do something that is genuinely difficult: to read critically from a set of sources that may mislead them, to evaluate and select, and to construct a coherent account of a complex topic. Teachers who review this work are doing something equally demanding: reconstructing a student's intellectual process from its written output, under time pressure, at scale. If AI support can make that teacher task more transparent, more efficient, and more reliably grounded in the student's actual work, without displacing the teacher's judgment or undermining their sense of professional agency, it will have served its purpose. This thesis is a step toward understanding how to design for that outcome.

References

[1] M. A. Britt and J.-F. Rouet, *"Multiple document comprehension,"* in Oxford Research Encyclopedia of Education, Oxford University Press, Apr. 2020.

[2] L. Primor and T. Katzir, *"Measuring multiple text integration: A review,"* Frontiers in Psychology, vol. 9, p. 2294, Nov. 2018.

[3] Ø. Anmarkrud, I. Bråten, and H. I. Strømsø, *"Multiple-documents literacy: Strategic processing, source awareness, and argumentation when reading multiple conflicting documents,"* Learning and Instruction, vol. 30, pp. 64–76, 2014.

[4] A. List and P. A. Alexander, *"Analyzing and integrating models of multiple text comprehension,"* Educational Psychologist, vol. 52, no. 3, pp. 143–147, 2017.

[5] Y. Wang, *"A systematic review of empirical studies using log data from open-ended learning environments,"* British Journal of Educational Technology, vol. 53, no. 2, pp. 333–351, 2022.

[6] S. Barzilai, D. Tal-Savir, F. Abed, S. Mor-Hagani, and A. R. Zohar, *"Mapping multiple documents: From constructing multiple document models to argumentative writing,"* Reading and Writing, vol. 36, pp. 809–847, 2023.

[7] S. Barzilai, S. Mor-Hagani, A. R. Zohar, T. Shlomi-Elooz, and R. Ben-Yishai, *"Making sources visible: Promoting multiple document literacy with digital epistemic scaffolds,"* Learning and Instruction, vol. 157, Art. 103980, 2020.

[8] M. T. McCrudden, I. Bråten, and L. Salmerón, *"Learning from multiple texts,"* in International Encyclopedia of Education, 4th ed., R. J. Tierney, F. Rizvi, and K. Ercikan, Eds. Amsterdam: Elsevier, 2022.

[9] N. Erdmann, *"Creating process indicators from learning traces,"* presented at Finnish Learning Analytics and Artificial Intelligence in Education Conference, Joensuu, Finland, 2024. Conference proceedings not publicly available]

[10] M. A. Chatti, A. Muslim, U. Schroeder, and M. Wosnitza, *"How to design effective learning analytics indicators? A human-centered indicator design approach,"* " in Proc. Workshop on

Human-Centered Learning Analytics, 15th European Conference on Technology Enhanced Learning (EC-TEL 2020), 2020.

[11] R. Scherer, S. Greiff, and J. Hautamäki, *"Exploring the relation between time on task and ability in complex problem solving,"* Intelligence, vol. 48, pp. 37–50, 2015

[12] R. A. Dourado, R. L. Rodrigues, N. Ferreira, R. F. Mello, A. S. Gomes, and K. Verbert, *"A teacher-facing learning analytics dashboard for process-oriented feedback in online learning,"* in Proc. 11th Int. Learn. Anal. Knowl. Conf. (LAK'21), New York: ACM, Apr. 2021, pp. 482–489

[13] K. Wiley, Y. Dimitriadis, and M. Linn, "*A human-centred learning analytics approach for developing contextually scalable K-12 teacher dashboards,"* British Journal of Educational Technology, 2023.

[14] R. Kaliisa and J. A. Dolonen, *"CADA: a teacher-facing learning analytics dashboard to foster teachers' awareness of students' participation and discourse patterns in online discussions,"* Technology, Knowledge and Learning, vol. 28, pp. 937–958, 2023.

[15] N. B. C. Nguyen, M. Lithander, C. Östlund, and T. Karunaratne, *"TEADASH: Implementing and evaluating a teacher-facing dashboard using design science research,"* Informatics, vol. 11, no. 3, Art. 61, Aug. 2024.

[16] B. Rienties, C. Herodotou, T. Olney, M. Schencks, and A. Boroowa, *"Making sense of learning analytics dashboards: A technology acceptance perspective of 95 teachers,"* International Review of Research in Open and Distributed Learning, vol. 19, no. 5, 2018.

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, *"Attention is all you need,"* in Advances in Neural Information Processing Systems, vol. 30, 2017, vol. 30, 2017

[18]T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "*Language models are few-shot learners*," in Advances in Neural Information Processing Systems, vol. 33, pp. 1877–1901, 2020.

[19] R. Bommasani, D. A. Hudson, E. Aditi, R. Altman, S. Arora, S. von Arx, et al., *"On the opportunities and risks of foundation models,"* Stanford Center for Research on Foundation Models, arXiv:2108.07258, 2021.

[20] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, Y. Du, C. Yang, Y. Chen, Z. Chen, J. Jiang, R. Ren, Y. Li, X. Tang, Z. Liu, P. Liu, J.-Y. Nie, and J.-R. Wen, *"A survey of large language models,"* arXiv:2303.18223, 2023.

[21] J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, E. H. Chi, T. Hashimoto, O. Vinyals, P. Liang, J. Dean, and W. Fedus, *"Emergent abilities of large language models,"* Transactions on Machine Learning Research, 2022.

[22] H. Khosravi, S. B. Shum, G. Chen, C. Conati, Y.-S. Tsai, J. Kay, S. Knight, R. Martinez-Maldonado, S. Sadiq, and D. Gašević, *"Explainable artificial intelligence in education,"* Computers and Education: Artificial Intelligence, vol. 3, Art. 100074, 2022..

[23] A. Darvishi, H. Khosravi, S. Sadiq, and D. Gašević, *"Incorporating AI and learning analytics to build trustworthy peer assessment systems,"* British Journal of Educational Technology, vol. 53, no. 4, pp. 844–875, 2022.

[24] S. Adams, *"Ethical principles for artificial intelligence in K-12 education,"* Computers and Education: Artificial Intelligence, vol. 4, Art. 100053, 2023.

[25] European Parliament and Council of the European Union, Regulation (EU) 2024/1689 of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act). Official Journal of the European Union, Aug. 2024.

[26] M. Heikkilä, M. Häkkinen, A. Vidbäck, M. Mikkilä-Erdmann, and I. E. Sääksjärvi, "*Student teachers' conceptual understanding of biodiversity loss in a multiple-source learning environment,"* Environmental Education Research, vol. 31, no. 7, pp. 1356–1370, 2025.

[27] A. Vidbäck, T. Iiskala, and M. Mikkilä-Erdmann, *"Teachers' experiences of the challenges of teaching science literacy in primary school,"* Ainedidaktiikka / Subject Didactics, vol. 7, no. 2, pp. 3–24, 2023

[28] N. Erdmann and M. Mikkilä-Erdmann, "Learning science supported by a digital learning environment," University of Turku, FINSCI Project Documentation, Parts 1–2, 2023. [Internal project document]

[29] M. Tapola, *"Using low-tech prototype to study children's preferences for UI components, case KidNet,"* M.S. thesis, Univ. Turku, Turku, Finland, Jul. 2023.

[30] K. Rautio, *"Design and implementation of a web application for practising internet reading skills,"* M.S. thesis, Univ. Turku, Turku, Finland, May 2022.

[31][31] M. Heikkilä, A. Vidbäck, M. Mikkilä-Erdmann, K. Rautio, and N. Erdmann, *"Student teachers' syntheses of knowledge on biodiversity loss from relevant, irrelevant, and fake sources,"* International Journal of Science Education, early online, 2026, doi: 10.1080/09500693.2026.2617917.

[32] D. J. Leu, C. K. Kinzer, J. Coiro, J. Castek, and L. A. Henry, *"New literacies: A dual-level theory of the changing nature of literacy, instruction, and assessment,"* Journal of Education, vol. 197, no. 2, pp. 1–18, 2017.

[33] R. K. Yin, Case Study Research: Design and Methods, 5th ed. Thousand Oaks, CA: SAGE Publications, 2014.

[34] A. R. Hevner, S. T. March, J. Park, and S. Ram, *"Design science in information systems research,"* MIS Quarterly, vol. 28, no. 1, pp. 75–105, 2004.

[35] K. Peffers, T. Tuunanen, M. A. Rothenberger, and S. Chatterjee, *"A design science research methodology for information systems research,"* Journal of Management Information Systems, vol. 24, no. 3, pp. 45–77, 2007.

[36] C. B. Seaman, *"Qualitative methods,"* in *Guide to Advanced Empirical Software Engineering,* F. Shull, J. Singer, and D. I. K. Sjøberg, Eds. London, UK: Springer, 2008, pp. 35–62.

[37] European Parliament and Council of the European Union, Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing

Directive 95/46/EC (General Data Protection Regulation), *Official Journal of the European Union,* L 119, pp. 1–88, Apr. 2016.

[38] M. Pedaste, M. Mäeots, L. A. Siiman, T. de Jong, S. A. N. van Riesen, E. T. Kamp, *et al.*, *"Phases of inquiry-based learning: Definitions and the inquiry cycle,"* Educational Research Review, vol. 14, pp. 47–61, 2015.

[39] J. Frerejean, J. L. H. van Strien, P. A. Kirschner, and S. Brand-Gruwel, *"Embedded instruction to learn information problem solving: Effects on secondary school students,"* Computers in Human Behavior, vol. 102, pp. 187–196, 2020.

UNIVERSITY OF TURKU

Appendix A

Semi-Structured Post-Session Interview Guide

Estimated duration: 10-15minutes.

Thank you for participating in today's session. I will now ask you some questions about your experience with both review conditions. There are no right or wrong answers; I am interested in your genuine reactions. The interview will be audio-recorded with your permission.

A. Overall impressions

- **a.** How would you describe your overall experience with the augmented panel compared to the manual review condition?
- **b.** Which parts of the panel did you find yourself using or returning to most? Did anything go unused?

B. Transparency and panel elements (RQ1)

a. How well did the AI synthesis evaluation reflect your own reading of the student's text?

b. Did the evidence previews — the quoted or paraphrased passages linked to each coverage judgment — affect how confident you felt in your assessment? In what way?

c. Were there any panel elements that felt unclear, redundant, or that you did not trust?

C. AI-teacher alignment (RQ2)

a. Looking at the sources, which passages or ideas would you expect a strong student synthesis to address?

b. When you compare those to the AI coverage assessments you saw during the session, how well did they correspond?

D. Usefulness and efficiency (RQ3)

a. Did reviewing with the augmented panel feel faster than without it? At which points in the review did you notice a difference?

b. How would you rate the cognitive effort of interpreting what the panel displays? Was anything effortful to make sense of?

E. Technical experience (RQ4)

a. Did you notice any loading delays or other technical issues during the augmented condition?
b. Did any AI output seem inaccurate, incomplete, or not supported by the student's text?

F. Closing

a. If you could change one thing about the panel design, what would it be?
b. Is there anything else about your experience today that you would like to add?