Exposía: Teaching and Assessment of Academic Writing Skills for Research Project Proposals and Peer Feedback

Dennis Zyska[1], Alla Rozovskaya[2], Ilia Kuznetsov[1], Iryna Gurevych[1]
[1]Ubiquitous Knowledge Processing Lab (UKP Lab), Department of Computer Science and Hessian Center for AI (hessian.AI), Technical University of Darmstadt
[2]Department of Computer Science at Queens College, City University of New York (CUNY)
[1]www.ukp.tu-darmstadt.de

Abstract

We present Exposía, the first public dataset[1] that connects writing and feedback in higher education, enabling research on educationally grounded computational approaches to teaching and evaluating academic writing. Exposía includes student research project proposals and peer and instructor feedback consisting of comments and free-text reviews. The dataset was collected in the "Introduction to Scientific Work" course of the Computer Science program that focuses on teaching *academic writing skills* and *providing peer feedback on academic writing*. Exposía reflects the multi-stage nature of the academic writing process that includes drafting, receiving feedback, and revising the writing based on the feedback received. Both the project proposals and peer feedback are accompanied by human assessment scores based on a fine-grained, pedagogically-grounded schema for writing and feedback assessment that we develop.

We use Exposía to benchmark state-of-the-art large language models (LLMs) on two tasks: automated scoring of (1) the proposals and (2) the student reviews. We find that the two tasks are best served by different LLMs. Furthermore, closed-source models consistently outperform open-weight models, motivating further research on improving the performance of open-weight models preferred in classroom settings. Finally, we establish that a prompting strategy that scores multiple aspects of the writing together is the most effective, paving the way for more effective classroom deployment of modern LLMs.

1 Introduction

Writing a research project proposal or exposé[2] is a key challenge in academic education and represents a pressing concern across the academic community (Muneer et al., 2020). It is also an essential skill required for conducting research and for achieving academic success (Kivunja, 2016; Grech, 2018). However, many students lack the necessary competencies to develop a research proposal and its core components, including identifying a research question, formulating appropriate research methodology, and defining hypotheses and methods to solve the problem (Rastri et al., 2023).

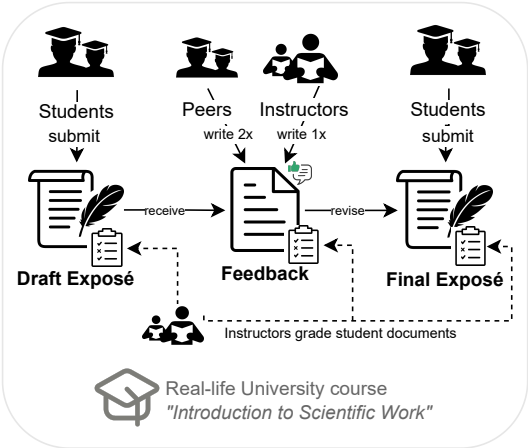


Figure 1: Overview of Exposía. Students submit a *draft exposé*, receive *feedback* from instructors and peers in the form of *comments* and free-text *reviews*, revise and submit a *final exposé*. Instructors grade draft and final exposés *and* student feedback.

At the same time, higher education faces another pressing challenge: the peer review crisis (Lu et al., 2025). Peer feedback is a powerful pedagogical tool that improves the writing performance of both authors and reviewers (Wu and Schunn, 2021; Jiang et al., 2024). Yet, students are rarely taught this skill in a classroom. Notably, research shows *a strong link* between academic writing skills and the ability to provide effective feedback (Huisman et al., 2019), underscoring the need for instructional approaches that *simultaneously support research project proposal writing and the development of effective peer review skills.*

[1]https://tudatalib.ulb.tu-darmstadt.de/handle/tudatalib/4978

[2]The term exposé is used in many universities to refer to a *research project proposal*, a structured document that outlines a planned research project. We will use *research project proposal*, *(research project) exposé* and exposé interchangeably.

Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing, pages 1076–1130
October 24-29, 2026 ©2026 Association for Computational Linguistics

Teaching students how to write a good exposé and how to provide high-quality feedback is very challenging. *Formative assessment*, which is used during the learning process, is a fundamental component of effective teaching (Black and Wiliam, 1998; Gass and Mackey, 2012). However, human evaluation of exposés and peer feedback requires substantial effort (Grabau and Wilson, 1995). This applies to *holistic scoring of exposés and feedback*, *fine-grained aspect-based assessment*, and especially to *the provision of textual feedback in the form of comments and reviews*. These limitations necessitate the development of *computational approaches* – not only for automated scoring of writing and feedback quality, but also for generating detailed text-level feedback on academic writing that evaluates and supports learning. In addition to the lack of appropriate benchmarks both for holistic and fine-grained scoring, providing textual feedback requires modeling complex cross-document relationships between proposal drafts, feedback, and resulting revisions. Curating such data is a time-consuming process. Finally, collecting data in a live course comes with ethical and legal challenges (see *Limitations*). As a result, existing resources for computational writing assistance and assessment are fragmented across subtasks such as automated essay scoring (Ke and Ng, 2019), peer feedback analysis (Staudinger et al., 2024), or revision modeling (Ruan et al., 2024).

We present Exposía, the first corpus that integrates drafting, feedback, revision, and fine-grained exposé and feedback assessment scores within a single dataset (Figure 1). Exposía is *continuously* collected in a newly developed university Computer Science course. The course focuses on teaching students (1) *how to write a research project exposé* and (2) *how to write peer feedback*. While this is an undergraduate Computer Science course, the skills it develops generalize across disciplines (Section 4) and are equally applicable to graduate students and other early-career researchers.

Exposía captures a complete research-proposal writing workflow. As illustrated in Figure 1, each student first produces an exposé **draft**. The draft receives (i) **peer feedback** from fellow students and (ii) **instructor feedback**, both provided as (a) **inline comments** anchored to draft passages and (b) a **free-text review**. The student revises the draft based on the feedback to produce the **final** exposé version. Figure 2 (top) details each component.

We further design fine-grained grading criteria that target the key aspects of exposé and feedback writing (Figure 2, bottom). These criteria are grounded in established research on exposé writing and feedback assessment and are tailored to the academic disciplinary context of Exposía. The exposé drafts, the final exposés, and the student feedback are graded by course instructors, who provide scores on each of the criteria.

To the best of our knowledge, Exposía is the first dataset, collected in an educational setting, that features this unique class of cross-document relations, *enabling new applications and tasks* for teaching the skills for writing research proposals and providing effective feedback (see Section 7). We further establish baseline models with LLMs for two tasks in this space: **exposé scoring** and **feedback scoring**, based on the fine-grained grading schema for writing and feedback assessment.

Contributions The main contribution of this work is the introduction of a novel problem definition for teaching and assessment of exposé and feedback writing skills in an educational setting. The result is the first *unique* public dataset of student research exposés and feedback, both accompanied by human assessment scores, based on the fine-grained grading schema that we develop. We present in-depth analyses of the dataset, and models for the corresponding tasks of exposé and feedback scoring, establishing baselines that the community can build on.

2 Related Work

We focus on exposé and feedback writing skills, which are distinct from other (academic) writing skills. Below, we discuss related research directions and the gaps in existing datasets.

Automated essay scoring (AES) AES (Page, 1966) is an established NLP task that focuses on writing assessment. However, AES benchmarks consist of persuasive essays in general-domain language, manually annotated with a holistic score, complemented by scores for a small set of assessment criteria (e.g., content, organization, language) (Mathias and Bhattacharyya, 2018; Gaudeau, 2025; Li and Ng, 2024; Xue et al., 2021; Crossley et al., 2023; Ishikawa, 2023). Importantly, these benchmarks do not include human feedback, revisions, or feedback assessment scores.

Peer feedback Because Exposía links each exposé draft with feedback text and assessment scores, it is related to scholarly peer-review datasets (Kang et al., 2018; Dycke et al., 2023). However, these datasets provide only coarse-grained annotations.

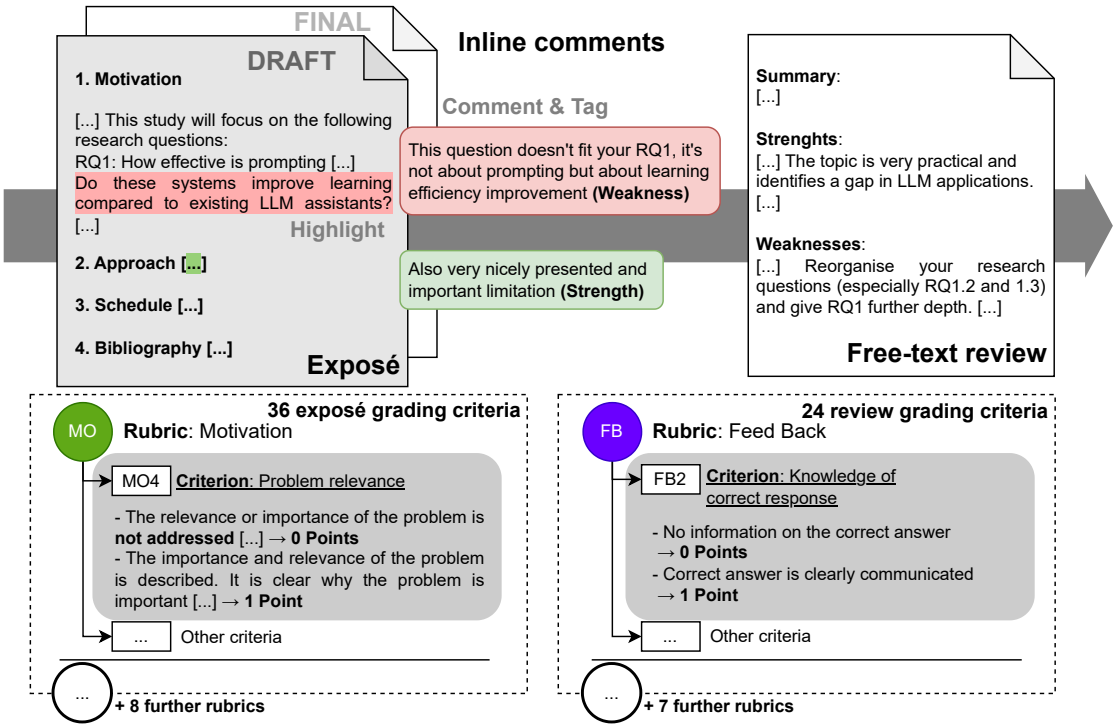


Figure 2: Example instance from Exposía. Top: A student **draft exposé** (top-left) is reviewed by an instructor or a student peer and receives **feedback**. The feedback consists of (1) passage-anchored **inline comments** (shown in the top middle in red and green), each associated with a **tag** (e.g., Weakness, Strength); and (2) a free-text **review** (top-right). Bottom: Exposés and reviews are graded by instructors and receive scores on multiple aspects (**criteria**) organized into high-level topics (**rubrics**). The **exposé grading criteria** and the **review grading criteria** are shown in the bottom left and bottom right of the figure, respectively.

Crucially, they serve fundamentally different purposes, as they focus on scientific communication in a professional decision-making context, whereas Exposía is situated in an educational environment with a focus on learning and assessment.

Connecting writing, feedback, and revisions A small number of resources connect student writing with peer or instructor feedback and/or revision (Zhang et al., 2017; Kashefi et al., 2022; Pilan et al., 2020; Rietsche et al., 2022), enabling analyses of how feedback relates to revision outcomes. However, existing datasets still focus on persuasive essays in general-domain language. Furthermore, they lack fine-grained annotations that allow for evaluating the impact of feedback on revisions and writing outcomes, and that are provided in Exposía: (i) passage-anchored inline comments tied to draft spans, (ii) role-separated *peer and instructor* feedback, and (iii) criterion-level assessment scores on the text to be assessed and the feedback.

3 Course Setting and Study Protocol

Exposía was collected across two iterations of the newly introduced Bachelor-level Computer Science course *Introduction to Scientific Work*: the first cohort in the winter term 2024/25 and the second cohort in the winter term 2025/26.[3]

Course structure In the first cohort, the students wrote an exposé on one of three topics, covering various areas of Computer Science and Artificial Intelligence (AI):[4] *Holographic AI Assistant* (Human-Computer Interaction), *Interactive LLM-based Teaching Agents* (Natural Language Processing), or *Generative AI in Healthcare* (Data Mining). In the second cohort, the students worked on three new topics: *Generative AI for Creative Collaboration* (Natural Language Processing), *Interactive Service Agents* (Human-Computer Interaction), or *Ethical and Transparent AI Systems* (Trustworthy AI). In addition, the second cohort introduced a German-language track alongside the English-language track.

In the beginning of the course,[5] alongside the lectures, the students received guidelines for writing an exposé (Appendix B) and the exposé grading criteria (Section 4.2). Each student produced an exposé draft. This was followed by a peer-review

[3]The institutional IRB has approved the related proposal.
[4]Topics were evenly assigned based on preference voting.
[5]The syllabus is provided in Appendix A.

phase, where each student wrote double-blind reviews for two exposés written by their peers, selected uniformly at random. The students were provided with review grading criteria (Section 4.3). The instructional staff also provided a review for each exposé. This was followed by a revision phase to produce the final exposé, based on the feedback. Both the drafts and final versions, and the student reviews were graded by the instructional staff. The final grade was calculated as follows: draft exposé 30%, peer reviews 30%, and final exposé 40%.[6] Reviewing was facilitated through the open-access CARE platform (Zyska et al., 2023, 2025).

In the first cohort, out of 193 students enrolled in the course,[7] 159 completed all requirements. The second cohort was substantially larger: 445 students submitted a draft exposé, and 436 completed all requirements. The Exposía dataset refers to the *open data* that participants agreed to contribute for public release. All 12 instructors in the first cohort and 23 instructors in the second cohort consented to share their materials. Among students who completed the course, 34% in the first cohort and 39% in the second cohort consented to public release. Consent was requested independently of grading and had no effect on course assessment.

Grade distributions of consenting and non-consenting students Since Exposía only includes data from consenting students, we compare the grade distribution of consenting students with that of the full course population. We find that in both cohorts, the distributions between consenting students and the full course population are similar, and the average grade is the same, indicating that the released subset is not primarily composed of high-achieving students.

Continuous large-scale data collection Exposía is collected continuously and will be released incrementally. The *first release*, covering the winter term 2024/25, contains 55 exposés, 306 reviews, and 2,253 inline comments. The first cohort serves as the basis for the analyses and experimental results reported in the main paper. The *second release*, covering the winter term 2025/26, was finalized shortly before camera-ready and adds 171 exposés, 767 reviews, and 4,773 inline comments. Appendix C documents the second cohort and provides a cross-cohort analysis. The second cohort substantially extends the resource, by providing a larger, bilingual collection, criterion-level scores accompanied by instructor justifications, and information about whether human assessments were conducted with or without LLM support. These additions enable the scaling of peer-feedback processes through AI support, as well as future research on cross-lingual writing assessment and the effects of AI-supported criterion-based assessment. The course runs annually, and the dataset will continue to grow, following similar collection efforts (Dycke et al., 2023).

Consistency protocol for future cohorts We develop a protocol to ensure a consistent data collection and assessment procedure across changes in staff and exposé topics. Specifically, all instructors use the same exposé and review criteria documented in the next section and Appendix D. Instructors are introduced to the criteria before grading begins, and discuss difficult cases in calibration sessions. For future cohorts, a recorded calibration video and representative examples are provided to support onboarding of new instructors.

4 The Exposía Dataset

4.1 Exposía Dataset Overview

Figure 2 illustrates two main components of the Exposía dataset: *exposés* and *feedback*.

Exposés Each exposé is available in two versions: a *draft* and a revised *final version* produced after receiving feedback. The exposés follow a standard Computer Science research proposal template with the following sections: (1) *Motivation*; (2) *Approach* divided into three parts – *Related Work*, *Theoretical Framework*, and *Methodology*; (3) a *Schedule* outlining the week-by-week plan; and (4) a *Bibliography* (Appendix B).

Feedback: inline comments and reviews Each draft exposé received feedback from instructors and other students. Each feedback consists of *inline comments* and a free-text *review*. An *inline comment* consists of a span-level *highlight* on the draft text and a free-text *comment* connected to it (Figure 2). Each inline comment is also associated with a *tag* from a fixed taxonomy $\mathcal{Y} = \{Strength, Weakness, Highlight, Other\}$. A reviewer first adds inline comments. Then the reviewer uses them as the basis for composing a *review*, a document-level free-text feedback on the draft. Appendix D provides more detail.

Grading of the exposés and student reviews The exposés and student reviews are graded by instructors on 36 and 24 assessment criteria, respectively.

[6] Grades for each component are derived from criterion-based scores (see Section 4 and Appendix G.3).

[7] Enrollment was limited because the course was offered for the first time as a pilot.

The same criteria and scoring scales apply to both cohorts. The second cohort additionally includes written justifications for most criterion scores. The criteria are designed to reflect traits that could impact an overall quality of the exposé and review, respectively, and were developed in consultation with the instructors of the course. The criteria are grounded in established research on exposé and feedback writing, as detailed below (see Appendix E for annotation guidelines). Before the course was run, the criteria were assessed by the instructors and their definitions were adjusted for clarity (for more detail, see Appendix H.1).

4.2 Exposé Assessment Criteria and Rubrics

Both the draft and the final exposé are graded independently by an instructional staff member and include 36 per-criterion *assessment scores* (see Table 8 for guidelines on grading). Each criterion is rated on an ordinal scale: [0,1], [0,1,2], or [0,-1,-2].

Designing exposé assessment criteria and rubrics The criteria were derived using 18 *cross-disciplinary guidance documents* for writing thesis exposés from major universities in Germany and Austria (Appendix F), supplemented by observations from instructors in our institution, based on over 100 exposés previously submitted in the Computer Science Department. The 36 criteria cover different aspects of writing a scientific proposal, with a focus on *content* assessment. Importantly, the criteria are *topic-independent* (rather than specific to the research topic of the proposal). This approach is in line with aspect-based assessment used in automated essay scoring (Li and Ng, 2024) and focuses assessment on *transferable proposal-writing skills* (e.g., motivation and methodology). Accordingly, the same 36 criteria are applied uniformly across the proposal topics.

Guided by prior research in rubric construction and fairness-oriented assessment (Jonsson and Svingby, 2007; Mayring, 2000; Kuckartz and Rädiker, 2023), we organize the criteria into nine high-level categories referred to as *rubrics*[8] (Table 2), by grouping together criteria that assess related aspects of the exposé writing.

4.3 Review Assessment Criteria and Rubrics

Each *student review* was independently graded by an instructional staff member on 24 criteria (see Table 9 for the guidelines on grading) organized into eight rubrics (Table 2), with each criterion rated on an ordinal scale of [0,1], [0,1,2], or [-1,1].

Designing review assessment criteria and rubrics The review assessment criteria are grounded in established pedagogical theory and, similar to the exposé criteria, focus on feedback skills that are not tied to the proposal topic (e.g., argumentation, actionable suggestions). The criteria operationalize key educational principles of effective feedback and are primarily informed by the feedback model in Hattie and Timperley (2007).

Drawing on Narciss (2008), we integrate the components of feedback as criteria within the eight rubrics. We further incorporate insights from D'Arcy et al. (2024), which analyzes revisions in response to peer reviews in scientific writing and shows that *actionable comments* offering concrete suggestions are more likely to lead to substantive text revisions. As such, our rubrics emphasize argumentation, accuracy, and clear suggestions for improvement as indicators of pedagogically effective feedback (Shute, 2008). Appendix E.2 provides more detail.

Topic	Exposés	IC	Reviews
T1	19	798	43
T2	18	785	44
T3	18	670	35
Orphaned reviews[*]			184
Total Σ	**55**	**2,253**	**306**

Table 1: Statistics on the Exposía corpus for the first cohort (winter term 2024/25). Topics: T1=*The Holographic AI Assistant*, T2=*Interactive LLM-based teaching agents*, T3=*Generative AI in Healthcare*. IC=Inline Comments. [*]Orphaned reviews are reviews for which no corresponding exposé exists.

5 Exposía Dataset Analysis

5.1 Exposía Dataset Statistics

The dataset statistics for the first cohort are in Table 1. The 55 exposés are evenly distributed across three CS/AI topics: Human–Computer Interaction, Natural Language Processing, and Data Mining / AI in Healthcare, with 19, 18, and 18 exposés, respectively. This provides topical variation while keeping the task, rubric, and feedback protocol fixed. As the dataset comprises only materials for which consent was granted, some exposés lack the full set of three reviews (Table 20), and there exist 184 *orphaned* reviews because the cor-

[8]Rubrics are an effective assessment tool and are constructed to ensure *consistency in evaluation across a diverse student body and multiple evaluators* (Northern Illinois University Center for Innovative Teaching and Learning, 2012). See Appendix E.1 for more details.

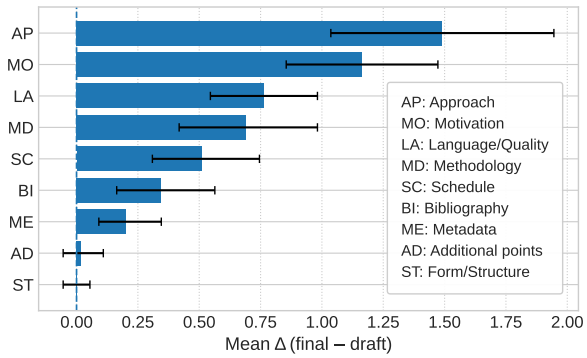


Figure 3: Per-rubric score improvements from draft to final exposé in the first cohort. Each bar shows the mean change in score. Error bars indicate nonparametric bootstrap 95% confidence intervals for each item.

responding exposés could not be made public. All released exposés nevertheless include the complete draft–feedback–revision structure with at least one review. The orphaned reviews further contribute student-written feedback and remain fully usable for review-level analyses and tasks, including review scoring, review-quality assessment, and inter-annotator agreement analysis on student feedback. The number of student reviews is 110, including 76 that are orphaned. Each draft exposé includes on average 41 inline comments. *Weakness* is the most frequent tag (59%) of all comments, followed by *Strength* (18%), *Other* (14%) and *Highlight* (9%).

Exposé revision analysis We analyze the impact of *feedback* by comparing how the final exposé changes relative to the draft: final versions increase in length from 528 to 649 words, on average. We also analyze revisions with respect to exposé criterion-based scores, averaged by rubric. The largest gains occur in *Approach*, *Motivation*, *Language/Quality* and *Methodology* (Figure 3). The most substantial revisions focus on conceptual and argumentative elements (see Appendix G.1).

Review analysis We compare reviews written by instructors vs. students (Appendix G.2): instructor reviews are almost twice as short as student reviews, but have higher lexical density and higher lexical diversity. Appendix G.2 provides more analyses.

5.2 Inter-Annotator Agreement

We compute inter-annotator agreement (IAA) on exposé and review criterion-based scores: each draft exposé and student review are independently graded by two instructors on 36 exposé and 24 review criteria, respectively (using two raters to compute IAA is common practice in similar settings such as AES (Li and Ng, 2024)), yielding 3,960 exposé assessment scores and 6,720 review assessment scores. This yields a substantial number of criterion-level judgments for reliability analysis.

Rubric	QWA	α	r
Exposé rubrics			
Language/Quality (#2)	0.85	0.10	0.19
Metadata (#2)	0.83	0.00	0.31
Form/Structure (#2)	0.97	0.33	0.34
Motivation (#7)	0.76	0.11	0.17
Approach (#8)	0.77	0.22	0.26
Methodology (#7)	0.62	0.18	0.22
Schedule (#4)	0.93	0.36	0.44
Bibliography (#2)	0.78	0.19	0.42
Addt'l points (#2)	0.98	0.32	0.34
All (avg.)	0.79	0.20	0.27
Review rubrics			
Structure/Clarity (#2)	0.86	0.27	0.38
Language (#2)	0.93	0.09	0.17
Feed Up (#2)	0.81	0.43	0.43
Feed Back (#6)	0.86	0.04	0.08
Feed Forward (#2)	0.78	0.02	0.04
Content Quality (#3)	0.94	0.22	0.24
Errors (#5)	0.93	0.10	0.12
Addt'l points (#2)	0.98	-0.01	—
All (avg.)	0.89	0.13	0.18

Table 2: IAA on exposé and review grading in the first cohort. Quadratic weighted agreement (QWA), Krippendorff's (α), and Pearson r. r for Addt'l points (review) is undefined due to zero variance and excluded from the average. Result for each rubric is averaged over all criteria in the rubric. The number of criteria per rubric is next to rubric name. Rubric definitions are in Tables 8 and 9.

Rater training and annotation effort During the semester, each draft exposé and student review were graded by an instructor only once; after the course, each item was graded again to calculate IAA. This annotation effort is substantial: grading one exposé requires assessing it on 36 criteria and providing feedback (inline comments and a review), taking 1–2 hours in total. Grading one review involves assessing it on 24 criteria and takes 20–40 minutes (see Appendix G.3 for more detail). Calibration was conducted before and during grading: Group 1 jointly graded initial items, attended weekly calibration sessions, and discussed difficult cases throughout the semester, while Group 2 received the same guidelines, a recorded calibration video, and an in-person discussion with experienced Group 1 instructors. See Appendix H.1 for

more detail on rater training.

IAA metrics We report Krippendorff's α (Krippendorff, 2019),[9] Pearson correlation coefficient r, and raw quadratic weighted percent agreement (QWA). We include all three metrics because α and r support comparison with prior work in automated writing assessment, while QWA is more informative under prevalence skew. We have several criteria with skewed distributions where the majority label accounts for over 90% cases (Tables 22 and 23). In this case, both α and r can be misleading (Jeni et al., 2013; Zhao et al., 2022), and QWA is the most informative metric. Appendix H.2 and H.3 provide more detail.

IAA on exposé assessment scores Table 2 (top) summarizes IAA results by exposé rubric. Overall agreement yields an average QWA of 0.787, with $\alpha = 0.198$, $r = 0.270$. The observed correlation levels are comparable to those reported on automated essay scoring (Persing and Ng, 2014, 2013) and reflect the subjectivity of the task. The rubrics that attain the lowest agreement levels are those related to content (Motivation, Approach, Methodology). Table 22 reports IAA for all 36 criteria.

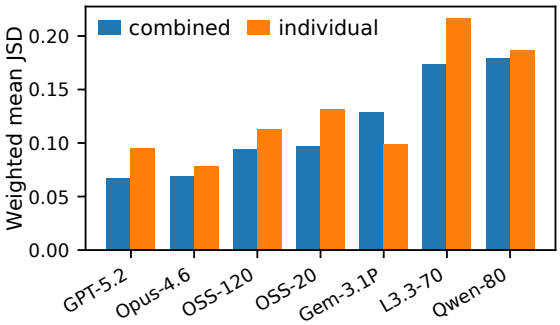
IAA on review assessment scores Table 2 (bottom) summarizes IAA results by review rubric, computed over student reviews. While QWA is higher for review rubrics than for exposé, α and Pearson coefficient are lower, due to the skewed distributions of many criteria (Appendix H.3). For criterion-level agreement and discussion, see Appendix H.2 and Appendix Table 23.

6 Experiments

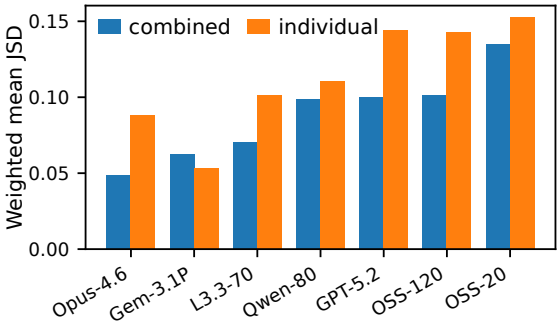
6.1 Experimental Setup

Tasks We use Exposía to benchmark LLM performance on two tasks: (i) criterion-based exposé assessment and (ii) criterion-based assessment of student reviews. We focus on zero-shot prompting to establish reproducible baselines for Exposía. This setting also reflects realistic classroom deployment scenarios, where annotated examples for a new cohort, topic, or criterion may not yet be available.

LLM prompting We assess LLM performance in a zero-shot setting across both open-weight and closed-source models (Appendix J). For both tasks, we use a unified prompting strategy and compare two prompting variants: (i) *single-criterion prompts*, which present and score one criterion at a time, and (ii) *combined prompts*, which list all criteria for the respective task in a single call. Cross-criterion synergy could potentially benefit the model, with a lower execution cost. Appendix K provides more detail about the prompts.



(a) Exposé scoring



(b) Review scoring

Figure 4: Weighted-mean binned JSD between human and LLM score distributions (*lower is better*), by prompting variant. Models: *Opus-4.6*=Claude Opus 4.6; *Gem-3.1P*=Gemini 3.1 Pro; *GPT-5.2*=GPT 5.2; *OSS-120*=GPT OSS (120B); *OSS-20*=GPT OSS (20B); *L3.3-70*=Llama 3.3 (70B); *Qwen-80*=Qwen 3 (80B).

Evaluation Following studies on criterion-based scoring (Pack et al., 2024; Huang and Wilson, 2025), we compare predicted criterion scores to human ratings using the protocol of Elangovan et al. (2025) who recommend using the binned Jensen–Shannon divergence (JSD) for capturing human–machine alignment, instead of agreement metrics such as Krippendorff, Pearson, or QWA, which are designed to account for errors in human annotation, but are not appropriate for computing human–machine agreement. For each criterion, we bin items by the median human score and aggregate the results over agreement and disagreement bins. We compute JSD between the empirical human score distribution and the model's predicted score distribution within each bin. We report the mean binned JSD (*lower is better*), aggregated over criteria (we also report QWA between models and hu-

[9] Krippendorff's α can be used in a setup such as ours, where we are comparing multiple raters where raters also belong to one of the two groups.

man raters for completeness). Results are reported on 55 exposé drafts and 110 student reviews. For each exposé and review, the corresponding model predicts 31 and 22 scores, respectively,[10] for a total of 1705 exposé and 2420 review criterion scores.

6.2 Results

Figure 4 shows key results on exposé and review assessment (averaged over all criteria). Also shown in Table 29, and Table 21 tabulates QWA results. We emphasize the JSD results, however, especially since many of our criteria exhibit relatively high human variation (Table 2), and when the proportion of samples with uncertainty in human-assigned labels is relatively high, machine labels may superficially appear to have similar or better correlation with the human majority label compared to the human-to-human correlation. This can create the illusion that labels from automatic evaluation approximate the human majority label (Elangovan et al., 2025), which is exactly what we observe when evaluating using QWA. Our key findings are as follows:

Combined prompting outperforms single-criterion prompting Across tasks and models, combined prompting yields higher alignments with human raters. Combined prompting is also more efficient computationally (Appendix L and M). This is important since student activity is deadline-driven (Figure 5). Thus, *combined prompting is recommended for classroom deployment*.

Closed-source models outperform open-weight models Across tasks, closed-source models tend to achieve *lower* binned JSD than open-weight models, indicating closer alignment with human score distributions. At the same time, educational deployments impose strict privacy constraints: students may not consent to their work being transmitted to public-facing model APIs. This creates a practical need for better-aligned, locally deployable open-weight models that support privacy-preserving educational applications.

Different models are better suited for exposé and review scoring We find that Claude exhibits consistently high alignment with humans on both tasks. However, other models are less consistent. For instance, *GPT 5.2* attains the best alignment on exposé scoring but is among the poorest models on review scoring, whereas *Llama 3.3* exhibits the opposite behavior. This suggests that different models may be better suited to the two tasks. In Appendix L, we identify criteria on which LLMs align most closely with human raters.

Model performance is stable on review scoring but not exposé scoring Finally, we observe that all models exhibit similar alignment levels on review scoring, whereas on exposé scoring the models exhibit less stability, and the difference between the top-performing model (*GPT 5.2*) and the most poorly-aligned *Qwen 3* is substantial.

Comparison with zero-shot prompting on automated essay scoring Recent zero-shot AES results provide a useful reference point. Lee et al. (2024) report average quadratic weighted kappa scores of 0.550 and 0.587 on ASAP and TOEFL11, respectively, with Mistral-7B-Instruct. These numbers are not directly comparable to the results on Exposía because the tasks, rubrics, and evaluation metrics differ. Nevertheless, they show that zero-shot LLM-based writing assessment remains challenging.

6.3 Illustrative Examples from Exposía

Appendix I presents two examples from Exposía. The examples illustrate how an exposé draft, reviewer feedback, and a revised exposé are connected in the dataset. We briefly discuss one example that focuses on the exposé criterion "Teaser" (Research Question). This example shows how the *Teaser* criterion is used to assess whether a proposal states a clear research question. The relevant excerpt from the draft proposal, including an inline comment from an instructor, is below:

[...] This research addresses the critical gap in AI-driven cancer research by integrating the impact of the tumor microenvironment into data analysis workflows for personalized treatment planning. We will [...] address the limitations of current AI models in incorporating TME data and explore the anticipated improvement of personalized treatment prediction. [WEAKNESS: The goals remain broad; the main research question is not clearly stated.][...]

In the draft, the research direction is understandable, but remains too broad because it describes goals and anticipated contributions rather than a distinct research question. This is reflected in the instructor reviews and the corresponding scores assigned on the *Teaser* criterion (1 and 0 for each instructor, for a maximum of 2, Table 18). We show a review excerpt from the instructor who gave a 1:

[...] Having an explicit, clearly defined research question gives a lot of structure to the overall expose... the exact goals and your main focus are not entirely clear to me. [...]

The final exposé (Appendix I) illustrates how the exposé was successfully revised, following the

[10] 5 exposé and 2 review criteria are excluded (Appendix J).

feedback, and assessed at 2 on this criterion.

Lastly, we analyze LLM-generated assessment scores for this example. We find that the models disagree in their scores. The best-performing model in Figure 4 (*GPT 5.2*) assigned a score of 0 with the following justification: "No explicit research question is stated.". Several models interpreted the passage as an implicit research direction: *Claude Opus* assigned a score of 1 and argued that the work was described in broad terms without a clearly articulated research question. Yet other models gave a score of 2. For example, *Llama 3.3* described the research question as clearly formulated.

Although the instructors disagree on the score, both identified the missing *Teaser*. In contrast, several LLMs, especially open-source ones, failed to detect the absence of a clearly formulated research question. This example highlights the limitations of LLM-based scoring and motivates future work on alternative approaches and improved prompting based on closer analysis of model justifications.

7 Tasks Enabled by Exposía

As discussed in Section 4, the annotations in Exposía are intentionally broad and support a wide range of tasks for scientific writing assistance and assessment, beyond those evaluated in this paper. This richness is a key strength of the dataset.[11] In particular, the dataset enables research on modeling draft–feedback–revision links across documents, with numerous applications. More broadly, the dataset supports research on the relationships among feedback, revision, and writing development in higher-education settings. Because the corpus contains paired drafts and final exposés, criterion-level assessments of both versions, and the corresponding feedback, it enables analyses of textual revisions and draft-to-final score changes with respect to the feedback received. In this paper, we analyze these patterns for the first cohort (Section 5), and find that the largest score increases occur for rubrics related to content and argumentation rather than form. Extending this analysis remains future work.

Exposía supports more fine-grained tasks such as linking inline comments to specific revisions, predicting revisions from feedback, and modeling feedback properties, associated with criterion-level improvements in revised exposés. Such models can help identify types of feedback that are most effective for improving specific writing dimensions. Exposía also enables feedback-generation tasks, where a model produces localized comments or document-level reviews from a draft exposé and the assessment criteria. This is an important application because good feedback can improve students' writing and, at the same time, help students improve their own feedback skills (Wu and Schunn, 2021; Jiang et al., 2024). We leave feedback-generation experiments to future work because responsible evaluation of generated feedback in educational settings requires a dedicated study of correctness, helpfulness, actionability, and potential harm. Future modeling work can also extend our zero-shot baselines with few-shot prompting, rationale-based prompting, prompt optimization, and fine-tuning.

Finally, automated scoring of exposés supports scalable assessment for research proposal writing, while automated review scoring supports scalable assessment of student ability to provide effective peer feedback (a key practical need given the substantial human grading effort; Appendix G.3).

All of these applications directly contribute to addressing the two challenges of (1) research proposal writing and (2) providing feedback on proposal writing. Exposía is the first dataset that explicitly connects the two target skills.

8 Conclusion

We introduce a novel task definition for teaching and assessment of exposé and feedback writing skills, and present the first public dataset, Exposía, of student exposés and feedback, accompanied by human assessment scores based on the fine-grained grading schema we develop.

By integrating drafting, feedback, and revision into a single resource, Exposía enables pedagogically grounded research on writing and feedback in higher education and provides a unified framework that allows for a joint study of feedback and exposé writing quality. We also outline a range of tasks enabled by Exposía. Although the dataset is collected in the Computer Science domain, the task formulation and the assessment framework are general enough to inform the development of models in other areas. We establish baseline performance with state-of-the-art LLMs on exposé and review scoring tasks. Our resource, findings and models pave the path for the holistic study of computational support for academic writing and feedback.

[11] Publishing a dataset with additional components that enable new tasks is consistent with other resource papers in CL venues (Baumel et al., 2014; Qin et al., 2018).

Limitations

This study has several limitations. One limitation is that Exposía is shaped by the course consent protocol: since only consented materials are included, not all exposés retain the complete set of three reviews. For the same reason, some of the reviews in Exposía may be *orphaned*, if the author of the corresponding exposé did not provide consent (Section 5.1). As such, the consent structure reduces the number of complete draft–review–revision instances in the released data.

Consent-based data release may also introduce self-selection bias, for example if students who are more confident in their writing or less concerned about data sharing are more likely to contribute. To assess whether the released subset primarily reflects high-performing students, we compared the grade distribution of consenting students with that of the full course population. The average grade is the same in both groups, and the grade distributions are similar. Nevertheless, grade similarity cannot rule out other forms of self-selection, such as differences in privacy attitudes or confidence. Consent was requested independently of grading outcomes and had no effect on course assessment.

Although Exposía remains modest in size compared with some other NLP datasets, it is substantial for the educational domain, particularly given its complexity and the richness of its annotations. We discuss the time investment required to create the Exposía components in Appendix G.3.

Another limitation concerns the generalizability of the dataset and the assessment criteria beyond Computer Science and AI, to other disciplinary contexts. We stress that the course is designed to teach research proposal writing and peer feedback skills, which generalize across disciplines and are applicable to early-career researchers in general. Furthermore, our assessment criteria are grounded in educational assessment research that is not specific to Computer Science (Section 4) and are derived from cross-disciplinary guidance documents (Appendix F). However, empirical validation of cross-disciplinary transferability is beyond the scope of this paper: it would require collecting multiple comparable draft–feedback–revision datasets across disciplines and evaluating whether human agreement patterns and model performance remain stable across domains. We therefore expect the proposed framework to generalize to other disciplines, although broader empirical validation remains important future work.

Additionally, we leave for future work the question of empirically validating the assessment criteria that we design for exposé and review grading. This is because validating the assessment framework would require a large-scale comparative study in which students are assessed using either the criteria developed in this work, or an alternative assessment scheme, and the outcomes of both groups would need to be compared. Such a study would take more than a year to complete and would require the involvement of several researchers. It is therefore outside the scope of the present work.

Another, related limitation is that Exposía focuses specifically on research proposal writing, which differs from other (academic) writing tasks in purpose, structure, and style (see Section 2). Similarly, exposé feedback differs from other forms of feedback studied in the NLP community, in particular, peer review for academic publishing. As a result, the assessment criteria and experimental findings reported in our study may not transfer directly to other writing tasks and require further validation. At the same time, we expect some aspects to generalize beyond this setting: several criteria, such as language quality and organization, overlap with other areas, for example, automated essay scoring. In contrast, other dimensions, including motivation, methodology, and approach, are more closely tied to the goals of proposal writing.

A further limitation is that open-source models do not yet perform on par with the strongest closed-source systems across the tasks considered in this work. This is especially relevant in educational settings, where student consent and privacy constraints may limit the use of external proprietary models. Future work should therefore focus on developing stronger open-source assessment models and privacy-preserving deployment strategies that better fit the practical and ethical constraints of educational data.

Finally, and most importantly, although LLM-based scoring can be attractive for reducing instructional workload, our results should not be interpreted as support for fully automated grading: the models exhibit systematic scoring tendencies and may miss substantive issues that are crucial for guiding revision, and automated grading can incentivize strategic "gaming" rather than genuine improvement (Alqahtani et al., 2023); we therefore view LLMs as decision-support tools, and as common in AES, recommend a human-in-the-loop workflow when using them for assessment.

Ethical Considerations

Collecting data in an educational setting is challenging due to the high sensitivity of the data, which raises serious privacy and ethical issues (Guan et al., 2023). The data collection was approved by the University's Institutional Review Board and Data Protection Office under IRB proposal **EK 77/2024**.

Data was collected based on informed consent (all instructors and 55 students), following a three-tier structure: (1) general consent for data processing within the review platform, (2) optional consent for contributing textual and highlight data, and (3) optional consent for contributing interaction data. Participation in the study was strictly voluntary and had no effect on grading for the students. No monetary compensation was provided to students, as participation was part of the course setting; importantly, contributing research data was optional. Instructors were compensated via regular University employment contracts, following institutional pay scales and policies. The compensation was not contingent on participation in the study.

We release Exposía to enable research on academic writing, peer feedback, and assessment in educational settings. We stress that our findings are not to be interpreted as justification for fully automatic grading: all assessment-oriented applications of AI should be restricted to decision-support roles with human oversight, in line with the limitations outlined above.

Acknowledgments

This work has been funded by the German Research Foundation (DFG) as part of the PEER project (grant GU 798/28-1), Hessian.AI Connectom Fund (LLMentor: Expert-AI Coteaching of "Introduction to Scientific Work") and by the European Union (ERC, InterText, 101054961). Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them. We thank Dr. Ekaterina Kochmar (Mohamed bin Zayed University of Artificial Intelligence) and Dr. Martin Greisel (University of Augsburg) for their valuable feedback and discussions.

References

Tariq Alqahtani, Hisham A. Badreldin, Mohammed Alrashed, Abdulrahman I. Alshaya, Sahar S. Alghamdi, Khalid bin Saleh, Shuroug A. Alowais, Omar A. Alshaya, Ishrat Rahman, Majed S. Al Yami, and Abdulkareem M. Albekairy. 2023. The emergent role of artificial intelligence, natural learning processing, and large language models in higher education and research. *Research in Social and Administrative Pharmacy*, 19(8):1236–1242.

Tal Baumel, Raphael Cohen, and Michael Elhadad. 2014. Query-chain focused summarization. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 913–922, Baltimore, Maryland. Association for Computational Linguistics.

Paul Black and Dylan Wiliam. 1998. Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice*, 5(1):7–74.

Scott Crossley, Yu Tian, Perpetual Baffour, Alex Franklin, Youngmeen Kim, Wesley Morris, Meg Benner, Aigner Picou, and Ulrich Boser. 2023. The english language learner insight, proficiency and skills evaluation (ellipse) corpus. *International Journal of Learner Corpus Research*, 9(2):248–269.

Mike D'Arcy, Alexis Ross, Erin Bransom, Bailey Kuehl, Jonathan Bragg, Tom Hope, and Doug Downey. 2024. ARIES: A corpus of scientific paper edits made in response to peer reviews. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6985–7001, Bangkok, Thailand. Association for Computational Linguistics.

Jiangshu Du, Yibo Wang, Wenting Zhao, Zhongfen Deng, Shuaiqi Liu, Renze Lou, Henry Peng Zou, Pranav Narayanan Venkit, Nan Zhang, Mukund Srinath, Haoran Ranran Zhang, Vipul Gupta, Yinghui Li, Tao Li, Fei Wang, Qin Liu, Tianlin Liu, Pengzhi Gao, Congying Xia, and 21 others. 2024. LLMs assist NLP researchers: Critique paper (meta-)reviewing. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5081–5099, Miami, Florida, USA. Association for Computational Linguistics.

Nils Dycke, Ilia Kuznetsov, and Iryna Gurevych. 2023. NLPeer: A unified resource for the computational study of peer review. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5049–5073, Toronto, Canada. Association for Computational Linguistics.

Aparna Elangovan, Lei Xu, Jongwoo Ko, Mahsa Elyasi, Ling Liu, Sravan Babu Bodapati, and Dan Roth. 2025. Beyond correlation: The impact of human uncertainty in measuring the effectiveness of automatic evaluation and LLM-as-a-judge. In *The Thirteenth International Conference on Learning Representations*.

Rudolph Flesch. 1948. A new readability yardstick. *Journal of Applied Psychology*, 32(3):221–233.

Susan M. Gass and Alison Mackey, editors. 2012. *The Routledge handbook of second language acquisition*. Routledge.

Gabrielle Gaudeau. 2025. Beyond the gold standard in analytic automated essay scoring. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 18–39, Vienna, Austria. Association for Computational Linguistics.

Larry Grabau and Patricia Wilson. 1995. Jumping on thin ice: Values argument writing assignment for a large enrollment plant science class. *Journal of Natural Resources and Life Sciences Education*, 24(2):185–189.

Victor Grech. 2018. Wasp (write a scientific paper): To strive, to seek, to find and not to yield. *Early Human Development*, 127:96–97.

Xiu Guan, Xiang Feng, and A Y M Atiquil Islam. 2023. The dilemma and countermeasures of educational data ethics in the age of intelligence. *Humanities and Social Sciences Communications*, 10:1–14.

John Hattie and Helen Timperley. 2007. The power of feedback. *Review of Educational Research*, 77(1):81–112.

Yue Huang and Joshua Wilson. 2025. Evaluating LLM-based automated essay scoring: Accuracy, fairness, and validity. In *Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con): Works in Progress*, pages 71–83, Wyndham Grand Pittsburgh, Downtown, Pittsburgh, Pennsylvania, United States. National Council on Measurement in Education (NCME).

Bart Huisman, Nadira Saab, Paul van den Broek, and Jan van Driel. 2019. The impact of formative peer feedback on higher education students' academic writing: a meta-analysis. *Assessment & Evaluation in Higher Education*, 44(6):863–880.

Shin'ichiro Ishikawa. 2023. *The ICNALE Guide: An Introduction to a Learner Corpus Study on Asian Learners' L2 English*, 1st edition. Routledge.

László A. Jeni, Jeffrey F. Cohn, and Fernando De la Torre. 2013. Facing imbalanced data-recommendations for the use of performance metrics. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction, ACII 2013, Geneva, Switzerland, September 2-5, 2013*, pages 245–251. IEEE Computer Society.

Yang Jiang, Jiangang Hao, Michael Fauss, and Chen Li. 2024. Detecting chatgpt-generated essays in a large-scale writing assessment: Is there a bias against non-native english speakers? *Computers & Education*, 217(C).

Anders Jonsson and Gunilla Svingby. 2007. The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review*, 2(2):130–144.

Dongyeop Kang, Waleed Ammar, Bhavana Dalvi, Madeleine van Zuylen, Sebastian Kohlmeier, Eduard Hovy, and Roy Schwartz. 2018. A dataset of peer reviews (PeerRead): Collection, insights and NLP applications. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1647–1661, New Orleans, Louisiana. Association for Computational Linguistics.

Omid Kashefi, Tazin Afrin, Meghan Dale, Christopher Olshefski, Amanda Godley, Diane Litman, and Rebecca Hwa. 2022. Argrewrite v.2: an annotated argumentative revisions corpus. *Language Resources and Evaluation*, 56(3):881–915.

Zixuan Ke and Vincent Ng. 2019. Automated essay scoring: A survey of the state of the art. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 6300–6308. ijcai.org.

Charles Kivunja. 2016. How to write an effective research proposal for higher degree research in higher education: Lessons from practice. *International Journal of Higher Education*, 5:163–172.

Klára Kolářová and Timothée Schmude. 2026. Start using justifications when explaining ai systems to decision subjects. In *Digital Humanism*, pages 190–202, Cham. Springer Nature Switzerland.

Klaus Krippendorff. 2019. *Content Analysis: An Introduction to Its Methodology*, fourth edition edition. SAGE Publications, Inc., Thousand Oaks, California.

Udo Kuckartz and Stefan Rädiker. 2023. *Qualitative Content Analysis: Methods, Practice and Software*, 2 edition. SAGE Publications Ltd, London.

Sanwoo Lee, Yida Cai, Desong Meng, Ziyang Wang, and Yunfang Wu. 2024. Unleashing large language models' proficiency in zero-shot essay scoring. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 181–198, Miami, Florida, USA. Association for Computational Linguistics.

Vladimir I. Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet physics. Doklady*, 10:707–710.

Shengjie Li and Vincent Ng. 2024. ICLE++: Modeling fine-grained traits for holistic essay scoring. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8465–8486, Mexico City, Mexico. Association for Computational Linguistics.

Sheng Lu, Ilia Kuznetsov, and Iryna Gurevych. 2025. Identifying aspects in peer reviews. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 6145–6167, Suzhou, China. Association for Computational Linguistics.

Sandeep Mathias and Pushpak Bhattacharyya. 2018. ASAP++: Enriching the ASAP automated essay grading dataset with essay attribute scores. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Philipp Mayring. 2000. Qualitative content analysis. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research*, 1(2).

Rizwana Muneer, Rida Batool, and Saima Zehra. 2020. A qualitative study on challenges that post- graduate students face in research proposal writing at university level. *International Journal of Social Sciences: Current and Future Research Trends*, 5(1):1–6.

Susanne Narciss. 2008. Feedback strategies for interactive learning tasks. In J. M. Spector, M. D. Merrill, J. van Merriënboer, and D. M. Driscoll, editors, *Handbook of research on educational communications and technology*, pages 125–144. Routledge, New York.

Northern Illinois University Center for Innovative Teaching and Learning. 2012. Rubrics for assessment. `https://www.niu.edu/citl/resources/guides/instructional-guide`. Retrieved from Northern Illinois University Center for Innovative Teaching and Learning website.

OpenAI. 2025. gpt-oss-120b & gpt-oss-20b model card. *Preprint*, arXiv:2508.10925.

Austin Pack, Alex Barrett, and Juan Escalante. 2024. Large language models and automated essay scoring of english language learner writing: Insights into validity and reliability. *Computers and Education: Artificial Intelligence*, 6:100234.

Ellis B. Page. 1966. The imminence of grading essays by computer. *The Phi Delta Kappan*, 47(5):238–243.

Isaac Persing and Vincent Ng. 2013. Modeling thesis clarity in student essays. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 260–269, Sofia, Bulgaria. Association for Computational Linguistics.

Isaac Persing and Vincent Ng. 2014. Modeling prompt adherence in student essays. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1534–1543, Baltimore, Maryland. Association for Computational Linguistics.

Ildiko Pilan, John Lee, Chak Yan Yeung, and Jonathan Webster. 2020. A dataset for investigating the impact of feedback on student revision outcome. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 332–339, Marseille, France. European Language Resources Association.

Lianhui Qin, Lemao Liu, Wei Bi, Yan Wang, Xiaojiang Liu, Zhiting Hu, Hai Zhao, and Shuming Shi. 2018. Automatic article commenting: the task and dataset. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 151–156, Melbourne, Australia. Association for Computational Linguistics.

Afifa Rastri, Yanti Sri Rezeki, Urai Salam, Dwi Riyanti, and Surmiyati Surmiyati. 2023. An analysis of students' problems in writing a research proposal. *Acitya: Journal of Teaching and Education*, 5:57–71.

Roman Rietsche, Andrew Caines, Cornelius Schramm, Dominik Pfütze, and Paula Buttery. 2022. The specificity and helpfulness of peer-to-peer feedback in higher education. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pages 107–117, Seattle, Washington. Association for Computational Linguistics.

Qian Ruan, Ilia Kuznetsov, and Iryna Gurevych. 2024. Are large language models good classifiers? a study on edit intent classification in scientific document revisions. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15049–15067, Miami, Florida, USA. Association for Computational Linguistics.

Valerie J. Shute. 2008. Focus on formative feedback. *Review of Educational Research*, 78(1):153–189.

Moritz Staudinger, Wojciech Kusa, Florina Piroi, and Allan Hanbury. 2024. An analysis of tasks and datasets in peer reviewing. In *Proceedings of the Fourth Workshop on Scholarly Document Processing (SDP 2024)*, pages 257–268, Bangkok, Thailand. Association for Computational Linguistics.

Llama Team. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Qwen Team. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Yong Wu and Christian D. Schunn. 2021. The effects of providing and receiving peer feedback on writing performance and learning of secondary school students. *American Educational Research Journal*, 58(3):492–526.

Jin Xue, Xiaoyi Tang, and Liyan Zheng. 2021. A hierarchical bert-based transfer learning approach for multidimensional essay scoring. *IEEE Access*, 9:125403–125415.

Fan Zhang, Homa B. Hashemi, Rebecca Hwa, and Diane Litman. 2017. A corpus of annotated revisions for studying argumentative writing. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1568–1578, Vancouver, Canada. Association for Computational Linguistics.

Xinshu Zhao, Guangchao Charles Feng, Song Harris Ao, and Piper Liping Liu. 2022. Interrater reliability estimators tested against true interrater reliabilities. *BMC Medical Research Methodology*, 22(1):232.

Dennis Zyska, Nils Dycke, Jan Buchmann, Ilia Kuznetsov, and Iryna Gurevych. 2023. CARE: Collaborative AI-assisted reading environment. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 291–303, Toronto, Canada. Association for Computational Linguistics.

Dennis Zyska, Ilia Kuznetsov, Florian Müller, and Iryna Gurevych. 2025. Pull requests from the classroom: Co-developing curriculum and code. In *Proceedings of the Mensch Und Computer 2025*, MuC '25, pages 689–693, New York, NY, USA. Association for Computing Machinery.

A Course Syllabus

This appendix documents the syllabus of the undergraduate computer science course *Introduction to Scientific Work*, which provided the instructional setting for the data collection described in Section 3.

The course covers foundational concepts and practical skills required for conducting scientific research. Here we present the weekly lectures:

1. What is science? Course overview and introduction
2. Empirical methods and research design
3. Literature review and academic search strategies
4. Project proposals (exposés); *start writing exposé draft*
5. Scientific writing
6. LaTeX for academic writing
7. Research Questions and practical exercise on academic literature search
8. Introduction to peer review
9. Feedback; *start writing peer feedback*
10. Generative AI in scientific work
11. Research ethics I
12. Research ethics II
13. Workshop: Software development
14. Workshop: Research and project management

B Exposé Template and Writing Guidelines

Panel *Exposé template* shows the content of the LaTeX exposé template that students received as part of the course and used to create their exposés. While the complete instructions were provided in an extensive slide-based lecture (Appendix A), the excerpt reproduced here is the concise reference sheet embedded directly in the template. Because this template guided how students organized and formatted their drafts, we include it to ensure transparency and to document the pedagogical context in which the dataset was created.

C The Second Cohort: Winter Term 2025/26

The second Exposía cohort became available after the paper had completed review and was released shortly before camera-ready. Consequently, the analyses and experiments in the main paper remain based on the first cohort. We nevertheless include the second cohort in the present dataset release and document it here because it constitutes a substantial extension of the resource.

Course and participation The course was offered for a second time in the winter term 2025/26, following the same draft–feedback–revision structure, the same exposé and review criteria, and the same reviewing workflow as in the first cohort. The students worked on three new research topics: *Generative AI for Creative Collaboration* (Natural Language Processing), *Interactive Service Agents* (Human–Computer Interaction), and *Ethical and Transparent AI Systems* (Trustworthy AI). The third topic introduces Trustworthy AI, an area not represented among the three topics in the first cohort.

Of the 445 students who submitted a draft exposé, 436 completed all course requirements. Among these students, 39% consented to the public release of their materials, compared with 34% in the first cohort. Consent rates were similar across the three topics, ranging from 32% to 37%. The resulting release contains 171 exposés, 767 reviews, and 4,773 inline comments. Table 3 reports the distribution across topics using the same definitions as the dataset overview in the main paper.

Bilingual course tracks The second release additionally contains German-language exposés and reviews from participants who consented to the public release of their materials. The exposé and review assessment criteria are introduced in Sections 4.2 and 4.3 and documented in detail in Appendix E.1. The two language versions use the same rubric codes, criterion order, scoring levels, and point ranges. Both language versions of the exposé and review criteria are included in the release, and each record identifies the language in which it was produced.

Exposé template

1 Motivation

Hook – Introduce the topic and catch the reader's interest.

- *Concrete*: Quote, example, real-world event.
- *Helpful*: Directly useful or valuable information.
- *Convincing*: Strong argument or empirical fact.
- *Shocking*: Unexpected or counterintuitive finding.
- *Personal*: Research journey or applied experience.
- *Understandable*: Link complex ideas to everyday life.

Anchor – Situate the problem concisely.

- Discipline, field, or domain.
- How the problem has been addressed so far.
- Current academic relevance.
- Deficits or criticism in existing literature.

Teaser – Identify the gap and position your research.
Research Questions

- Break the main RQ into meaningful sub-questions.
- Clarify terminology; define if needed.
- Set clear boundaries and scope.

2 Approach

State of the Art (SOTA)

- Refer to relevant existing work.
- Highlight research deficits and criticism.
- Combine different traditions or perspectives.

Theoretical Framework

- Introduce theories used in the study.
- Explain key concepts and definitions.
- Describe how the theory informs the RQs.

Note: Theories need not be SOTA but should be connected to it.
Methodology

- Specify method (quantitative, qualitative, mixed, source research).
- Define target population and sampling.

- State data sources and availability.
- Clarify agreements or permissions.
- Anticipate potential difficulties.
- Discuss methodological limitations.
- Explain relevance and scope.
- Describe variables and their relationships.
- Outline practical implementation (software, documents, helpers).
- Clarify epistemic goals (knowledge to be gained).
- If building on prior work, state progress so far.

3 Schedule

Provide a chronological work plan with realistic milestones.
Tips:

- Break tasks into concrete steps.
- Include checkpoints and buffer time.
- Use weekly or monthly units depending on project size.
- Seek advice from experienced researchers.

Example Schedule

Month	Week	Task	Description
1	1	Planning	Work plan, coordination
1	2–3	Literature Review	Search and initial analysis
1	4	Literature Evaluation	Critical review and selection
2	5–6	Theoretical Framework	Develop theory and methodology
2	7	Concept Development	Draft interview/ analysis guide
2	8	Scenario Planning	Design simulations or experiments
3	9–10	Recruitment	Acquire participants
3	11	Data Collection	Conduct interviews/ experiments
3	12	Data Preparation	Clean and pre-analyze data
4	13–14	Qualitative Analysis	In-depth qualitative analysis
4	15–16	Quantitative Analysis	Statistical analysis
5	17	Writing Results	Compose results section
5	18	Discussion	Interpret findings
5	19	Methodology Review	Finalize methodology
5	20	Full Revision	Edit and review all chapters
6	21	Final Editing	Final proofreading, formatting
6	22–24	Finalization	Buffer, printing, submission

Reference

[1] Helmut Balzert, Christian Schäfer, Marion Schröder, Uwe Kern, Roman Bendisch, and Klaus Zeppenfeld. *Wissenschaftliches Arbeiten: Ethik, Inhalt und Form wissenschaftlicher Arbeiten, Handwerkszeug, Quellen, Projektmanagement, Präsentation*. W3L-Verlag, Herdecke/Witten, 2011. ISBN: 978-3-937137-59-9.

Topic	Exposés	IC	Reviews
T1	63	1,511	102
T2	46	1,276	81
T3	46	1,510	86
not recorded	16	476	25
*Orphaned reviews**	–	–	473
Σ	**171**	**4,773**	**767**

Table 3: Statistics for the second Exposía cohort. Topics: T1=*Generative AI for Creative Collaboration*, T2=*Interactive Service Agents*, and T3=*Ethical and Transparent AI Systems*. IC=Inline Comments. *Orphaned reviews are reviews for which the corresponding exposé is not part of the public release.

Of the 171 released exposés, 149 are associated with the English-language track and 22 with the German-language track. The review data comprise 673 English-language reviews and 96 German-language reviews.

LLM-assisted grading The second cohort additionally included an LLM-assisted condition for the instructor assessment of draft exposés. As described in Section 4, instructors assess each exposé using the criterion-based grading scheme introduced in Section 4.2 and documented in detail in Appendix E.1.

Before an instructor assessed a draft exposé, an on-premise Qwen3-14B model processed its LaTeX source together with the corresponding assessment criterion and its scoring descriptors. For each criterion, the model was instructed to assign an integer score and provide a one–two sentence justification grounded in specific evidence from the exposé. The model additionally generated draft textual feedback. Criterion-level assessments were generated separately but batched into a single inference call per submission. We used deterministic decoding (temperature 0), and the criterion definitions were matched to the language of the submission.

Instructors were assigned either to an assisted or a control condition. In the assisted condition, the model-generated scores, justifications, and draft feedback were shown pre-filled in the grading interface; instructors could accept or modify them before submitting their assessment. In the control condition, instructors used the same grading interface, but the model-generated pre-assessment was not shown. The condition was assigned at the instructor level rather than independently for each exposé. The use of LLM-assisted grading in the course was approved by the university's ethics board under approval EK 56/2025.

In the second cohort, instructors also provided short written justifications for most criterion-level scores. These justifications are included in the dataset together with their provenance, distinguishing instructor-written justifications from model-generated suggestions that were accepted verbatim. For example, for the criterion *Hook existing*, one model-generated justification reads: "The exposé opens with a compelling statistic: 'Only 13% of germans trust news media which were mainly created by artificial intelligence (AI).' This directly addresses the topic's relevance and establishes a clear problem statement, making it a suitable hook."

The release contains 84 assisted draft gradings. In these assessments, 1,048 of 2,694 recorded justifications were submitted unchanged from the model-generated suggestion, and 2,278 of 2,602 criterion scores (88%) were identical to the model's proposed score. In the control condition, where the same model pre-assessment was generated but not shown to instructors, 77% of the human-assigned scores were identical to the model proposal.

Comparison of the releases Table 4 compares the two released cohorts under consistent counting rules. The second release is substantially larger in absolute terms and introduces written criterion-level justifications, which are available for 16,160 of its 19,464 numeric criterion scores (83%).

In the first cohort, a dedicated second grading pass was collected after the course specifically to estimate inter-annotator agreement (IAA). Since the second cohort used the same assessment criteria and annotation protocol, and the reliability of this protocol had already been evaluated on the first cohort, we did not consider a second grading pass necessary for the purposes of the present release.

The larger cohort did not produce proportionally more feedback per document. Inline comments decreased from 41.0 to 27.9 per released exposé, while reviews per exposé decreased from 2.22 to 1.72. The composition of the inline feedback also shifted: the proportion labelled as a weakness decreased from 59% to 50%, while the proportion labelled as a strength increased from 18% to 26%.

Cross-cohort LLM baseline To examine whether the first-cohort baseline transfers to the second cohort, we repeated a reduced version of the automated assessment experiment from Section 6. We evaluated GPT-5.2 under zero-shot combined prompting, using the same prompt construction and the same 31 exposé and 22 review criteria as in the

	WS 24/25	WS 25/26
Exposés	55	171
Reviews	306	767
orphaned	184	473
Inline comments	2,253	4,773
per exposé	41.0	27.9
Weakness	59%	50%
Strength	18%	26%
Other	14%	12%
Highlight	8%	10%
Annotations	5,765	11,584
Comments	5,797	11,782
Comment votes	592	105
Display states†	–	562
Gradings	450	649
raw	538	778
drafts	110	162
finals	55	162
reviews	285	325
Criterion scores	12,779	19,464
justified	–	16,160
Behaviour records	586,505	1,991,894
Languages	EN	EN, DE
Second grading pass (IAA)	yes	–

Table 4: Comparison of the two Exposía cohorts using the same counting procedure. Gradings are deduplicated by author, grader, type, review, and criterion; raw counts include duplicated storage locations. Behaviour records cover both the course and the subsequent assessment. †Display states indicate whether a comment card was expanded or collapsed. EN = English; DE = German.

Rubric	WS 24/25	WS 25/26	Δ
Exposé drafts ($n = 55$ vs. 87, control condition)			
Language	0.88	0.86	−0.02
Metadata	0.90	0.92	+0.02
Motivation	0.83	0.87	+0.04
Approach	0.86	0.87	+0.01
Methodology	0.72	0.76	+0.04
Schedule	0.94	0.91	−0.03
Bibliography	0.57	0.56	−0.01
All criteria	**0.82**	**0.83**	**+0.01**
Reviews ($n = 106$ vs. 337)			
Structure and Clarity	0.84	0.88	+0.04
Language	0.94	0.92	−0.02
Feed Up	0.83	0.80	−0.03
Feed Back	0.88	0.85	−0.03
Feed Forward	0.87	0.60	−0.27
Content Quality	0.91	0.90	−0.01
Errors	0.85	0.91	+0.06
All criteria	**0.87**	**0.85**	**−0.02**

Table 5: Quadratic-weighted agreement between GPT-5.2 under combined prompting and human criterion scores in WS 2024/25 and WS 2025/26. The second-cohort exposé evaluation is restricted to the control condition, whose graders were not shown the LLM pre-assessment. The second cohort contains only one human grading pass, so no within-cohort human–human agreement reference can be computed.

first-cohort experiment. The exposé evaluation was restricted to the control condition, whose graders were not shown the LLM pre-assessment. The review evaluation used all graded reviews available in the experimental input, since no LLM-generated criterion scores were displayed during review grading. The English-language rubric descriptions were used in the prompts for both English- and German-language records. Table 5 reports quadratic-weighted agreement between the model and the first human grading group.

The binned Jensen–Shannon divergence analysis from Section 6 cannot be extended to the second cohort. Its agreement and disagreement bins depend on whether at least two independent human raters assigned the same score to an item. Duplicating the single available rating would place every item in the agreement bin by construction and would therefore not approximate the missing quantity.

Overall model–human agreement on exposé drafts remains similar across cohorts, increasing from 0.82 to 0.83. Agreement on reviews decreases slightly, from 0.87 to 0.85. The largest criterion-level change occurs for Feed Forward, where agreement decreases from 0.87 to 0.60. This is also one of the criteria for which the mean human-assigned score decreases most strongly in Table 7. Thus, the score comparison and the independently computed model-agreement comparison both identify Feed Forward as a principal difference between the two cohorts.

Score comparison Table 7 reports descriptive per-rubric scores for the two cohorts. For the first cohort, we use the in-semester grading pass (Group 1), which is the basis of the first-cohort descriptive score analysis and is directly comparable to the single grading pass available for the second cohort.

Restricting the second-cohort drafts to graders who were not shown the LLM pre-assessment (*control*, $n = 78$), the mean total draft score increases

from 33.00 to 35.14. When the 84 assisted gradings are included (*all*, $n = 162$), the mean increases further to 35.49. Mean review scores decrease from 22.33 to 21.27. The largest changes occur for the Feed Up and Feed Forward rubrics. Mean final-exposé scores increase from 38.20 to 39.52.

Limits on direct comparison The cross-cohort experiment should be interpreted as a transfer check rather than a full replication of the experiments in Section 6. It evaluates only GPT-5.2 under combined prompting, and the exposé evaluation is restricted to the control condition, in which graders were not shown the LLM pre-assessment. Because the second cohort contains only one human grading pass, no second-cohort human–human agreement reference can be computed. For the same reason, the binned Jensen–Shannon divergence analysis from Section 6 cannot be extended to this cohort, since its agreement and disagreement bins require at least two independent human ratings.

More generally, the cohorts should not be pooled without accounting for differences in language, research topics, grader composition, and assessment procedure. The second cohort includes both English- and German-language records and an LLM-assisted draft-grading condition. Because this condition was assigned at the grader level, LLM assistance is confounded with grader identity, preventing a clean estimate of the effect of visible model support. Moreover, the cross-cohort experiment uses English-language rubric descriptions in the prompts for both English- and German-language records and therefore does not isolate language effects.

The first cohort consequently remains the basis for the agreement analyses and model benchmarks in the main paper. The second cohort instead provides a transfer check together with a substantially larger bilingual collection, criterion-level written justifications, and explicit provenance for assessments produced with and without visible LLM support.

D Additional Details on the Exposía Dataset

Exposés Each exposé is available in two versions: a *draft* and a revised *final version* produced after receiving feedback.

Each version is provided as (i) a compiled PDF file and (ii) the original LaTeX source, including the separate bibliography file (`.bib`) and all textual content.

Reviews Every exposé draft is accompanied by two *student reviews* written by fellow students taking the course and at least one *instructor review*. Instructional staff included nine instructors and consisted of PhD students, postdoctoral researchers, and teaching assistants holding a Master's degree (Appendix H.1 (Group 1)). The feedback was double-blind among students, but not with respect to the instructional staff. During the revision process, students could identify whether a specific feedback came from an expert or a student.

Each review is released as (i) an HTML document that preserves the original formatting used on the review platform and (ii) a plain-text rendering. For all reviews, we provide basic metadata including a synthetic reviewer name, reviewer role, document size, the associated exposé (if available), and an anonymized identifier.

For all reviews for which the author provided additional consent for behavioral logging, we release the full edit history of the review text as stored by the platform, capturing the successive revisions made by the reviewer while composing their feedback. This results in sequences of review drafts that can be aligned chronologically via timestamps.

Inline comments and voting Beyond document-level feedback, the platform supports fine-grained, span-based comments on the exposé draft. For each exposé, we collect a set of inline comments created by students and instructors while reading the draft. Each comment is attached to a character-level span in the PDF-rendered text of the exposé and is associated with (i) optional free-text content, (ii) a categorical tag from a fixed taxonomy $\mathcal{Y} = \{\textit{Strength}, \textit{Weakness}, \textit{Highlight}, \textit{Other}\}$ and (iii) metadata about the author role (student or instructor) and the underlying review document it belongs to.

The platform further allows other participants, but not the comment author, to cast binary votes (*thumbs up* / *thumbs down*) on each comment. These votes were intended as judgments of comment helpfulness. However, interviews with participants indicate that student authors sometimes used the voting mechanism as a personal *checklist* during revision (marking comments as *done* rather than as helpful), despite explicit instructions from the instructors to use it purely as a helpfulness signal. We therefore provide the raw voting data, together with information about voter role, but caution that votes do not form a clean ground truth for perceived usefulness.

Action	Students (n=45)	Instructors (n=9)	Total
Mouse move	73,811	131,272	205,083
PDF page show/hide	53,378	102,120	155,498
Annotation Page scroll	18,944	31,080	50,024
Browser Tab show/hide	16,810	31,657	48,467
PDF (Browser) resize	13,377	25,475	38,852
Text select	7,403	8,043	15,446
Sidebar scroll	2,445	7,248	9,693
Navigation step	2,705	3,742	6,447
Sidebar click	2,882	3,382	6,264
Topbar click	1,826	3,512	5,338
PDF scroll	2,206	2,179	4,385
Copy text	1,403	2,371	3,774
Table click	1,811	1,164	2,975
Text paste	1,208	1,435	2,643
Browser load	715	1,364	2,079
Modal show	889	324	1,213
Modal hide	396	304	700
Search used	102	369	471
Modal Button click	259	176	435
Total	202,570	357,217	559,787

Table 6: Number of behaviour data rows per action, split by user group for the first cohort.

Behavioral interaction logs For users who provided explicit additional consent, we release interaction logs from the review platform, covering 45 students and 9 instructors. Each record contains an `action` type, a JSON-encoded `data` payload, a `timestamp`, and an anonymized `user` identifier. Logged events are timestamped UI interactions (Table 6); an overview of activity over the semester is shown in Figure 5 and discussed in Appendix N . Actions including navigation and visibility changes (e.g., `Navigation step`, `PDF page show/hide`, `Browser Tab show/hide`), scrolling and resizing (e.g., `Annotation Page scroll`, `PDF scroll`, `Sidebar scroll`, `PDF (Browser) resize`), interface controls and modal events (e.g., `Sidebar click`, `Topbar click`, `Modal show`/`Modal hide`, `Modal Button click`), and lightweight editing interactions (e.g., `Text select`, `Copy text`, `Text paste`).

Authorship metadata and anonymization Each exposé, review, inline comment, vote, and interaction event is associated with an anonymized author identifier and a role label (*student*, *junior*, *senior*). To support more interpretable examples and qualitative analysis, we replace real names with consistently applied synthetic names generated with `Faker`.[12] No raw platform identifiers, email addresses, or other personally identifying information are released. This setup enables the analysis of role-specific behavior and bias while preserving participant privacy.

[12] https://faker.readthedocs.io/en/master

E Guidelines on Criteria and Rubrics for Exposé and Review Assessment

This appendix provides an overview of exposé and review criteria and the guidelines used by the instructors when grading the exposés and student reviews, respectively.

Expertise levels Each criterion is assigned an *expertise level*[13], indicating the type of judgment needed to assess the quality of the exposé or review with respect to this criterion. We introduce a three-level expertise scheme (expertise levels by criteria are shown in Tables 8 and 9 for exposé and review criteria, respectively):

- **Level 0 (formal).** No academic or expert knowledge required. Largely objective. Sample exposé criteria: *Metadata available*; *Template used*; *Number of pages*. Review criteria: None.
- **Level 1 (academic).** Comprehension-based judgments of clarity, coherence, and basic research sense-making, but no specific in-domain expertise is required. Sample exposé criteria: *Language quality*; *Hook existing*. Sample review criteria: *Summary available*; *Language*.
- **Level 2 (expert).** Expert judgment of the soundness and adequacy of the reasoning and its alignment with disciplinary standards. Sample exposé criteria: *Teaser (RQ)*; *Relevance of theoretical framework*. Sample review criteria: *Knowledge of correct response*; *Learning goals*.

E.1 Annotation Guidelines: Exposé Grading Criteria and Rubrics

Table 8 lists all the 36 exposé assessment criteria, grouped into nine high-level rubrics. These criteria define the scoring rules applied to each submitted exposé.

[13] Two co-authors inspected all criteria and labeled each for the expertise level required for scoring an exposé for a given criterion.

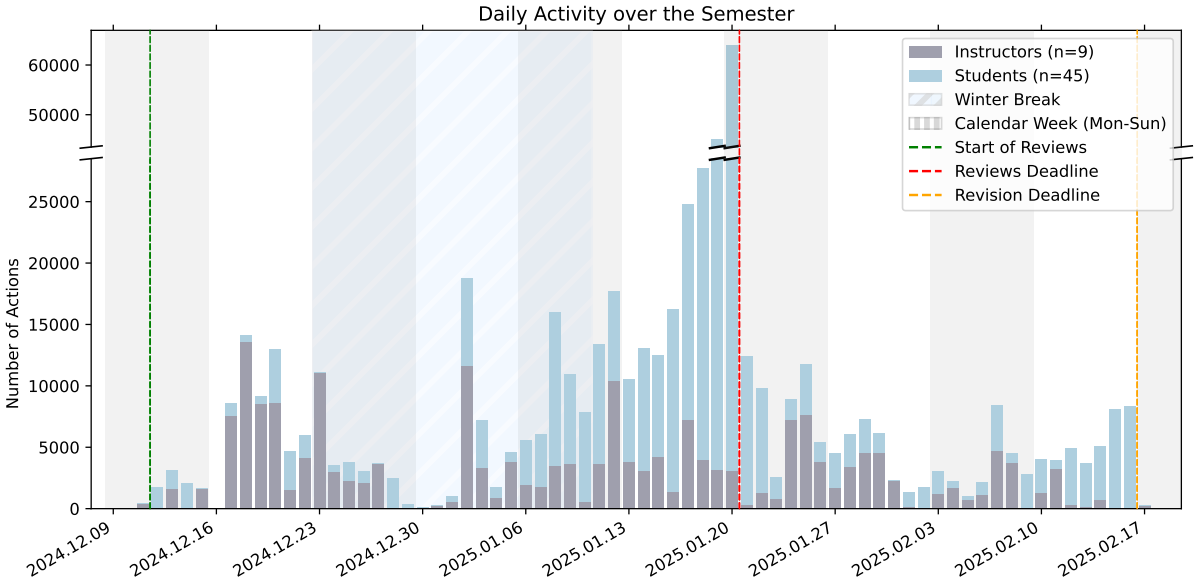


Figure 5: Daily number of actions across the semester for instructors (n=9) and students (n=45). Bars show per-day activity (instructors in gray, students in blue). Alternating vertical shading indicates calendar weeks (Mon–Sun), and the hatched region marks the winter break. Dashed lines mark the start of reviews (green), the review deadline (red), and the revision deadline (orange). The y-axis is broken to accommodate the peak activity before the review deadline.

Refinement of criteria and rubrics for exposé assessment Before the course was run, the designed criteria and rubrics were piloted on exposés previously submitted within the department: two members of the instructional staff assessed together a small sample of old student exposés, and other instructional staff members went over these to give feedback. This feedback was used to verify that every criterion is assessable and the criteria definitions are clear. Criteria definitions were adjusted accordingly for clarity and consistency.

E.2 Annotation Guidelines: Review Grading Criteria and Rubrics

Table 9 lists all the 24 review assessment criteria grouped into eight high-level rubrics. These criteria define the scoring rules applied to each submitted student review. While actionable feedback promotes learning, ineffective or misleading reviews can have the opposite effect. To this end, inspired by Du et al. (2024), the criteria include a penalty for unhelpful feedback patterns (e.g., neglect of key elements, unsupported claims, or contradictory advice).

Refinement of criteria and rubric for review assessment As with the exposé rubrics, all criteria and rubrics were reviewed by instructors and piloted on sample reviews before course deployment.

	WS 2024/25	WS 2025/26	
Rubric	*Group 1*	*Control*	*All*
Exposé drafts ($n = 55$, 78, and 162, respectively)			
Language / Quality	2.69 ($\pm$1.03)	3.22 ($\pm$0.80)	3.33 ($\pm$0.75)
Metadata	2.75 ($\pm$0.52)	2.73 ($\pm$0.45)	2.73 ($\pm$0.44)
Form / Structure	1.96 ($\pm$0.19)	1.92 ($\pm$0.27)	1.94 ($\pm$0.24)
Motivation	6.93 ($\pm$1.48)	7.81 ($\pm$1.46)	7.84 ($\pm$1.41)
Approach	7.91 ($\pm$2.27)	8.90 ($\pm$2.57)	8.88 ($\pm$2.38)
Methodology	4.07 ($\pm$1.33)	4.05 ($\pm$1.48)	4.01 ($\pm$1.34)
Schedule	4.15 ($\pm$0.85)	4.29 ($\pm$1.34)	4.46 ($\pm$1.07)
Bibliography	2.56 ($\pm$0.86)	2.44 ($\pm$0.86)	2.47 ($\pm$0.78)
Additional points	-0.02 ($\pm$0.36)	-0.22 ($\pm$0.64)	-0.18 ($\pm$0.58)
Total	**33.00 ($\pm$6.20)**	**35.14 ($\pm$6.74)**	**35.49 ($\pm$5.91)**
Exposé finals ($n = 55$ and 162, respectively)			
Language / Quality	3.45 ($\pm$0.79)	3.75 ($\pm$0.50)	
Metadata	2.95 ($\pm$0.23)	2.85 ($\pm$0.36)	
Form / Structure	1.96 ($\pm$0.19)	1.98 ($\pm$0.16)	
Motivation	8.09 ($\pm$1.02)	8.52 ($\pm$0.91)	
Approach	9.42 ($\pm$2.07)	10.20 ($\pm$1.44)	
Methodology	4.76 ($\pm$0.58)	4.72 ($\pm$0.74)	
Schedule	4.65 ($\pm$0.55)	4.81 ($\pm$0.48)	
Bibliography	2.91 ($\pm$0.29)	2.75 ($\pm$0.51)	
Additional points	0.00 ($\pm$0.27)	-0.05 ($\pm$0.44)	
Total	**38.20 ($\pm$3.87)**	**39.52 ($\pm$3.24)**	
Reviews ($n = 112$ and 325, respectively)			
Structure and Clarity	3.26 ($\pm$1.04)	3.67 ($\pm$0.73)	
Language	2.82 ($\pm$0.47)	2.79 ($\pm$0.46)	
Feed Up	2.37 ($\pm$1.49)	1.59 ($\pm$1.62)	
Feed Back	3.96 ($\pm$0.31)	3.93 ($\pm$0.35)	
Feed Forward	1.74 ($\pm$0.50)	1.14 ($\pm$0.73)	
Content Quality	4.51 ($\pm$0.79)	4.46 ($\pm$0.88)	
Errors	3.71 ($\pm$0.62)	3.78 ($\pm$0.60)	
Additional points	-0.04 ($\pm$0.19)	-0.08 ($\pm$0.29)	
Total	**22.33 ($\pm$3.46)**	**21.27 ($\pm$3.48)**	

Table 7: Per-rubric scores by cohort and assessment stage. Values are means ($\pm$SD). *Control* includes gradings by graders who did not see the LLM pre-assessment, whereas *All* also includes LLM-assisted draft gradings, whose scores may be anchored toward the model proposal.

Code	Criterion	Expert level	Point descriptors
Rubric: Language/Quality (LA) *Assessment of language quality and structural coherence of the text.*			
LA1	Language quality	1	**0:** many linguistic errors (e.g. wrong tense, wrong personal pronouns) **1:** few linguistic errors, clear and comprehensible sentences **2:** no linguistic errors, clear and comprehensible sentences with consistent terminology
LA2	Common Thread	1	**0:** No recognizable common thread, the structure is unclear and jumpy **1:** Partly recognizable, but the structure is not continuous **2:** Text has a continuous structure that is easy to follow, common thread is clear and recognizable from beginning to end
Rubric: Metadata (ME) *Evaluation of title appropriateness and creativity.*			
ME1	Preliminary title	1	**0:** Title does not match the topic **1:** Title fits the topic; similar title to the one in the topic suggestions **2:** Title fits the topic; creative title changed in a novel way and/or fits into the story of the exposé
ME2	Metadata available	0	**0:** Name, date and matriculation number are not entered correctly/changed **1:** Name, date, matriculation number are correct and updated
Rubric: Form/Structure (ST) *Evaluation of document formatting and structural requirements.*			
ST1	Template used	0	**0:** Latex template is not used **1:** Latex template is used
ST2	Number of pages	0	**0:** > three pages (motivation + approach only) **1:** <= three sides (motivation + approach only)
Rubric: Motivation (MO) *Assessment of problem identification, context establishment, and research question formulation.*			
MO1	Hook existing	1	**0:** Hook not available **1:** Hook present and suitable for the topic
MO2	Anker	1	**0:** The problem is not mentioned at all. **1:** The problem is stated in general terms without specific details. It is recognizable which problem is being addressed, but important information is omitted. **2:** The problem is described in detail. All relevant aspects of the problem are mentioned and clarify its scope and possible effects.
MO3	Domain	1	**0:** The domain or context in which the problem occurs is not mentioned.

(continued on next page)

Code	Criterion	Expert level	Point descriptors
			1: The domain or the relevant expert area of the problem is mentioned. It is clear in which environment the problem exists.
MO4	Problem relevance	1	**0:** The relevance or importance of the problem is not addressed. It remains unclear why this problem is relevant or worth discussing. **1:** The importance and relevance of the problem is described. It is made clear why the problem is important and what impact or consequences it could have (e.g. for specific individuals, organisations or society as a whole).
MO5	Problem handling	1	**0:** current handling of the problem or current criticism is not given **1:** current handling of the problem or current criticism is given
MO6	Teaser (RQ)	2	**0:** There is no research question available **1:** A research question exists, but it has weaknesses, e.g. due to unclear terminology or vague formulation. It is difficult to recognise what is to be investigated and how the question fits the problem description. **2:** The research question is formulated clearly and precisely. It clearly conveys what is to be investigated, and fits thematically with the problem description.
MO7	RQ Limitation	2	**0:** The research question is either too general or too broad, so that it does not seem realistic to answer it within the scope of a Bachelor's thesis, to answer it in the context of a Bachelor's thesis. **1:** The research question is precise and clearly defined so that it can realistically be addressed in the context of a Bachelor's thesis. The question is formulated in concrete terms and describes specifically which aspects are to be investigated without deviating from the topic.
Rubric: Approach (AP) *Assessment of state of the art analysis, theoretical framework.*			
AP1	SOTA	1	**0:** SOTA not available **1:** SOTA present and correctly cited
AP2	SOTA Relevance	1	**0:** The relevance of SOTA is not present or unclear. **1:** The relevance of SOTA is mentioned, but is only partially comprehensible or weakly presented. **2:** It is clearly shown how SOTA is relevant to the topic of the exposé and why it is relevant to the research.
AP3	SOTA Weaknesses	1	**0:** SOTA's weak points are not mentioned **1:** SOTA's weaknesses are mentioned, but only superficially or unclearly described **2:** SOTA's weaknesses are clearly identified and precisely explained; it is clear which weaknesses are relevant for own research

(continued on next page)

Code	Criterion	Expert level	Point descriptors
AP4	SOTA Delimitation	1	**0:** No differentiation of own performance from the current state of research **1:** The own achievement is differentiated from SOTA; it is explicitly emphasized, what is new or unique about your own research
AP5	SOTA Combination	2	**0:** No meaningful combination of existing material (even if reference is made to existing material, but its combination or consolidation is unconvincing or unclear) **1:** Existing material is clearly and meaningfully combined; the added value of the combination is clear.
AP6	Theoretical Framework	1	**0:** The theoretical framework does not exist or is completely misleading **1:** The theoretical framework is present, but unclear, poorly structured or difficult to understand **2:** The theoretical framework is clearly and logically structured. The theoretical approaches are presented in a understandable and comprehensible
AP7	Relevance of theoretical framework	2	**0:** The theoretical framework shows no connection to the research topic, or the theory is irrelevant **1:** The theoretical framework clearly demonstrates how the selected theories directly relate to the research topic and support the research question
AP8	Methodology Availability	1	**0:** Methodology not available **1:** Methodology available and suitable for answering the RQ(s)
Rubric: Methodology (MD) *Assessment of the methodology.*			
MD1	Methodology Completeness	1	**0:** Methodology is available, but not all RQ(s) are addressed **1:** Methodology is available and addresses all points of the research questions
MD2	Methodology Relevance	1	**0:** The relevance of the methodology for the research project is not clear or is unclear **1:** It is clearly and precisely demonstrated why the chosen methodology is relevant to the research project and how it contributes to the research questions
MD3	Methodology Target group	1	**0:** The target group is not specified or there is no precise differentiation **1:** The target group is clearly and precisely defined and it is clear why it is relevant to the methodology
MD4	Methodology Existing Material	1	**0:** No reference is made to the existing material or the material is inappropriate **1:** The existing material is appropriately and accurately related to the topic; it is clear that this material is relevant to the methodology

(continued on next page)

Code	Criterion	Expert level	Point descriptors
MD5	Methodology Difficulties	2	**0:** Potential difficulties in the practical implementation of the methods are not addressed. Potential challenges that could hinder the research process are overlooked. **1:** Potential difficulties in the practical implementation of the methods are described clearly and precisely. Concrete solutions or strategies to overcome these difficulties are also presented, in order to organize the research process as efficiently and target-oriented as possible.
MD6	Methodology Possibilities/restrictions	2	**0:** The possibilities or limitations of the methods used are not addressed, or they are only mentioned superficially, without concrete details. It remains unclear how these methods contribute to answering the research question and which limitations could possibly influence the validity of the results. **1:** The possibilities and limitations of the methods are described clearly and precisely. It is clear what strengths the methods bring to the study and what weaknesses or limitations could possibly influence the results. limitations could possibly influence the results. The choice of methods is critically reflected upon and related to the research project.
MD7	Methodology Details	2	**0:** No additional explanations are given on the methodology or implementation **1:** The methodology is explained comprehensively and precisely with additional explanations (e.g. details on implementation), so that the procedure is clearly comprehensible
Rubric: Schedule (SC) *Assessment of project timeline structure, completeness, and realism.*			
SC1	Schedule Availability	0	**0:** Tabular schedule not available **1:** Tabular schedule in chronological order available
SC2	Schedule Completeness	1	**0:** Tabular schedule contains gaps **1:** Tabular schedule available from start to finish and reasonably divided into blocks
SC3	Schedule Block Description	0	**0:** not available or only headlines available **1:** individual blocks described (not only headlines)
SC4	Schedule Realistic Relevance	1	**0:** The time distribution is obviously unrealistic and important work steps are missing. The research questions are not or only insufficiently addressed **1:** The timeline contains the most important steps, but some of the timelines are unrealistic or unclear. It is partially unclear how the steps contribute to answering the research questions. The timeline does not appear to be fully aligned with the Bachelor's thesis context.

(continued on next page)

Code	Criterion	Expert level	Point descriptors
			2: The timeline is clearly structured, realistic and aligned with the scope of the Bachelor's thesis. All important work steps are included and organised in a sensible time frame. Each step contributes to answering the research questions, and the entire plan fits coherently into the exposé and the context of the thesis.
Rubric: Bibliography (BI) *Assessment of bibliography consistency and literature relevance.*			
BI1	Bibliography Consistency	0	**0:** Bibliography with different citation styles **1:** Bibliography is standardized (consistent citation style) and essentially all information on the literature is available
BI2	Key literature	2	**0:** Literature does not fit the topic **1:** Literature fits the topic **2:** Literature fits the topic and most of it is highly relevant
Rubric: Additional points (AD) *Allow additional points for very good submissions and major mistakes in the submission.*			
AD1	Additional points	1	**0:** No additional strengths beyond existing criteria. **1:** Some additional strengths beyond existing criteria. **2:** Clear exceptional strengths beyond existing criteria.
AD2	Negative points	1	**0:** No major issues beyond existing criteria. **-1:** Notable issues beyond existing criteria. **-2:** Serious issues that strongly reduce overall quality.

Table 8: Guidelines for exposé grading. 36 criteria used to grade student exposés. We organize the criteria into nine rubrics, by grouping together criteria that evaluate related aspects of exposé writing.

Code	Criterion	Expert	Point descriptors
Rubric: Structure and Clarity (SC) *Assessment of overall structure and clarity.*			
SC1	Summary available	1	**0:** No summary available or, the summary does not reflect understanding of the performance **1:** Simple summary that demonstrates basic understanding of the performance **2:** Concise and accurate summary that clearly reflects understanding of performance and builds confidence in feedback
SC2	Clarity	2	**0:** Feedback needs considerable improvement in clarity and comprehensibility, comments are disjointed or difficult to follow **1:** Feedback is largely clear, but sometimes lacks specific references to particular lines, pages, sections or figures/tables **2:** Feedback is clear, well-structured and easy to understand, with coherent comments and precise references to specific points in the text
Rubric: Language (LG) *Assessment of language quality and tone.*			
LG1	Language	1	**0:** Many linguistic errors (e.g. wrong tense, wrong personal pronouns) **1:** Almost no linguistic errors, clear and comprehensible sentences with standardised terminology
LG2	Tone	1	**0:** The tone is inappropriate, sounds judgemental or reproachful, criticism comes across as destructive or personal **1:** The tone is generally respectful, avoids personal attacks, but remains unclear or less friendly in places **2:** The tone is consistently calm, respectful and polite; criticism is constructive, factual and only directed at the text, not the person
Rubric: Feed Up (FU) *Assessment of learning goals and success criteria.*			
FU1	Learning Goals	2	**0:** No learning objectives, unclear what is to be achieved, no orientation **1:** Partly unclear learning objectives, generally formulated, offer only limited orientation **2:** Clear learning objectives, specific, challenging, comparable and provide clear direction
FU2	Success Criteria	1	**0:** No success criteria, success unclear or vaguely formulated, objectives not defined **1:** Partly unclear success criteria, available but imprecise or general, orientation partly missing **2:** Clear success criteria, concrete, measurable goals and clear definition of success
Rubric: Feed Back (FB) *Assessment of feedback quality and knowledge transfer.*			

(continued on next page)

Code	Criterion	Expert	Point descriptors
FB1	Knowledge of result	1	**0:** No information as to whether the task has been solved or not **1:** Information given as to whether the task has been solved or not
FB2	Knowledge of correct response	2	**0:** No information on the correct answer **1:** Correct answer is clearly communicated
FB3	Knowledge about task constraints	1	**0:** No information on rules, requirements or restrictions of the task **1:** Clear information on rules, requirements or restrictions of the task (e.g. specifications)
FB4	Knowledge about concepts	1	**0:** No information on conceptual knowledge **1:** Basic information on relevant concepts is provided
FB5	Knowledge about mistakes	1	**0:** No information about errors **1:** Information on the number, location, source or type of errors
FB6	Self-feedback	1	**0:** Self-feedback is available that has not been combined with task, process or self-regulation (including feedback on the learner's abilities (i.e. intelligence or talent)) **1:** No self-feedback is available or, if available, in combination with task, process or self-regulation
Rubric: Feed Forward (FF) *Assessment of future learning guidance.*			
FF1	Self-regulation	1	**0:** No indication of how to approach the task, no recognisable approaches to self-monitoring or self-assessment **1:** Either indications of strategies for working on the task and recognising errors (e.g. suggestions for improvement, targeted search for information) or approaches to self-monitoring and self-assessment (e.g. reflection on own understanding, strategies or willingness to seek help)
FF2	Learning Skills	1	**0:** No information on how to develop or improve learning skills **1:** Information on how to improve the learning objective and how to deal with challenges
Rubric: Content Quality (CQ) *Assessment of content accuracy and usefulness.*			
CQ1	Correctness	2	**0:** The content has significant deficiencies in terms of correctness **1:** The content is largely correct **2:** The content is completely correct
CQ2	Actionability	1	**0:** No specific instructions available, feedback is difficult to implement **1:** Some instructions available, but only partially clear or specific

(continued on next page)

Code	Criterion	Expert	Point descriptors
			2: Clear, specific feedback with actionable instructions, e.g. comparisons with existing methods, application to other tasks, definitions, explanations, discussions, additional analyses or specific suggestions to remove confusion
CQ3	Argumentation	1	**0:** Statements or conclusions are not supported by evidence or comprehensible explanations. The argumentation remains superficial or unsystematic, making the content unconvincing. **1:** Statements are supported by comprehensible evidence, examples or explanations. The argumentation shows a clear link between evidence and conclusions, which strengthens the quality of the content.
Rubric: Errors (ER) *Assessment of potential errors and issues in feedback.*			
ER1	Neglect	2	**0:** All important details have been considered and no unnecessary questions are asked **-1:** Important details have been overlooked, leading to unnecessary questions or criticism
ER2	Vague Critic	1	**0:** Criticism is specific and clearly states what is missing or can be improved **-1:** Unspecific criticism, mentions missing components without clearly stating what is missing
ER3	Out-of-Scope	1	**0:** Proposals remain within the scope of the task and are relevant **-1:** Suggestions refer to methods or analyses that go beyond the intended scope
ER4	Missing Reference	1	**0:** Alternative methods, etc. (if available) are supported with justification or references **-1:** Alternative methods, etc. (if available) are suggested without justification or references
ER5	Contradiction	1	**0:** Feedback is consistent in itself, no contradictory statements **-1:** Contradictory feedback, e.g. criticism and praise for the same methods
Rubric: Additional points (AD) *Allow additional points for very good reviews or deduct points for major mistakes.*			
AD1	Additional points	1	**0:** No additional strengths identified. **1:** Notable additional strengths beyond other criteria.
AD2	Negative points	1	**0:** No additional major issues identified. **-1:** Significant issues beyond other criteria.

Table 9: Guidelines for review grading. 24 criteria used to grade student reviews. We organize the criteria into eight rubrics, by grouping together criteria that evaluate related aspects of review.

F Summary of Collected Sources for Exposé Criteria

Across the full corpus of 110 exposé guidelines we reviewed, we identified materials from universities and higher-education institutions in Germany, Austria, and Switzerland, complemented by student-facing writing portals (e.g., Scribbr, Finito24, Studieren.de), which we further investigated. In terms of disciplinary scope, our primary sources are published by faculties and units in the humanities and education, the social sciences (including sociology and political science), medicine, economics and business, as well as applied/STEM fields (i.e., architecture, computer science, and civil engineering). Roughly three-quarters of the sources are published in English, primarily from internationalized programs (e.g., TU Berlin, Universität Wien, TU Dresden, Uni Magdeburg), while the remainder appear in German and serve national undergraduate or graduate audiences. About one-fifth of the documents explicitly target Bachelor- or Master-level theses, and a smaller subset focuses on doctoral or dissertation exposés, such as those from the University of Leipzig Graduate Centre and the University of Magdeburg Medical Faculty. Only a few pages, most notably the TU Berlin departmental guide, explicitly include project or grant proposal contexts.

Thematically, nearly all guidelines converge on the same structural expectations: a clear *Motivation* or problem statement, a theoretically informed *Approach* or methodology, a feasible *Schedule* or timeline, and a complete *Bibliography*. Additional sections, such as *Research Questions*, *Related Work*, and *Formal Aspects* (including format, citation, or layout), occur frequently, especially in English-language handbooks and institutional templates.

Across this heterogeneity, the documents show remarkable consistency in pedagogical intent: *to ensure students articulate a focused question, justify methodological choices, and plan realistic execution.* These recurrent patterns directly informed the rubric domains of *Motivation*, *Approach*, *Schedule*, and *Bibliography* used for annotation and model benchmarking in our dataset.

Links for Table 10:

1. https://www.tu.berlin/en/ak/study-and-teaching/final-theses/writing-an-expose
2. https://www.oei.fu-berlin.de/soziologie/studiumlehre/handreichungen_zum_studium/How-to-write-an-expose.pdf
3. https://home.uni-leipzig.de/~gsgas/fileadmin/Recommendations/Expose_Recommendations.pdf
4. https://expae.med.ovgu.de/unimagdeburg_mm/Downloads/Kliniken/KPAE/EXPAE/Guidelines+for+writing+an+expos%C3%A9+for+a+doctoral+project.pdf
5. https://www.wvf.uni-freiburg.de/promhabil/merkblatt-erstellung-expose
6. https://www.h2.de/fileadmin/user_upload/Promotionszentren/PromZ_SGW/4eng_Expose-Hinweise_PromZ_SGW.pdf
7. https://www.archland.uni-hannover.de/fileadmin/archland/FORSCHUNG/promotionsausbildung/Requirements_admission_and_expose__.pdf
8. https://tu-dresden.de/ing/informatik/smt/im/studium/theses-and-research-projects/writing-an-expose
9. https://www.hfp.tum.de/fileadmin/w00cjd/governance/pdf/Instructions_for_Preparing_an_Expose.pdf
10. https://www.gsi.uni-muenchen.de/personen/wiss_mitarbeiter/kruck/general_guidelines.pdf
11. https://sts.univie.ac.at/fileadmin/user_upload/i_sts/Studium/Master_STS/05_Services_for_current_students/Master_Thesis/Guidelines_for_Writing_a_Master_Thesis_Expose.pdf
12. https://www.uni-kassel.de/fb11agrar/index.php?eID=dumpFile&t=f&f=572&token=eb5a0183632b6762477026bbd8f1694660f1b3f6
13. https://www.bwl.uni-mannheim.de/media/Lehrstuehle/bwl/Schoen/Teaching/Master/Expose_Guide/Master_Thesis_Expose_Guide_v6.pdf
14. https://www.fh-erfurt.de/fileadmin/Bilder/Seitenbereiche/Fakultaeten/BKR/B/Studium/MA_Sustainable_Engineering_of_Infrastructure/Expose_Recommendations_SEI.pdf
15. https://www.scribbr.de/anfang-abschlussarbeit/expose-bachelorarbeit/
16. https://www.wu.ac.at/fileadmin/wu/d/i/ign/IGN_BA-Expos%C3%A9_Requirements_De.pdf
17. https://www.finito24.de/expose-schreiben/
18. https://www.uni-bremen.de/fileadmin/user_upload/fachbereiche/fb7/gscm/Dokumente/Structure_of_an_Expose.pdf

G Exposía Dataset Statistics

This section provides additional descriptive statistics for Exposía that complement the summary in Section 5 of the main paper. To give a more detailed view of the dataset's composition, we report basic statistics on the exposés, reviews, and assessment scores.

Source	Lang	BA	MA	Diss.	Grant/Proj.	Link
TU Berlin	EN	✓	✓	✓	✓	[1]
FU Berlin	EN	✗	✓	✓	✗	[2]
Uni Leipzig	EN	✗	✗	✓	✗	[3]
Uni Magdeburg	EN	✗	✗	✓	✗	[4]
Uni Freiburg	DE/EN	✗	✗	✓	✗	[5]
Hochschule Magdeburg	EN	✗	✗	✓	✗	[6]
Uni Hannover	EN	✗	✗	✓	✗	[7]
TU Dresden	EN	✓	✓	✗	✗	[8]
Hochschule für Politik München	EN	✓	✓	✗	✗	[9]
LMU München	EN	✓	✓	✗	✗	[10]
Uni Wien	EN	✗	✓	✗	✗	[11]
Uni Kassel	EN	✗	✓	✗	✗	[12]
Uni Mannheim	EN	✗	✓	✗	✗	[13]
FH Erfurt	EN	✗	✓	✗	✗	[14]
Scribbr (DE)	DE	✓	✗	✗	✗	[15]
WU Wien	DE	✓	✗	✗	✗	[16]
Finito24	DE	✓	✗	✗	✗	[17]
Uni Bremen	EN	✗	✗	✗	✗	[18]

Table 10: University and institutional exposé guidelines with explicit scope for Bachelor, Master, Dissertation, or Grant/Project proposals.

G.1 Exposé Revision Statistics

Table 11 reports section-wise text statistics for *Motivation*, *Approach*, and *Schedule* in draft vs. final versions (mean ± SD). Raw LaTeX is converted to plain text with `pylatexenc`[14]; sentence/word segmentation uses spaCy. We report **length/structure** (#Words, #Sentences and average word length), **lexical/readability** (type–token ratio (TTR) and Flesch Reading Ease (FRE); Flesch, 1948 via `textstat`[15]), and **draft–final similarity** (cosine similarity of SentenceTransformer embeddings[16] and normalized token-level Levenshtein similarity; Levenshtein, 1966).

Across sections, final exposés are longer in words and sentences, with the largest expansion in *Approach*; average word length changes only marginally. Lexical diversity (TTR) decreases slightly from draft to final in all three sections. The negative FRE values for *Schedule* likely reflect formatting artifacts (fragmentary, table-like content) interacting with plain-text conversion and segmentation.

[14] https://github.com/phfaist/pylatexenc
[15] https://github.com/textstat/textstat
[16] sentence-transformers/all-mpnet-base-v2

Similarity and topic structure High embedding cosine similarities in Table 11 indicate strong semantic continuity between draft and final, even where token-level edit similarity is lower (notably for *Motivation* and *Approach*). Table 12 reports pairwise embedding cosine similarities within and across topics: within-topic pairs are generally more similar than cross-topic pairs, with the largest within-topic variance for *Generative AI in Healthcare*.

Bibliography size and stability Table 13 reports topic-wise bibliography statistics for draft vs. final exposés (mean ± SD), measured as the number of distinct BibTeX entry keys and their draft–final overlap (Jaccard similarity over key sets). Draft bibliography size differs substantially by topic (T1 < T2 < T3). Despite these initial gaps, revisions show a consistent pattern across topics: authors add about two references on average while removing fewer than one, suggesting modest expansion rather than wholesale replacement. Correspondingly, mean draft–final Jaccard similarity remains relatively high (overall 0.78), indicating that bibliographies are generally stable across revisions with targeted additions.

Metric	Motivation		Approach		Schedule	
	Draft	Final	Draft	Final	Draft	Final
Length and structure						
#Words	402 (±133)	428 (±125)	748 (±258)	851 (±221)	248 (±83)	270 (±131)
#Sentences	21.8 (±8.5)	24.0 (±7.8)	40.9 (±18.4)	47.7 (±15.3)	7.3 (±8.3)	8.4 (±8.7)
Avg. word length	5.38 (±0.45)	5.43 (±0.44)	5.41 (±0.49)	5.46 (±0.50)	5.55 (±0.37)	5.50 (±0.36)
Lexical and readability						
Lexical diversity	0.54 (±0.06)	0.53 (±0.06)	0.45 (±0.07)	0.43 (±0.06)	0.60 (±0.06)	0.59 (±0.06)
Flesch reading ease	43.0 (±14.5)	41.6 (±13.3)	42.1 (±15.3)	41.0 (±15.9)	-50.4 (±90.4)	-34.2 (±90.1)
Draft–Final similarity						
Cosine similarity	0.96 (±0.04)		0.95 (±0.06)		0.93 (±0.14)	
Levenshtein similarity	0.64 (±0.24)		0.62 (±0.26)		0.81 (±0.24)	

Table 11: Chapter-level textual statistics for exposés by version (draft vs. final). Entries show mean (±SD) for length, lexical diversity, readability, and draft–final similarity metrics.

Topic	Motivation		Approach		Schedule	
	Draft	Final	Draft	Final	Draft	Final
T1	0.59 (±0.13)	0.58 (±0.13)	0.55 (±0.14)	0.56 (±0.14)	0.52 (±0.16)	0.52 (±0.16)
T2	0.60 (±0.12)	0.66 (±0.09)	0.60 (±0.11)	0.62 (±0.12)	0.59 (±0.13)	0.61 (±0.12)
T3	0.48 (±0.14)	0.47 (±0.14)	0.40 (±0.17)	0.41 (±0.16)	0.54 (±0.13)	0.50 (±0.11)
Overall	0.37 (±0.18)	0.38 (±0.18)	0.35 (±0.18)	0.35 (±0.19)	0.48 (±0.16)	0.46 (±0.15)

Table 12: Pairwise cosine similarity between exposés (each exposé is compared with all others within the same topic and version). Entries show mean (±SD) over all unordered document pairs, computed separately per chapter. Topic mapping: T1=*The Holographic AI Assistant*, T2=*Interactive LLM-based teaching agents*, T3=*Generative AI in Healthcare*.

Draft–final improvements on rubric scores Table 14 reports per-rubric statistics for draft and final exposés. It complements Section 5.1 by additionally showing (i) the remaining gap between the mean final score and the rubric maximum and (ii) the observed score ranges (min–max) for each version. As in the main text (Section 5.1), the largest mean gains occur in conceptually central rubrics such as *Approach* and *Motivation*. In contrast, smaller changes in *Metadata*, *Bibliography*, and *Form/Structure* largely reflect limited headroom: these rubrics have low maxima, and draft scores are already close to the ceiling. Finally, *Additional points* changes little because it represents bounded instructor adjustments rather than points that students can systematically earn through revision (see Appendix G.3 for the corresponding scoring rules). Overall, the revision signal is therefore most visible in rubrics with substantial remaining headroom, and is concentrated in substantive content and argumentation.

G.2 Review Statistics

Table 15 reports text statistics for review documents by author role (mean ± SD), computed analogously to the exposé statistics. Student reviews are substantially longer than instructor reviews, while instructors show higher lexical diversity.

Instructor roles We distinguish between *junior* and *senior* instructors based on their academic stage and experience in reviewing and research. Junior instructors are recruited Master's students who have completed a Bachelor's thesis and thus have basic academic writing experience, whereas senior instructors are course staff (PhD researchers and a postdoc) with substantially more experience in academic peer review and supervision; see Appendix H.1 for details on rater composition and training.

Inline comment distributions Table 16 summarizes comment volume and length by role: junior

Topic	Draft $\bar{n}_{\text{draft}}$	Final $\bar{n}_{\text{final}}$	$\bar{n}_{\text{added}}$	$\bar{n}_{\text{removed}}$	Similarity Jaccard
T1	6.74 ± 6.02	8.63 ± 5.63	2.16 ± 3.86	0.26 ± 0.56	0.73 ± 0.40
T2	9.22 ± 6.39	10.78 ± 6.31	2.06 ± 2.62	0.50 ± 1.25	0.76 ± 0.27
T3	11.83 ± 8.01	13.06 ± 8.74	1.89 ± 2.97	0.67 ± 2.83	0.84 ± 0.26
Overall	9.22 ± 7.04	10.78 ± 7.11	2.04 ± 3.15	0.47 ± 1.77	0.78 ± 0.32

Table 13: Bibliography size and draft–final changes by topic. We report mean (±SD) number of entries in draft and final versions, average additions/removals, and mean draft–final entry-set Jaccard similarity computed over BibTEX keys. Topic mapping: T1=*The Holographic AI Assistant*, T2=*Interactive LLM-based Teaching Agents*, T3=*Generative AI in Healthcare*.

Rubric	Mean (±SD) Draft	Final	Δ	Gaps Gap	Max	Observed bounds Draft	Final
Approach	7.91 (±2.25)	9.40 (±2.06)	1.49 (±1.74)	1.60	11.00	1.00–11.00	1.00–11.00
Motivation	6.93 (±1.46)	8.09 (±1.01)	1.16 (±1.17)	0.91	9.00	4.00–9.00	6.00–9.00
Language/Quality	2.69 (±1.02)	3.45 (±0.78)	0.76 (±0.81)	0.55	4.00	0.00–4.00	1.00–4.00
Methodology	4.07 (±1.32)	4.76 (±0.57)	0.69 (±1.08)	0.24	5.00	0.00–5.00	3.00–5.00
Schedule	4.15 (±0.84)	4.65 (±0.55)	0.51 (±0.83)	0.35	5.00	0.00–5.00	3.00–5.00
Bibliography	2.56 (±0.85)	2.91 (±0.29)	0.35 (±0.77)	0.09	3.00	0.00–3.00	2.00–3.00
Metadata	2.75 (±0.51)	2.95 (±0.23)	0.20 (±0.48)	0.05	3.00	1.00–3.00	2.00–3.00
Additional points	-0.02 (±0.36)	0.00 (±0.27)	0.02 (±0.30)	2.00	2.00	-2.00–1.00	-1.00–1.00
Form/Structure	1.96 (±0.19)	1.96 (±0.19)	0.00 (±0.19)	0.04	2.00	1.00–2.00	1.00–2.00

Table 14: Per-rubric statistics for draft and final exposé scores. We report mean (±SD), average remaining gap to the rubric maximum, and observed score bounds (min–max) for each version.

instructors contribute the most comments overall, while students and senior instructors write longer comments on average. Table 17 reports the distribution of span-level tags. Totals are similar but not identical across the two tables because span annotations are only recorded for comments that are attached to a specific text span; page-level comments have no associated span annotation. Across roles, *Weakness* is the most frequent tag, followed by *Strength*, *Other*, and *Highlight*.

G.3 Assessment scores coverage

Table 18 summarizes the number of assessments that include a complete set of criterion-based scores for exposé drafts and final submissions, by rater group. Draft exposés were graded by instructors during the semester (Group 1) and re-scored after the course by Group 2 to compute inter-annotator agreement (Section 5.2). Final exposés, in contrast, were not re-scored post-semester: due to the substantial time required for criterion-based grading, Group 2 focused on a draft subset rather than duplicating the full set of final-submission grades. A similar pattern holds for student review grading (Table 19): reviews were graded during the semester and selectively re-scored for agreement analyses, but criterion-based scores are only available for student-written reviews (not for instructor-written reviews).

Annotation effort We emphasize that constructing this dataset is unusually labor-intensive. In Group 1, producing a full review with inline comments and criterion-based scores typically required on the order of 2–4 hours early in the semester, decreasing to roughly 1–2 hours as raters gained experience. Criterion-based grading itself is also high effort: grading a single exposé typically takes about 1–2 hours, while grading a single review requires 20–40 minutes. Despite these restrictions, the dataset currently comprises 55 exposés and 306 reviews and includes unusually rich annotation artifacts: inline comments and comment tags, free-text reviews, draft and final exposé versions, and hundreds of instructor assessment scores across 36 exposé and 24 review criteria, which we use throughout our analyses.

Metric	Review author role		
	Student	Junior Instructor	Senior Instructor
Length and structure			
#Words	680 (±455)	285 (±134)	347 (±116)
#Sentences	33.7 (±21.3)	16.7 (±8.4)	18.7 (±6.3)
Avg. word length	5.31 (±0.44)	5.16 (±0.28)	5.18 (±0.49)
Lexical and readability			
Lexical diversity	0.46 (±0.08)	0.56 (±0.09)	0.55 (±0.08)
Flesch reading ease	50.2 (±13.3)	58.9 (±9.9)	54.3 (±14.2)

Table 15: Textual statistics of review texts by author role. Entries show mean (± SD).

Role	Comments	Avg chars	Avg words
St. (n=31)	640	77.0	12.6
Jr. (n=7)	1121	53.8	9.1
Sr. (n=5)	492	68.8	12.0
Σ (n=43)	**2253**	**63.6**	**10.7**

Table 16: Basic statistics of comments by role, including total counts and average length in characters and words. St.=Students, Jr.=Junior Instructors, Sr.=Senior Instructors.

Scoring and special rubric behavior Within each submission (draft exposé, final exposé, peer feedback), criterion-level scores are summed to yield rubric points, and rubric points are summed to obtain total submission points. Most rubrics use simple addition; we introduce a small number of rubric-specific rules to maintain consistency across topics and feedback situations. The exposé rubric *Methodology* is capped at 5 points, and the review rubric *Feedback* is capped at 4 points. For example, in the exposé rubric, *Existing Material* may be inapplicable when the proposed methodology relies on newly collected primary data rather than reusing an existing dataset; the cap ensures that full credit remains attainable by satisfying an appropriate subset of criteria. The review rubric *Errors* employs a deduction-only scheme, starting at 4 points and decreasing by 1 point for each identified error, with a minimum score of 0. Finally, the instructor-only rubric *Additional points* permits bounded adjustments in exceptional cases (Exposé: $[-2, +2]$, Review: $[-1, +1]$); these adjustments do not increase the maximum attainable points.

H Inter-annotator agreement

This section documents the computation of inter-annotator agreement (IAA), complementing the criterion-based scoring description in Section 5.2 for Exposía. It also provides the reference point for interpreting LLM–human agreement in the main experiments. Because our criterion scores are discrete and ordinal, we prioritize agreement measures that account for ordered categories and report distributional diagnostics for criteria affected by prevalence skew.

H.1 Raters, training, and scoring design

Raters and training Twelve raters participated and were organized into two groups. In total, 7 were *junior instructors* (student helpers) and 5 were *senior instructors* (4 PhD researchers and 1 postdoc). One PhD researcher participated in both groups.

We refer to instructors who graded during the semester as *Group 1*. They attended weekly sessions and discussed edge cases. Group 1 raters were instructed to resolve uncertainty in favor of the student, because grades were at stake. *Group 2* includes a different subset of instructors (with one who also participated in Group 1). They received the same guidelines, but they were not instructed to resolve uncertainty in favor of the student.

Group 1 (in-semester): 6 student helpers, 2 PhD researchers, 1 postdoc; weekly calibration (jointly graded initial exposés, then addressed edge cases). Because grades were at stake, Group 1 was instructed to resolve uncertainty in favor of the student.

Group 2 (post-semester): 1 student helper, 3 PhD researchers; scored subsets of draft exposés and

Tag	Metric	St. (N=31)	Jr. (N=7)	Sr. (N=5)	Σ
Highlight	Raw count	90	28	58	**176**
	Avg. length	112.68 (±188.62)	122.79 (±149.17)	168.34 (±227.29)	132.63 (±198.48)
Other	Raw count	107	145	52	**304**
	Avg. length	120.66 (±236.88)	41.77 (±61.91)	90.83 (±132.55)	77.93 (±160.87)
Strength	Raw count	131	200	73	**404**
	Avg. length	254.63 (±263.11)	178.40 (±239.87)	196.52 (±285.21)	206.39 (±258.44)
Weakness	Raw count	308	691	290	**1289**
	Avg. length	79.20 (±106.17)	59.17 (±127.02)	97.88 (±154.42)	72.66 (±130.22)
Σ	**Raw count**	**636**	**1064**	**473**	**2173**

Table 17: Distribution of span-level annotation tags by role. We report raw counts and average span length (±SD). St.=Students, Jr.=Junior Instructors, Sr.=Senior Instructors.

	Group 1			Group 2		
Version / Topic	Jr.	Sr.	Σ	Jr.	Sr.	Σ
Draft (N = 55)						
T1 (N=19)	14	5	**19**	2	17	**19**
T2 (N=18)	14	4	**18**	4	14	**18**
T3 (N=18)	15	3	**18**	2	16	**18**
Σ **Draft**	**43**	**12**	**55**	**8**	**47**	**55**
Final (N = 55)						
T1 (N=19)	15	4	**19**	0	0	**0**
T2 (N=18)	14	4	**18**	0	0	**0**
T3 (N=18)	15	3	**18**	0	0	**0**
Σ **Final**	**44**	**11**	**55**	**0**	**0**	**0**

Table 18: Distribution of criterion-based exposé assessments across topics (T1–T3), rater roles (Jr. Junior Instructors, Sr. Senior Instructors), rater groups (Group 1 in-semester, Group 2 post-semester), and versions (draft vs. final). T1: "The Holographic AI Assistant"; T2: "Interactive LLM-based teaching agents"; T3: "Generative AI in Healthcare".

reviews using a recorded calibration video prepared by a Group-1 senior instructor.

Rater groups and coverage Table 18 indicates that Group 1 grades are predominantly assigned by junior instructors, whereas Group 2 grades are predominantly assigned by senior instructors. This shift mainly reflects post-semester availability: once the course ended, much of the teaching staff was no longer available, so re-grading was carried out primarily by senior instructors; final submissions were graded only during the semester.

	Group 1			Group 2		
Reviews	Jr.	Sr.	Σ	Jr.	Sr.	Σ
Student reviews (N = 110)						
E (N = 34)	28	6	**34**	0	59	**59**
$\neg E$ (N = 76)	67	9	**76**	0	111	**111**
Σ	**95**	**15**	**110**	**0**	**170**	**170**

Table 19: Distribution of criterion-based student review assessments by instructor role (Jr. Junior Instructors, Sr. Senior Instructors) and rater group (Group 1 vs. Group 2), stratified by whether the reviewed author has an associated exposé (E) or not ($\neg E$).

Scope of IAA Our agreement analyses are necessarily restricted to items (exposés and reviews) that have assessments from two instructors. For exposés, we report agreement only for *draft* submissions, as final submissions were not scored post-semester due to the substantial time required for criterion-based assessment (Appendix G.3). For reviews, we focus on the subset that has criterion-based grades from Group 1 (student reviews). Finally, for a subset of reviews ($n = 60$), we collected duplicate retrospective ratings from Group 2, enabling a within-group agreement comparison in addition to cross-group agreement (Table 24). Beyond criterion-level results, we also report aggregate agreement (i) by rubric domain and (ii) by criterion expertise level (see Appendix E).

H.2 Human–human agreement

Agreement reporting We report raw quadratic weighted agreement (QWA), Krippendorff's α, Pearson correlation coefficient r, and simple mean

Reviews per exposé	Group 1					Group 2			
	N	Jr.	Sr.	St.	Σ	**N**	Jr.	Sr.	Σ
Exposés (total N = 55)									
0 review(s)	**0**	0	0	0	**0**	**30**	0	0	**0**
1 review(s)	**23**	17	6	0	**23**	**24**	4	20	**24**
2 review(s)	**23**	21	7	18	**46**	**1**	0	2	**2**
3 review(s)	**9**	8	3	16	**27**	**0**	0	0	**0**
Σ **totals**	**55**	**46**	**16**	**34**	**96**	**55**	**4**	**22**	**26**
Orphaned reviews (total N = 184)									
Σ totals	**184**	88	20	76	**184**	**0**	0	0	**0**

Table 20: Distribution of reviews per exposé by review group (Group 1 vs. Group 2). In the Exposés block, **N** is the number of exposés in the bucket; role columns (Jr./Sr./St.) are counts of reviews written by that role (columns that are all-zero for a group are omitted); and Σ is the total number of reviews. Orphaned reviews are reported separately.

Rater / Model	**Exposé** (#31)		**Review** (#22)	
	All	Ind.	All	Ind.
Human raters	0.76		0.88	
Claude Opus 4.6	**0.80**	**0.79**	**0.90**	**0.85**
Gemini 3.1 Pro	0.78	**0.79**	**0.90**	**0.89**
GPT-5.2	**0.80**	0.78	0.87	0.79
GPT OSS 120B	0.77	0.75	0.78	0.72
GPT OSS 20B	0.77	0.76	0.74	0.69
Llama 3.3 70B	0.76	0.75	**0.90**	0.82
Qwen3 80B	0.72	0.75	0.89	0.84

Table 21: Quadratic weighted agreement (QWA) for the two tasks, reported over all criteria only. Prompting variants: *all* (combined) and *ind.* (individual). Each LLM value shows human–LLM QWA averaged over both human raters and all criteria. The number of criteria are shown in the title (for exposés and reviews respectively). Bold indicates the best model performance per column.

difference ("mean bias") between groups. **QWA** is the observed agreement between two raters on an ordinal scale, where disagreements are weighted by squared distance so that larger score differences are penalized more strongly than adjacent-category mismatches. Let x_i, y_i be two raters' scores for item i on an ordinal scale with discrete domain $\mathcal{D} = \{d_{\min}, \ldots, d_{\max}\}$ and span $\Delta = d_{\max} - d_{\min} > 0$. We use the quadratic weight

$$w(x_i, y_i) = \max\left(0,\, 1 - \left(\frac{x_i - y_i}{\Delta}\right)^2\right),$$

and define raw quadratic weighted agreement as

$$\mathrm{QWA} = \frac{1}{N}\sum_{i=1}^{N} w(x_i, y_i),$$

where N is the number of co-rated items.

We use quadratic weighting because it preserves the magnitude of disagreement and remains interpretable under strong prevalence skew, a setting in which chance-corrected coefficients (e.g., κ or α) can be misleading. Accordingly, for highly skewed criteria (e.g., *Number of pages*, *Template used*, *Negative points*), we emphasize QWA.

α values range between -1 and 1, with 1 being perfect agreement, 0 being agreement by chance, and negative values indicating systematic disagreement; r measures the correlation between a system's predicted scores and the annotator-assigned scores; it ranges from -1 to 1. A positive (negative) value implies that the two sets of predictions are positively (negatively) correlated. Inter-annotator agreement results for exposé criteria are reported in Table 22, while agreement for review criteria is summarized in Tables 23 and 24.

H.3 Failure modes of chance-corrected agreement

This section interprets the agreement patterns observed in the exposé and review analyses (see Section 5.2 in the main paper). In particular, we focus on cases where chance-corrected reliability or correlation is negative, undefined, or unexpectedly low, despite high raw agreement, and relate these outcomes to prevalence effects, calibration differences between rater groups, and ambiguities in rubrics or training.

Prevalence effects under extreme class imbalance ("rare-event" criteria) For highly skewed binary (or near-binary) criteria, the expected agreement under a chance model is already very high, so a handful of rare, non-overlapping deviations can yield $\alpha < 0$ even when raw agreement is high. This explains the slightly negative values for *Template used* (ST1) and *Negative points* (AD2), where almost all submissions receive the default label, but the few exceptional ratings occur on different submissions. In such settings, Pearson's r is typically close to zero because the rating vectors are nearly constant, and mean bias $\Delta\mu$ helps diag-

nose whether the deviations are symmetric (ST1: $\Delta\mu \approx 0$) or slightly one-sided (AD2: small negative $\Delta\mu$ consistent with marginally more penalizing scores in one group).

Annotation guideline/interpretation issues (training or rubric ambiguity) Some negative/undefined reliability values are attributable to concrete, identifiable rubric failures. For *Metadata available* (ME2), metadata was present for all submissions; all disagreements originated from a single rater in Group 2, indicating a training/interpretation error (i.e., the rater did not apply the intended check). This also leads to an undefined Pearson correlation because one group's ratings are effectively constant (zero variance). For *Template used* (ST1), the few "0" assignments were triggered by superficial deviations (e.g., renamed chapter headings or bibliography compilation issues) despite the underlying template being used, revealing an unclear rubric description about what counts as a template violation. For *Methodology Availability* (AP8), we found edge cases where the methodology is present only as a very short list; these items were inconsistently judged as available vs. unavailable across groups, suggesting that the criterion definition is underspecified (i.e., availability vs. sufficiency/quality).

Systematic calibration shifts and low signal in subjective criteria For *Anker* (MO2), the negative (near-zero) α is consistent with a systematic calibration shift between groups: Group 2 assigns lower scores on average (positive mean bias $\Delta\mu = 0.309$), and disagreements are predominantly adjacent-category mismatches (e.g., 2 vs. 1), which remain relatively benign under quadratic weighting but still depress chance-corrected reliability. For *Problem relevance* (MO4), negative coefficients are driven by sparse and subjective "0" decisions: the negative label is rare and tends to be applied to different submissions across groups.

Prevalence effects under extreme class imbalance in review scoring A closely related prevalence-driven failure mode appears in review scoring for error and penalty criteria (Tables 23 and 24). These criteria are effectively *rare-event detection* tasks on a highly skewed label space: the non-penalizing default (typically 0) dominates, while penalizing assignments (e.g., -1 in *Errors (ER)*) occur only rarely. As a result, chance-corrected reliability and Pearson correlation become unstable or uninformative. A small number of non-overlapping penalizations can yield $\alpha \leq 0$ despite high raw agreement, and near-constant rating vectors drive r toward zero or render it undefined when one rater shows zero variance.

Overall, we emphasize quadratic-weighted raw agreement as our primary agreement signal and use α, r, $\Delta\mu$, and the pooled distributions to (i) flag prevalence-induced artifacts, (ii) diagnose calibration differences, and (iii) identify rubric/training issues that can be addressed in the future.

I Illustrative Examples

This chapter presents two representative examples from Exposía. The examples illustrate how exposé drafts, reviewer comments, and revised final versions are connected in the dataset. Each example is shown in three parts: an annotated draft excerpt, a short summary of the reviewer feedback, and the corresponding final excerpt. Passages highlighted in red indicate text spans that received inline comments from the reviewer, in this case, marking weaknesses that were addressed during revision.

Example 1: Anchor This example illustrates how reviewers ask students to strengthen the *anchor* of a proposal by making the motivation more concrete and evidence-based. In the draft, the problem context is understandable, but still too general: the text claims that museums have experienced a major decline in visitor numbers and need more engaging approaches, yet the argument is not substantiated with references or concrete examples. This is reflected not only in the human assessment, but also in the automatic scoring of the draft, where models differed in whether they interpreted the problem description as already sufficiently detailed or still somewhat too general. The automatic assessments of the draft show a clear split, which helps explain why the human evaluation was not entirely uniform either. First, most models already viewed the problem statement as sufficiently detailed. For example, *Claude Opus 4.6* assigned a score of 2 and argued that "the problem is described in detail," explicitly pointing to the post-COVID decline in visitor numbers, the reliance on traditional text-panel-based interaction, and the limited current use of holograms. Second, a smaller group of models judged the same passage more critically because the motivation remained somewhat general. For instance, *GPT OSS (20B)* assigned a score of 1 and summarized the draft as "Problem stated with some detail but missing deeper specifics."

The student review is especially informative

when viewed through *Feed Up*, *Feed Back*, and *Feed Forward*. It clearly identifies the goal of the passage, namely to motivate the topic convincingly, which is reflected in the score for *Learning Goals* (Score 2). It also explains what is missing in the current text by pointing out the unsupported claim and the need for a citation, corresponding to positive scores across the *Feed Back* criteria. Finally, it provides forward-looking guidance by recommending that important factual claims should be supported with evidence and statistics in future revisions, which is reflected in the scores for *Self-regulation* and *Learning Skills* (Score 1 each).

Example 2: Teaser (Research Question) This example shows how reviewers use the *teaser* criterion to assess whether a proposal states a clear research question. In the draft, the overall research direction is understandable, but the teaser remains too broad because it describes goals and anticipated contributions rather than a distinct research question. This is reflected especially clearly in the human reviews: one reviewer assigned a score of 1 and noted that a clearly formulated research question would give the exposé more structure, while another reviewer assigned a score of 0 and criticized the absence of a proper research question altogether. The automatic scoring of the draft points to the same central issue. First, several models took a strict view and penalized the text because no explicit research question was stated. For example, *GPT 5.2* assigned a score of 0 and wrote that "No explicit research question is stated." Second, some models interpreted the passage more leniently and treated it as an implicit research direction, while still criticizing the lack of precision. For instance, *Claude Opus 4.6* assigned a score of 1 and argued that the work was described in broad terms without a clearly articulated research question. Third, a smaller group of models inferred a sufficiently clear question from the surrounding context. For example, *Llama 3.3* assigned a score of 2 and described the research question as clearly formulated and aligned with the problem description.

J Experimental Details

Models We evaluate a diverse set of recent instruction-tuned LLMs spanning different architectures, scales, and deployment regimes. Our open-weight on-premise selection includes (i) a large dense decoder-only Transformer from the Llama 3.3 family (meta-llama/Llama-3.3-70B-Instruct; 70B parameters) (Team, 2024), (ii) a high-sparsity Mixture-of-Experts (MoE) model with long-context support (Qwen/Qwen3-Next-80B-A3B-Instruct; 80B total parameters with 3B active) (Team, 2025), as well as an independent MoE reasoning model in two variants, (iii) a high-capacity reasoning model (openai/gpt-oss-120b; 117B total parameters with 5.1B active) and (iv) a smaller low-latency variant (openai/gpt-oss-20b; 21B total parameters with 3.6B active) (OpenAI, 2025). These open-weight models can be executed locally on two NVIDIA A100 80GB GPUs, which is important for classroom deployment where data protection and limited consent often motivate on-premise inference rather than cloud-based APIs.

To contextualize results against strong proprietary baselines, we additionally include three closed-source, API-accessed models: OpenAI GPT 5.2 (`openai/gpt-5.2`), Anthropic Claude Opus 4.6 (`anthropic/claude-opus-4-6`), and Google Gemini 3.1 Pro Preview (`gemini/gemini-3.1-pro-preview`). Closed-source models are queried via the LiteLLM[17] interface, enabling a unified prompting and logging pipeline across providers.

Inference for open-weight models is performed using the Hugging Face TRANSFORMERS library[18] with deterministic greedy decoding (`do_sample=False`). For API-based models, we likewise use greedy decoding by default (temperature $= 0$; and where supported, `top_p` $= 1$).

Prompts For both tasks, we use a unified prompting strategy (Appendix K) with a system message that (i) assigns the model the role of an expert evaluator and (ii) enumerates the relevant criteria, including their names, short descriptions, allowed integer score ranges, and the verbal descriptions associated with each score level (Tables 8 and 9). The model is required to return a JSON list with one object per criterion, where each object contains the fields `criterion_name`, `assigned_score`, and `justification`. The additional justifications are requested to explain the decision-making process, as it is crucial in an environment where people are affected by the LLM outcome (Kolářová and Schmude, 2026).

Criteria excluded from the LLM experiments The exposé and review rubrics contain structural or format-related checks (e.g., LaTeX template usage, page limits) and content-related criteria. We

[17] https://www.litellm.ai
[18] https://huggingface.co/docs/transformers

Example 1: Excerpt (Criteria: Anker)

Draft Exposé

"The modern museum has multiple purposes - to curate and preserve, to research, and to reach out to the public. They challenge us and ask us to question our assumptions about the past or the world around us." [1] [...] But espacially the part about reaching out to the public doesn't hold truth nowadays. Since the COVID-19 pandemic, museums worldwide have faced a dramatic decline in visitor numbers, highlighting the need for more engaging and innovative approaches to attract audiences. [WEAKNESS: Where does this information come from? Cite.] Despite their [WEAKNESS: Unclear reference; use a precise noun phrase such as "the museums".] efforts to adapt many institutions still rely on traditional methods of interaction [WEAKNESS: Underline why improving human interaction is important.], by simply showcasing an object and additional information written on text panels.[...]

Review 1 (Junior Instructor)

[...] The anchor section provides a general context for the problem. However, the problem lacks detailed evidence or why it is important to research on it; for instance, while it mentions a "dramatic decline" since COVID-19, no quantitative data or case studies are included to support the claim. To strengthen the anchor, it is recommended to reference studies on the impact of the pandemic on cultural institutions and to underline its relevance/importance.[...]

Review 2 (Student)

[...] In your motivation, you wrote that since the COVID-19 pandemic, museums faced a dramatic decline in visitor numbers. You aim to motivate your topic by providing an up-to-date example and an alarming development. This is important to make the reader clear why you think that this topic is relevant.[FEED UP] To ensure you provide valid and important facts, you should always provide a citation or source of such claims. This ensures that the reader will trust your judgment and understand the importance of your work. In this case, you do not provide any source or citation for the claim. For a sophisticated paper, you should avoid having unproofed claims in your work.[FEED BACK] I would suggest citing some paper that proves your claim with statistics. This will lead to a better, more understandable paper and avoid having false claims in your work, which could be an undesired situation.[FEED FORWARD] [...]

Final Exposé

[...] But especially the part about reaching out to the public doesn't hold truth nowadays. Since the COVID-19 pandemic, museums worldwide have faced a dramatic decline in visitor numbers [2], highlighting the need for more engaging and innovative approaches to attract audiences. For example, the National Gallery in London still had a 55% drop in visitors two years after the pandemic [3]. To counter this decline, museums need to innovate and create more engaging ways to interact with their content, because meaningful interaction is the key driver for museum visits. Despite the efforts to adapt, many institutions still rely on traditional methods of interaction[...]

exclude the first one, as the LLM cannot successfully assess it because the generated PDF is needed for correct evaluation. We also exclude additional points that can be given by the human evaluator, as these occur only in isolated cases and are intended for exceptional circumstances. The following exposé criteria are excluded: Metadata available; Template used; Number of pages; Additional points; Negative points. The following review criteria are excluded: Additional points; Negative points.

Example 2: Excerpt (Criteria: Teaser (RQ))

Draft Exposé

> [...] This research addresses the critical gap in current AI-driven cancer research by integrating the impact of the tumor microenvironment into data analysis workflows for personalized treatment planning. [WEAKNESS: This states a research aim rather than an explicit research question.] We will discuss the landscape of adapting current AI-enhanced data analysis to the importance of TME, address the limitations of current AI models in incorporating TME data and explore the anticipated improvement of personalized treatment prediction. [WEAKNESS: The goals remain broad; the main research question is not clearly stated.][...]

Review 1 (Junior Instructor) *Score: 1*

> [...] Having an explicit, clearly defined research question gives a lot of structure to the overall expose, especially to the methodology section as your methods should tackle all details of all RQs. As you describe your work instead, the exact goals and your main focus are not entirely clear to me. When making your RQ more explicit, also keep in mind to check whether your methodology addresses all relevant sub-parts of your RQ(s). [...]

Review 2 (Senior Instructor) *Score: 0*

> [...] Also, a major issue is the lack of a distinct research question. The Motivation raises a question right at the end of the first paragraph. However, this question is rather a rhetorical question than a research question. Authors should include at least one sound research question that is answered in the work. [...]

Final Exposé *Score: 2*

> [...] The research will explore the following key questions:
>
> - How does the heterogeneity of the tumor microenvironment impact treatment response predictions?
> - What are the limitations of current AI-driven models in incorporating TME data, and how can these be addressed?
> - How can integrating TME characteristics improve the predictive accuracy of personalized cancer treatments? [...]

K Prompt Templates

This appendix provides the exact prompt formats used in all baseline experiments. The rubric text included in the templates is automatically generated from the released JSON rubric files, ensuring that the prompts align with the annotation scheme used in Exposía.

For both tasks, we use a unified prompt structure. The system message (i) assigns the model the role of an expert evaluator and (ii) enumerates the relevant criteria, including their names, short descriptions, allowed integer score ranges, and the verbal descriptions associated with each score level (Tables 8 and 9). The model is required to return a JSON *list* with one object per criterion, containing the fields `criterion_name`, `assigned_score`, and a short `justification`. The additional justifications are requested to explain the decision-making process, as it is crucial in an environment where people are affected by the LLM outcome (Kolářová and Schmude, 2026).

K.1 Expose Scoring Prompt

The user message provides the topic, the full exposé text, and the bibliography.

System message

```
You are an expert evaluator specializing
    in assessing student exposes.
Your task is to carefully evaluate the
    given expose based on the specified
grading criteria.

The grading criteria are as follows:

## <Rubric Name>
<Rubric description>

### Criterion: <Criterion Name>
Description: <Criterion description>
Allowed scores:
- <Points> points: <Description>
...

## <Next Rubric>
...

You must:
  1. Evaluate the expose **criterion by
     criterion**.
  2. For **every** listed criterion,
     assign exactly **one** score using
     **only** the scoring options and
     point values provided.
  3. Justify each score with a clear,
     concise and **short** explanation.

Your response MUST be a list of
    dictionaries in **valid JSON** schema:

[
  {
    "criterion_name": "<the EXACT
    criterion name from the criteria
    list, without any other words>",
    "assigned_score": "<an INTEGER, one of
    the allowed point values for this
    criterion>",
    "justification": "<SHORT justification
    of why you decided to give this
    score>"
  },
  ...
]

Constraints:
- Return **one object per criterion**; do
    not skip any criteria.
- Use only the criteria and scoring
    options that are explicitly provided.
- Return a response that consists solely
    of valid JSON, with no additional
    explanatory text before or after it.
- Make sure the JSON is syntactically
    valid (no trailing commas, correct
    quotation marks, etc.).
```

User message

```
The topic is: "<TOPIC>".

Here is the expose you have to grade:

<EXPOSE TEXT>

Bibliography:
<BIBLIOGRAPHY>
```

K.2 Review Scoring Prompt

The user message provides the full written review text.

System message

```
You are an expert evaluator specializing
    in assessing the quality of student
    peer reviews.
Your task is to carefully evaluate the
    reviewer's performance based on the
    specified grading criteria.

Grading criteria:

## <Rubric Name>
<Rubric description>

### Criterion: <Criterion Name>
Description: <Criterion description>
Allowed scores:
- <Points> points: <Description>
...

## <Next Rubric>
...

You must:
1. Evaluate the peer review **criterion by
    criterion**.
2. For **every** listed criterion, assign
    exactly **one** score using **only**
    the scoring options and point values
    provided for that criterion.
3. Base your evaluation strictly on the
    quality of the full review text.
4. Justify each score with a clear,
    concise and **short** explanation.

Output format:
Return a list of objects following this
    **valid JSON** schema:

[
  {
    "criterion_name": "<EXACT criterion
    name from the criteria list>",
    "assigned_score": <INTEGER, one of the
    allowed point values for this
    criterion>,
    "justification": "<SHORT justification
    of why you decided to give this
    score>"
  },
```

```
  ...
]

Constraints:
- Return **one object per criterion**; do
    not skip any criteria.
- Use only the criteria and scoring
    options that are explicitly provided.
- The response must consist solely of
    valid JSON, with no additional
    explanatory text.
- Make sure the JSON is syntactically
    valid (no trailing commas, correct
    quotation marks, etc.).
```

User message

```
Here are the materials of the peer review
    you have to grade.

=== Full written text of the review ===
<FULL_REVIEW_TEXT>
```

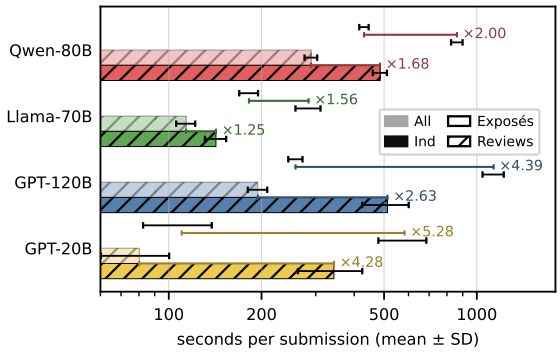


Figure 6: Runtime per submission for criterion-based assessments across open-weight models. Bars show mean seconds per submission with standard deviation bars. Within each task, **All** denotes combined prompting (light fill) and **Ind** denotes single-criterion prompting (dark fill). Review runs are hatched; Exposé runs are unhatched. Ratios show $\times(\mathrm{Ind}/\mathrm{All})$, highlighting the additional latency introduced by single-criterion prompting.

L Additional Experimental Results

Criterion-wise alignment Figures 7 and 8 report weighted mean binned JSD at the criterion level (darker = lower/better), complementing the domain-level averages in Table 29. Across both tasks, criteria on the 0–1–2 scale show a consistently higher JSD baseline than 0–1 criteria (Table 29), reflecting the increased degrees of freedom and typically higher annotator dispersion in three-way ratings, which makes distributional matching more challenging. For exposé, most criteria exhibit low divergence, but mismatches concentrate in the more interpretive rubric components—especially SOTA- and methodology-related criteria—with *Anker* emerging as uniformly difficult and *Schedule Realistic Relevance* showing a clear spike for some model/prompt settings. Reviews are more homogeneous overall: stylistic dimensions (language, tone) remain low across models, whereas divergence is driven primarily by epistemic "knowledge" judgments and reference-related criteria (notably *Missing Reference*).

Scalability and classroom deployment Across all models and both tasks, combined prompting is more efficient in most cases than single-criterion prompting: it reduces mean runtime per submission by about $1.3\times$–$5.3\times$ and avoids the linear cost growth from issuing one model call per criterion (Figure 6). Single-criterion prompting also substantially increases total token processing because the full input is repeated across prompts, with particularly large overhead for long exposés and the review setting. These efficiency differences become especially consequential under the *observed course usage dynamics*. As shown by the behavioural interaction logs (Appendix Figure 5), student activity is strongly deadline-driven, with a pronounced spike in actions immediately before the review deadline, whereas instructor activity is relatively steady. This implies that in practice, LLM-based scoring must be provisioned for short periods of peak demand rather than average-day usage. As such, the overhead of single-criterion prompting is particularly problematic, while combined prompting better supports batching and throughput during deadline-induced bursts. Consequently, *combined prompting is the recommended operating point for classroom deployment under constrained on-premise resources, with model choice primarily determining the remaining latency–quality trade-off*. Runtime and token usage are reported in Appendix M (Tables 30–31).

Operational latency under peak-demand grading windows In classroom deployment under constrained on-premise resources, using LLM scores as instructor support (e.g., batch pre-scoring or rubric-aligned decision support) introduces a practical scheduling choice: scores can either be *pre-computed* ahead of grading, or compute capacity must be provisioned to serve requests *on demand* throughout the grading period. Under

deadline-driven bursts (Appendix Figure 5), it can translate directly into waiting times, because even combined prompting incurs minute-level latency per submission and shows substantial min–max spread (Tables 30–31). To illustrate, consider a cohort of N=600 exposés scored on the same two-GPU setup used in our experiments. For exposés, the fastest combined configuration in Table 30 (GPT OSS 20B, All) yields $600 \times 110.2\text{s} \approx 18.4$ hours on average; larger models further increase this budget (e.g., Qwen3 80B, All: ≈ 71.7 hours). For student reviews, the volume is larger ($2N$ submissions); in the worst-case combined configuration, runtimes of ≈ 289.6s per review (Table 31) translate into ≈ 96.5 hours of wall-clock time, making peak-time live scoring particularly susceptible to queueing delays.

Finally, beyond latency, we observed occasional out-of-memory failures for unusually long inputs (e.g., exposés with many and very long comments), reinforcing the need for robust deployment to plan around worst-case input lengths (e.g., advance computation, batching, and input-length safeguards) rather than relying on unconstrained live scoring near deadlines.

Is review text enough for review assessment? Overall, we did not find an indication that review scoring requires context beyond the review text. In supplementary experiments, we augmented the review input jointly with additional context from the exposé and inline comments, but did not observe consistent improvements over the review-only setting.

M Runtime and Token Consumption

In this appendix, we focus exclusively on inference-time cost (runtime and token usage) under these fixed conditions. Prompting variants (combined vs. single-criterion) are defined in Section 6.1. For each model/condition, we measure wall-clock runtime both per full submission (`s/assessment`) and per model call (`s/prompt`). Token counts are computed with the respective model tokenizer on the fully formatted prompt (system + user messages) and the generated JSON response, and reported as `tokens/prompt` split into `tokens/input` and `tokens/output`. All values are aggregated across evaluation instances and reported as mean (SD) and [min–max], capturing variation in input length.

LLM scoring throughput: combined prompting is essential Tables 30 and 31 report runtime and token usage for exposé and review scoring, respectively. Across both tasks, combined prompting is essential for throughput: it requires a single forward pass per submission and therefore yields the lowest latency (110–430s on average across models for exposé scoring). In contrast, single-criterion prompting multiplies calls per submission, increasing total runtime to 284–1133s on average for exposés, with worst cases up to 23 minutes. The same multiplicative effect appears in token processing: although single-criterion outputs are shorter per call, repeating the full input across criteria yields an order-of-magnitude increase in total processed tokens per submission (approximately $13\times$ in our setup). Review scoring exhibits the same trend (Table 31), with combined prompting consistently yielding lower latency and substantially fewer processed tokens per submission than single-criterion prompting.

N Behaviour Data: Interaction Logs and Semester Dynamics

To complement the textual artefacts in Exposía, we also record fine-grained *behavioural interaction logs* from the review platform used in the course.

Temporal dynamics across the semester Figure 5 shows daily action counts for students and instructors. Two qualitative patterns stand out.

First, student activity is strongly deadline-driven: actions rise sharply in the days immediately preceding the review deadline (red dashed line). In contrast, instructor activity is comparatively smoother and distributed across the semester, consistent with instructional work that is scheduled and paced over time rather than concentrated in the final days.

Second, activity drops to near-zero during the winter break (hatched region), including the Christmas/New Year period, and then resumes afterward.

Implications for LLM-assisted assessment workflows The pronounced pre-deadline student spike (Figure 5) suggests that any LLM-based support integrated into the reviewing/assessment pipeline (e.g., rubric guidance, feedback scaffolding, or automated checks) will face *highly non-uniform demand*. Compute budgets and latency constraints should be evaluated under peak load conditions rather than on average-day usage.

Limitations These behavioural patterns require *particular* caution because they reflect only *partial observation*. Not all students consented to

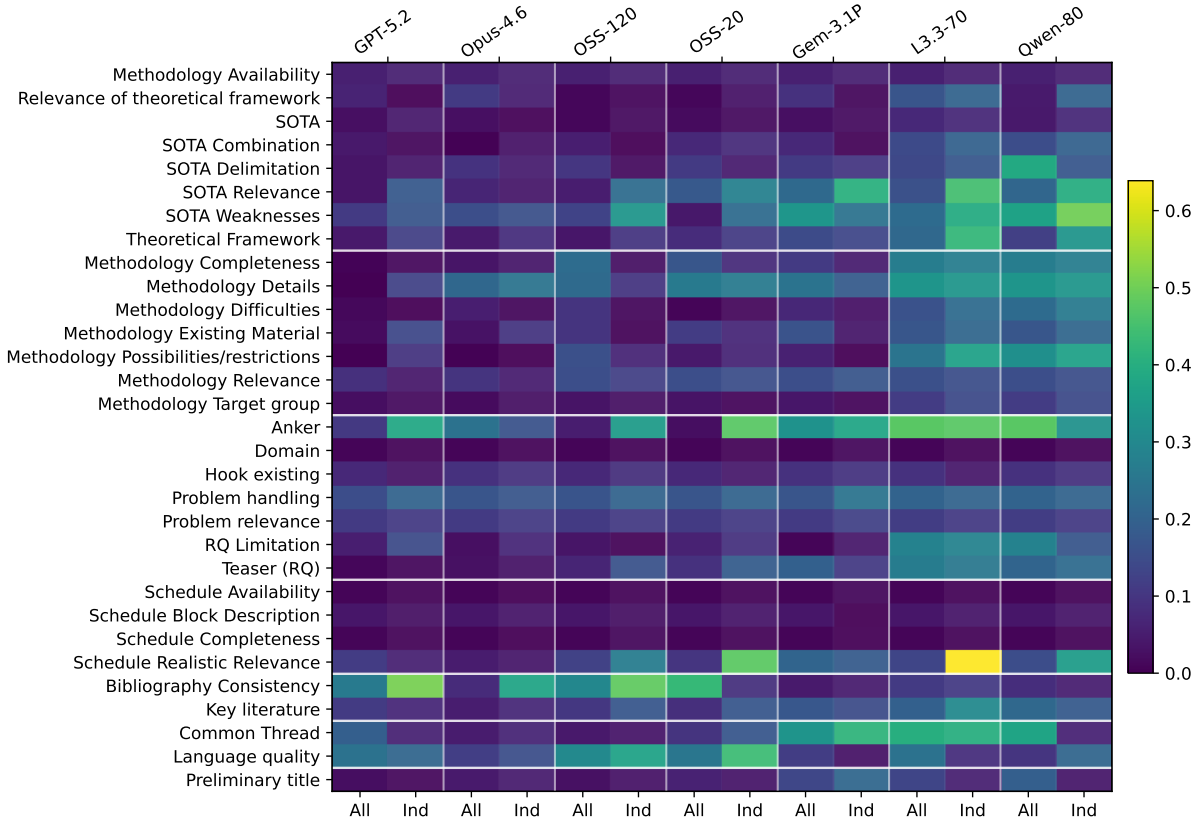


Figure 7: **Exposé: Criterion × model heatmap (JSD).** Cell values show weighted mean JSD per criterion and model (weights: items_in_bin). Lower is better. Model abbreviations: *Opus-4.6*=Claude Opus 4.6; *Gem-3.1P*=Gemini 3.1 Pro; *GPT-5.2*=GPT 5.2; *OSS-120*=GPT OSS (120B); *OSS-20*=GPT OSS (20B); *L3.3-70*=Llama 3.3 (70B); *Qwen-80*=Qwen 3 (80B).

behavioural logging, so the curves describe only the consenting subset and may not generalize to the full cohort. The instructor group is also small (n=9), limiting statistical power and making aggregates sensitive to individual work styles. Moreover, instructor activity is not directly comparable to student activity because instructors carried a substantially higher workload: while students completed only two reviews without grading, instructors collectively reviewed across the cohort and additionally assigned criterion-based grades. Finally, logs capture only on-platform actions; substantial reviewing work may occur offline (e.g., reading drafts externally or drafting feedback elsewhere). Thus, peaks and troughs should be interpreted as *platform activity* rather than complete measures of effort or time-on-task.

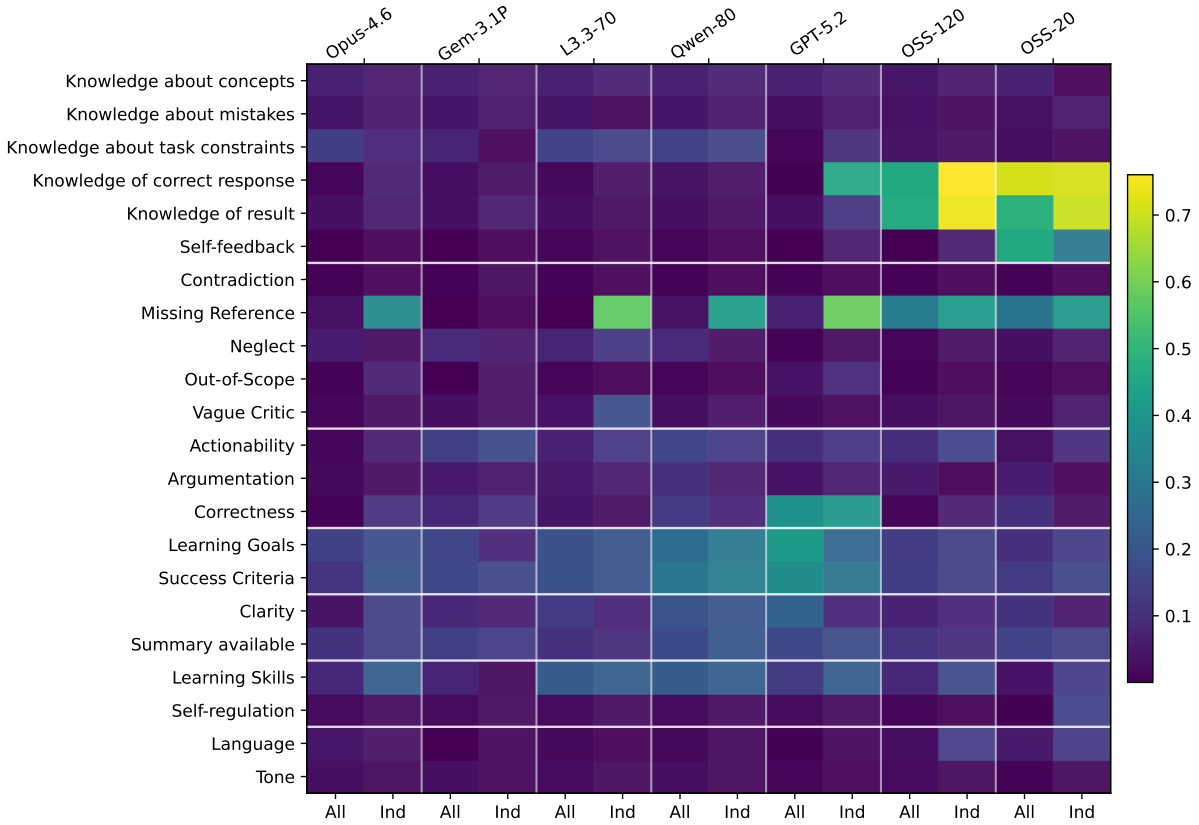


Figure 8: **Reviews: Criterion × model heatmap (JSD).** Cell values show weighted mean JSD per criterion and model (weights: items_in_bin). Lower is better. Model abbreviations: *Opus-4.6*=Claude Opus 4.6; *Gem-3.1P*=Gemini 3.1 Pro; *GPT-5.2*=GPT 5.2; *OSS-120*=GPT OSS (120B); *OSS-20*=GPT OSS (20B); *L3.3-70*=Llama 3.3 (70B); *Qwen-80*=Qwen 3 (80B).

Code	Criterion	Distribution (%)	QWA	α	r	$\Delta\mu$
LA1	Language quality	0 (15), 1 (55), 2 (30)	0.82	0.15	0.17	0.13
LA2	Common Thread	0 (4), 1 (65), 2 (32)	0.87	0.05	0.20	0.38
Average Language			0.85	0.10	0.19	0.25
ME1	Preliminary title	0 (3), 1 (42), 2 (55)	0.86	0.10	0.31	0.44
ME2*	Metadata available	0 (10), 1 (90)	0.80	-0.10	—	0.20
Average Metadata			0.83	0.00	0.31	0.32
ST1*	Template used	0 (2), 1 (98)	**0.96**	-0.01	-0.02	0.00
ST2*	Number of pages	0 (3), 1 (97)	**0.98**	0.66	0.70	0.02
Average Form/Structure			0.97	0.33	0.34	0.01
MO1	Hook existing	0 (15), 1 (85)	0.76	0.10	0.10	0.05
MO2	Anker	0 (5), 1 (63), 2 (33)	0.85	-0.00	0.06	0.31
MO3*	Domain	0 (5), 1 (95)	0.95	0.38	0.49	0.05
MO4	Problem relevance	0 (17), 1 (83)	0.65	-0.20	-0.21	-0.02
MO5	Problem handling	0 (35), 1 (65)	0.64	0.20	0.24	0.18
MO6	Teaser (RQ)	0 (7), 1 (56), 2 (36)	0.85	0.15	0.21	0.22
MO7	RQ Limitation	0 (46), 1 (54)	0.58	0.17	0.25	0.27
Average Motivation			0.76	0.11	0.16	0.15
AP1	SOTA	0 (13), 1 (87)	0.85	0.35	0.37	0.07
AP2	SOTA Relevance	0 (9), 1 (56), 2 (35)	0.86	0.28	0.39	0.36
AP3	SOTA Weaknesses	0 (23), 1 (55), 2 (22)	0.83	0.25	0.30	0.27
AP4	SOTA Delimitation	0 (30), 1 (70)	0.73	0.36	0.40	0.16
AP5	SOTA Combination	0 (34), 1 (66)	0.65	0.23	0.29	0.20
AP6	Theoretical Framework	0 (20), 1 (47), 2 (33)	0.78	0.15	0.20	0.29
AP7	Relevance of theoretical framework	0 (35), 1 (65)	0.64	0.20	0.20	0.00
AP8*	Methodology Availability	0 (9), 1 (91)	0.82	-0.09	-0.08	0.11
Average Approach			0.77	0.22	0.26	0.18
MD1	Methodology Completeness	0 (45), 1 (55)	0.69	0.38	0.46	0.24
MD2	Methodology Relevance	0 (26), 1 (74)	0.73	0.30	0.30	0.05
MD3	Methodology Target group	0 (25), 1 (75)	0.73	0.27	0.30	0.13
MD4	Methodology Existing Material	0 (35), 1 (65)	0.55	0.02	0.03	-0.13
MD5	Methodology Difficulties	0 (44), 1 (56)	0.56	0.12	0.19	0.25
MD6	Methodology Possibilities/restrictions	0 (56), 1 (44)	0.60	0.19	0.27	0.25
MD7	Methodology Details	0 (53), 1 (47)	0.49	-0.01	0.01	0.18
Average Methodology			0.62	0.18	0.22	0.14
SC1*	Schedule Availability	0 (4), 1 (96)	0.96	0.49	0.57	0.04
SC2*	Schedule Completeness	0 (4), 1 (96)	0.96	0.49	0.57	0.04
SC3*	Schedule Block Description	0 (6), 1 (94)	0.91	0.24	0.28	-0.05
SC4	Schedule Realistic Relevance	0 (7), 1 (71), 2 (22)	0.90	0.21	0.36	0.25
Average Schedule			0.93	0.36	0.44	0.07
BI1	Bibliography Consistency	0 (19), 1 (81)	0.80	0.36	0.49	0.20
BI2	Key literature	0 (15), 1 (35), 2 (49)	0.76	0.03	0.35	0.64
Average Bibliography			0.78	0.19	0.42	0.42
AD1*	Additional points	0 (97), 1 (2), 2 (1)	**0.99**	0.67	0.70	0.00
AD2*	Negative points	-2 (2), -1 (1), 0 (97)	0.96	-0.02	-0.03	-0.02
Average Additional points			0.98	0.32	0.34	-0.01
Average all criteria			0.79	0.20	0.27	0.16

Table 22: Quadratic weighted agreement (QWA) between Group 1 and Group 2 raters for each exposé criterion, with rubric-wise and overall averages. Columns show agreement (Agr), reliability (α), correlation (r), and mean bias ($\Delta\mu$). Rows marked with * have a highly skewed class distribution ($\max_cp(c) \geq 90\%$).

Code	Criterion	Distribution (%)	QWA	α	r	$\Delta\mu$
SC1	Summary available	0 (14), 1 (30), 2 (56)	0.90	0.49	0.61	0.11
SC2	Clarity	0 (10), 1 (38), 2 (53)	0.82	0.04	0.15	0.38
Average Structure and Clarity			0.86	0.27	0.38	0.25
LG1*	Language	0 (6), 1 (94)	0.88	-0.06	-0.06	-0.07
LG2*	Tone	0 (1), 1 (7), 2 (92)	0.98	0.23	0.39	-0.06
Average Language			0.93	0.08	0.17	-0.06
FU1	Learning Goals	0 (29), 1 (28), 2 (42)	0.83	0.51	0.50	-0.08
FU2	Success Criteria	0 (32), 1 (31), 2 (37)	0.79	0.35	0.36	0.05
Average Feed Up			0.81	0.43	0.43	-0.01
FB1*	Knowledge of result	0 (5), 1 (95)	0.95	0.26	0.40	-0.05
FB2*	Knowledge of correct response	0 (8), 1 (92)	0.90	-0.05	-0.05	-0.05
FB3	Knowledge about task constraints	0 (33), 1 (67)	0.65	0.07	0.08	0.07
FB4	Knowledge about concepts	0 (13), 1 (87)	0.82	-0.00	0.05	-0.12
FB5*	Knowledge about mistakes	0 (8), 1 (92)	0.90	-0.05	-0.05	0.01
FB6*	Self-feedback	0 (3), 1 (97)	0.97	-0.01	—	0.03
Average Feed Back			0.86	0.04	0.08	-0.02
FF1*	Self-regulation	0 (6), 1 (94)	0.95	-0.02	-0.02	-0.03
FF2	Learning Skills	0 (39), 1 (61)	0.60	0.06	0.10	0.17
Average Feed Forward			0.78	0.02	0.04	0.07
CQ1	Correctness	0 (1), 1 (28), 2 (71)	0.91	0.14	0.16	0.12
CQ2	Actionability	0 (1), 1 (32), 2 (66)	0.92	0.22	0.21	0.08
CQ3*	Argumentation	0 (10), 1 (90), 2 (0)	**0.98**	0.30	0.36	-0.06
Average Content Quality			0.94	0.22	0.24	0.05
ER1	Neglect	-1 (16), 0 (84)	0.86	0.40	0.41	-0.05
ER2*	Vague Critic	-1 (7), 0 (93)	0.91	0.12	0.12	0.00
ER3*	Out-of-Scope	-1 (4), 0 (96)	0.93	-0.03	-0.03	-0.01
ER4*	Missing Reference	-1 (2), 0 (98)	0.96	-0.01	—	-0.04
ER5*	Contradiction	-1 (2), 0 (98)	**0.98**	-0.00	-0.01	0.00
Average Errors			0.93	0.10	0.12	-0.02
AD1*	Additional points	0 (100), 1 (0)	**1.00**	—	—	0.00
AD2*	Negative points	-1 (2), 0 (98)	0.96	-0.01	—	-0.04
Average Additional points			0.98	-0.01	—	-0.02
Average all criteria			0.89	0.13	0.18	0.02

Table 23: Quadratic weighted agreement (QWA) between review Group 1 and Group 2 raters for each review criterion, with rubric-wise and overall averages. Columns show agreement (Agr), reliability (α), correlation (r), and mean bias ($\Delta\mu$). Rows marked with * have a highly skewed class distribution ($\max_cp(c) \geq 90\%$).

Code	Criterion	Distribution (%)	QWA	α	r	$\Delta\mu$
SC1	Summary available	0 (13), 1 (43), 2 (43)	0.91	0.60	0.62	0.00
SC2	Clarity	0 (12), 1 (48), 2 (40)	0.88	0.47	0.45	0.17
Average Structure and Clarity			0.89	0.53	0.54	0.08
LG1*	Language	0 (4), 1 (96)	0.95	0.38	0.38	0.02
LG2	Tone	0 (1), 1 (12), 2 (88)	0.93	0.02	0.00	0.07
Average Language			0.94	0.20	0.19	0.04
FU1	Learning Goals	0 (28), 1 (28), 2 (44)	0.75	0.31	0.33	0.25
FU2	Success Criteria	0 (34), 1 (35), 2 (31)	0.72	0.13	0.15	0.23
Average Feed Up			0.74	0.22	0.24	0.24
FB1*	Knowledge of result	0 (6), 1 (94)	0.88	-0.05	-0.06	-0.02
FB2	Knowledge of correct response	0 (11), 1 (89)	0.85	0.23	0.22	-0.02
FB3	Knowledge about task constraints	0 (37), 1 (63)	0.53	0.00	0.03	0.17
FB4	Knowledge about concepts	0 (13), 1 (87)	0.77	-0.00	-0.01	0.03
FB5*	Knowledge about mistakes	0 (10), 1 (90)	0.87	0.27	0.27	0.03
FB6*	Self-feedback	0 (4), 1 (96)	0.92	-0.03	-0.04	0.02
Average Feed Back			0.80	0.07	0.07	0.04
FF1*	Self-regulation	0 (9), 1 (91)	0.82	-0.09	-0.10	-0.02
FF2	Learning Skills	0 (48), 1 (52)	0.57	0.14	0.14	0.10
Average Feed Forward			0.69	0.02	0.02	0.04
CQ1	Correctness	0 (0), 1 (39), 2 (61)	0.91	0.27	0.27	0.02
CQ2	Actionability	0 (2), 1 (34), 2 (63)	0.89	0.23	0.22	0.05
CQ3	Argumentation	0 (12), 1 (88), 2 (0)	0.94	-0.12	-0.13	0.00
Average Content Quality			0.91	0.13	0.12	0.02
ER1	Neglect	-1 (15), 0 (85)	0.80	0.22	0.22	-0.03
ER2*	Vague Critic	-1 (5), 0 (95)	0.90	-0.04	-0.05	0.00
ER3*	Out-of-Scope	-1 (2), 0 (98)	0.95	-0.02	-0.02	-0.02
ER4*	Missing Reference	-1 (1), 0 (99)	**0.98**	0.00	—	0.02
ER5*	Contradiction	-1 (2), 0 (98)	0.95	-0.02	-0.02	-0.02
Average Errors			0.92	0.03	0.03	-0.01
AD1*	Additional points	0 (100), 1 (0)	**1.00**	—	—	0.00
AD2*	Negative points	-1 (1), 0 (99)	**0.98**	0.00	—	0.02
Average Additional points			0.99	0.00	—	0.01
Average all criteria			0.86	0.13	0.14	0.04

Table 24: Quadratic weighted agreement (QWA) within Group 2 raters for each review criterion, with rubric-wise and overall averages. Columns show agreement (Agr), reliability (α), correlation (r), and mean bias ($\Delta\mu$). Rows marked with * have a highly skewed class distribution ($\max_cp(c) \geq 90\%$).

Code	Qwen3 (80B)				Llama 3.3 (70B)				GPT OSS (120B)				GPT OSS (20B)			
	$A^{\text{all}}_{\text{G1}}$	$A^{\text{ind}}_{\text{G1}}$	$A^{\text{all}}_{\text{G2}}$	$A^{\text{ind}}_{\text{G2}}$	$A^{\text{all}}_{\text{G1}}$	$A^{\text{ind}}_{\text{G1}}$	$A^{\text{all}}_{\text{G2}}$	$A^{\text{ind}}_{\text{G2}}$	$A^{\text{all}}_{\text{G1}}$	$A^{\text{ind}}_{\text{G1}}$	$A^{\text{all}}_{\text{G2}}$	$A^{\text{ind}}_{\text{G2}}$	$A^{\text{all}}_{\text{G1}}$	$A^{\text{ind}}_{\text{G1}}$	$A^{\text{all}}_{\text{G2}}$	$A^{\text{ind}}_{\text{G2}}$
LA1	0.83	0.85	0.85	0.91	0.78	0.84	0.77	0.83	0.68	0.59	0.73	0.67	0.72	0.57	0.77	0.64
LA2	0.87	0.88	0.76	0.85	0.87	0.87	0.75	0.75	0.88	0.90	0.89	0.86	0.82	0.87	0.91	0.81
Language	0.85	0.87	0.80	0.88	0.82	0.85	0.76	0.79	0.78	0.75	0.81	0.76	0.77	0.72	0.84	0.73
ME1	0.91	0.92	0.82	0.86	0.91	0.90	0.84	0.87	0.91	0.84	0.87	0.84	0.82	0.81	0.90	0.86
Metadata	0.91	0.92	0.82	0.86	0.91	0.90	0.84	0.87	0.91	0.84	0.87	0.84	0.82	0.81	0.90	0.86
MO1	0.87	0.87	0.82	0.82	0.87	0.84	0.82	0.82	0.84	0.73	0.78	0.60	0.84	0.71	0.78	0.69
MO2	0.86	0.88	0.74	0.76	0.86	0.86	0.74	0.74	0.92	0.87	0.82	0.76	0.90	0.86	0.85	0.74
MO3	**0.96**	**0.98**	**0.91**	**0.93**	**0.98**	**0.98**	0.93	**0.93**	**0.98**	**0.98**	**0.93**	**0.93**	**0.98**	**0.98**	**0.93**	**0.93**
MO4	0.38	0.82	0.42	0.84	0.75	0.82	0.76	0.84	0.82	0.82	0.84	0.84	0.82	0.82	0.84	0.84
MO5	0.75	0.75	0.56	0.56	0.75	0.75	0.56	0.56	0.76	0.75	0.55	0.56	0.73	0.75	0.55	0.56
MO6	0.87	0.85	0.79	0.78	0.86	0.87	0.78	0.79	0.92	0.89	0.85	0.85	0.90	0.87	0.81	0.81
MO7	0.67	0.69	0.40	0.42	0.67	0.67	0.40	0.40	0.75	0.65	0.47	0.71	0.67	0.76	0.40	0.49
Motivation	0.77	0.83	0.66	0.73	0.82	0.83	0.71	0.72	0.85	0.81	0.75	0.75	0.83	0.82	0.74	0.72
AP1	0.89	0.91	0.87	0.84	0.91	0.91	0.84	0.84	0.93	0.85	0.85	0.82	0.93	0.89	0.89	0.85
AP2	0.86	0.85	0.76	0.71	0.87	0.84	0.79	0.70	0.89	0.88	0.81	0.77	0.86	0.86	0.91	0.74
AP3	0.75	0.70	0.64	0.60	0.75	0.72	0.72	0.65	0.78	0.77	0.73	0.66	0.85	0.75	0.87	0.67
AP4	0.15	0.78	0.11	0.62	0.80	0.78	0.64	0.62	0.82	0.76	0.65	0.67	0.84	0.82	0.67	0.65
AP5	0.62	0.76	0.49	0.56	0.80	0.76	0.60	0.56	0.84	0.75	0.56	0.62	0.82	0.84	0.62	0.67
AP6	0.83	0.77	0.76	0.66	0.80	0.76	0.74	0.63	0.81	0.84	0.88	0.80	0.78	0.83	0.88	0.82
AP7	0.16	0.65	0.15	0.65	0.64	0.65	0.64	0.65	0.67	0.71	0.71	0.60	0.56	0.75	0.65	0.71
AP8	**0.96**	**0.96**	0.85	0.85	**0.96**	**0.96**	0.85	0.85	**0.96**	**0.96**	0.85	0.85	**0.96**	**0.96**	0.85	0.85
Approach	0.65	0.80	0.58	0.69	0.82	0.80	0.73	0.69	0.84	0.82	0.76	0.72	0.82	0.84	0.79	0.75
MD1	0.67	0.67	0.44	0.44	0.67	0.67	0.44	0.44	0.69	0.64	0.45	0.58	0.65	0.78	0.49	0.58
MD2	0.73	0.76	0.69	0.71	0.75	0.76	0.69	0.71	0.76	0.75	0.71	0.69	0.76	0.76	0.71	0.71
MD3	0.80	0.82	0.67	0.69	0.80	0.82	0.67	0.69	0.85	0.78	0.73	0.69	0.87	0.78	0.75	0.73
MD4	0.56	0.58	0.73	0.71	0.56	0.58	0.73	0.71	0.62	0.64	0.71	0.73	0.60	0.58	0.73	0.67
MD5	0.71	0.69	0.45	0.44	0.71	0.71	0.49	0.45	0.78	0.64	0.56	0.64	0.82	0.70	0.71	0.70
MD6	0.58	0.56	0.33	0.31	0.58	0.56	0.36	0.31	0.67	0.73	0.45	0.55	0.64	0.65	0.56	0.55
MD7	0.56	0.56	0.38	0.38	0.56	0.56	0.38	0.38	0.58	0.60	0.44	0.45	0.60	0.60	0.42	0.42
Methodology	0.66	0.66	0.53	0.52	0.66	0.67	0.54	0.53	0.71	0.68	0.58	0.62	0.71	0.70	0.62	0.62
SC1	**0.98**	**0.98**	**0.95**	**0.95**	**0.98**	**0.98**	**0.95**	**0.95**	**0.96**	**0.96**	**0.96**	**0.96**	**0.98**	**0.96**	**0.95**	**0.96**
SC2	**0.98**	**0.98**	**0.95**	**0.95**	**0.98**	**0.98**	**0.95**	**0.95**	**0.98**	**0.95**	**0.95**	**0.91**	**0.98**	**0.98**	**0.95**	**0.95**
SC3	0.91	0.91	**0.96**	**0.96**	0.91	0.91	**0.96**	**0.96**	0.91	0.93	**0.96**	**0.98**	0.91	0.91	**0.96**	**0.96**
SC4	0.85	0.82	0.85	0.78	0.90	0.78	**0.95**	0.72	0.85	0.83	0.88	0.81	0.85	0.77	0.88	0.74
Schedule	0.93	0.92	0.93	0.91	0.94	0.91	0.95	0.89	0.93	0.92	0.94	0.92	0.93	0.91	0.93	0.90
BI1	0.93	0.72	0.72	0.59	0.91	0.91	0.71	0.71	0.27	0.13	0.40	0.33	0.20	0.49	0.40	0.55
BI2	0.90	0.92	0.68	0.73	0.90	0.88	0.71	0.66	0.86	0.91	0.80	0.72	0.91	0.88	0.75	0.70
Bibliography	0.91	0.82	0.70	0.66	0.91	0.90	0.71	0.68	0.57	0.52	0.60	0.52	0.56	0.69	0.57	0.62
All criteria	0.75	0.80	0.66	0.70	0.81	0.80	0.71	0.70	0.81	0.77	0.74	0.72	0.79	0.79	0.75	0.73

Table 25: Quadratic-weighted agreement between open-weight LLM scores and human ratings for exposé criteria. Columns report A_{G1} (Group 1 vs. LLM) and A_{G2} (Group 2 vs. LLM) for the prompting variants shown (all, ind). Bold/underlined cells mark column-wise maxima/minima. Average rows summarize agreement within rubrics and across all criteria.

Code	Claude Opus 4-6				Gemini 3.1 Pro				GPT-5.2			
	$A^{\text{all}}_{\text{G1}}$	$A^{\text{ind}}_{\text{G1}}$	$A^{\text{all}}_{\text{G2}}$	$A^{\text{ind}}_{\text{G2}}$	$A^{\text{all}}_{\text{G1}}$	$A^{\text{ind}}_{\text{G1}}$	$A^{\text{all}}_{\text{G2}}$	$A^{\text{ind}}_{\text{G2}}$	$A^{\text{all}}_{\text{G1}}$	$A^{\text{ind}}_{\text{G1}}$	$A^{\text{all}}_{\text{G2}}$	$A^{\text{ind}}_{\text{G2}}$
LA1	0.84	0.83	0.87	0.86	0.85	0.85	0.80	0.85	0.86	0.83	0.91	0.88
LA2	0.92	0.92	0.87	0.86	0.89	0.85	0.76	0.76	0.90	0.91	0.80	0.86
Language	0.88	0.87	0.87	0.86	0.87	0.85	0.78	0.81	0.88	0.87	0.85	0.87
ME1	0.93	0.92	0.88	0.89	0.92	0.94	0.84	0.83	0.90	0.89	0.89	0.88
Metadata	0.93	0.92	0.88	0.89	0.92	0.94	0.84	0.83	0.90	0.89	0.89	0.88
MO1	0.87	0.87	0.82	0.82	0.87	0.84	0.82	0.81	0.87	0.80	0.82	0.71
MO2	0.89	0.91	0.78	0.81	0.88	0.87	0.77	0.77	0.89	0.87	0.84	0.75
MO3	**0.98**	**0.98**	**0.93**	0.93	**0.98**	**0.98**	**0.93**	**0.90**	**0.98**	**0.98**	0.93	**0.93**
MO4	0.82	0.82	0.84	0.84	0.82	0.81	0.84	0.81	0.82	0.82	0.84	0.84
MO5	0.76	0.76	0.58	0.58	0.76	0.71	0.58	0.54	0.75	0.75	0.56	0.56
MO6	0.93	0.92	0.90	0.91	0.88	0.92	0.82	0.84	0.91	0.90	0.89	0.90
MO7	0.60	0.56	0.69	0.65	0.71	0.55	0.58	0.67	0.60	0.45	0.69	0.65
Motivation	0.84	0.83	0.79	0.79	0.84	0.81	0.76	0.76	0.83	0.80	0.79	0.76
AP1	0.91	0.93	0.87	**0.93**	0.91	0.95	0.87	0.83	0.91	0.89	0.87	0.85
AP2	0.91	0.90	0.85	0.89	0.88	0.87	0.79	0.74	0.92	0.89	0.89	0.78
AP3	0.80	0.84	0.77	0.85	0.75	0.80	0.67	0.70	0.88	0.85	0.84	0.79
AP4	0.84	0.82	0.67	0.69	0.80	0.78	0.64	0.66	0.80	0.84	0.67	0.67
AP5	0.76	0.62	0.67	0.67	0.78	0.68	0.65	0.66	0.80	0.73	0.64	0.71
AP6	0.85	0.85	0.85	0.82	0.80	0.79	0.77	0.81	0.87	0.88	0.87	0.84
AP7	0.67	0.67	0.67	0.64	0.69	0.66	0.69	0.80	0.73	0.71	0.69	0.64
AP8	**0.96**	**0.96**	0.85	0.85	**0.96**	**0.98**	0.85	0.84	**0.96**	**0.96**	0.85	0.85
Approach	0.84	0.82	0.78	0.79	0.82	0.81	0.74	0.75	0.86	0.84	0.79	0.77
MD1	0.73	0.76	0.64	0.64	0.71	0.68	0.55	0.56	0.65	0.69	0.64	0.64
MD2	0.78	0.78	0.73	0.69	0.76	0.71	0.71	0.71	0.78	0.75	0.73	0.69
MD3	0.85	0.73	0.65	0.64	0.84	0.79	0.71	0.67	0.85	0.80	0.69	0.71
MD4	0.65	0.62	0.71	0.75	0.62	0.54	0.75	0.83	0.69	0.58	0.71	0.75
MD5	0.84	0.80	0.69	0.73	0.85	0.66	0.67	0.59	0.71	0.65	0.67	0.76
MD6	0.69	0.71	0.76	0.75	0.71	0.68	0.60	0.66	0.71	0.73	0.75	0.55
MD7	0.65	0.62	0.47	0.44	0.62	0.71	0.44	0.49	0.64	0.64	0.67	0.49
Methodology	0.74	0.72	0.66	0.66	0.73	0.68	0.63	0.64	0.72	0.69	0.69	0.65
SC1	**0.96**	**0.98**	**0.96**	**0.98**	**0.96**	**0.97**	**0.96**	**0.93**	**0.96**	**0.96**	**0.96**	**0.96**
SC2	**0.98**	**1.00**	**0.95**	**0.96**	**0.98**	**0.97**	**0.95**	**0.93**	**0.98**	**0.98**	**0.95**	**0.95**
SC3	0.91	0.91	**0.96**	**0.96**	0.91	0.93	**0.96**	**1.00**	0.91	0.93	**0.96**	**0.98**
SC4	0.84	0.90	0.91	0.89	0.80	0.85	0.78	0.85	0.89	0.89	**0.95**	0.91
Schedule	0.92	0.95	0.95	0.95	0.91	0.93	0.91	0.92	0.94	0.94	0.95	0.95
BI1	0.58	0.25	0.56	0.35	0.62	0.62	0.67	0.62	0.27	0.15	0.40	0.35
BI2	0.87	0.83	0.86	0.90	0.92	0.90	0.72	0.72	0.86	0.89	0.87	0.84
Bibliography	0.72	0.54	0.71	0.62	0.77	0.76	0.70	0.67	0.57	0.52	0.63	0.59
All criteria	0.83	0.81	0.78	0.78	0.82	0.80	0.75	0.75	0.82	0.79	0.79	0.76

Table 26: Quadratic-weighted agreement between closed-source LLM scores and human ratings for exposé criteria. Columns report A_{G1} (Group 1 vs. LLM) and A_{G2} (Group 2 vs. LLM) for the prompting variants shown (all, ind). Bold/underlined cells mark column-wise maxima/minima. Average rows summarize agreement within rubrics and across all criteria.

Code	Qwen3 (80B)				Llama 3.3 (70B)				GPT OSS (120B)				GPT OSS (20B)			
	$A^{\text{all}}_{\text{G1}}$	$A^{\text{ind}}_{\text{G1}}$	$A^{\text{all}}_{\text{G2}}$	$A^{\text{ind}}_{\text{G2}}$	$A^{\text{all}}_{\text{G1}}$	$A^{\text{ind}}_{\text{G1}}$	$A^{\text{all}}_{\text{G2}}$	$A^{\text{ind}}_{\text{G2}}$	$A^{\text{all}}_{\text{G1}}$	$A^{\text{ind}}_{\text{G1}}$	$A^{\text{all}}_{\text{G2}}$	$A^{\text{ind}}_{\text{G2}}$	$A^{\text{all}}_{\text{G1}}$	$A^{\text{ind}}_{\text{G1}}$	$A^{\text{all}}_{\text{G2}}$	$A^{\text{ind}}_{\text{G2}}$
SC1	0.83	0.83	0.85	0.84	0.84	0.85	0.88	0.89	0.85	0.83	0.87	0.87	0.86	0.88	0.88	0.91
SC2	0.92	0.92	0.80	0.81	0.90	0.86	0.81	0.80	0.91	0.86	0.83	0.85	0.84	0.85	0.85	0.83
Structure and Clarity	0.88	0.87	0.82	0.83	0.87	0.85	0.84	0.85	0.88	0.84	0.85	0.86	0.85	0.87	0.87	0.87
LG1	0.90	0.91	0.98	0.97	0.91	0.88	0.97	0.94	0.86	0.67	0.85	0.62	0.75	0.69	0.75	0.64
LG2	**0.98**	**0.98**	**0.99**	**0.98**	**0.98**	**0.98**	0.99	**0.99**	**0.98**	**0.99**	**0.99**	**1.00**	**0.96**	**0.97**	**0.98**	**1.00**
Language	0.94	0.94	0.99	0.97	0.94	0.93	0.98	0.97	0.92	0.83	0.92	0.81	0.85	0.83	0.87	0.82
FU1	0.75	0.70	0.74	0.69	0.74	0.78	0.72	0.73	0.64	0.63	0.57	0.55	0.66	0.55	0.60	0.48
FU2	0.75	0.71	0.73	0.69	0.76	0.78	0.72	0.75	0.64	0.66	0.62	0.65	0.58	0.61	0.59	0.60
Feed Up	0.75	0.71	0.73	0.69	0.75	0.78	0.72	0.74	0.64	0.64	0.59	0.60	0.62	0.58	0.59	0.54
FB1	0.94	0.94	**0.99**	**0.99**	0.94	0.92	**0.99**	0.96	0.23	0.10	0.20	0.06	0.25	0.09	0.22	0.05
FB2	0.92	0.93	0.97	0.96	0.91	0.93	0.95	0.96	0.24	0.09	0.19	0.05	0.08	0.10	0.04	0.04
FB3	0.79	0.79	0.73	0.73	0.79	0.78	0.73	0.74	0.77	0.62	0.73	0.63	0.75	0.59	0.69	0.58
FB4	0.84	0.84	0.96	0.96	0.84	0.84	0.96	0.96	0.83	0.74	0.93	0.78	0.84	0.84	0.96	0.91
FB5	0.95	0.95	0.94	0.94	0.95	0.94	0.94	0.92	0.95	0.92	0.94	0.92	0.95	0.76	0.94	0.79
FB6	**1.00**	**0.99**	0.97	0.96	**0.99**	**0.99**	0.98	**0.98**	0.96	0.83	0.95	0.82	0.34	0.41	0.37	0.42
Feed Back	0.91	0.91	0.93	0.92	0.90	0.90	0.93	0.92	0.66	0.55	0.66	0.54	0.54	0.47	0.54	0.46
FF1	0.96	**0.96**	**0.99**	**0.99**	0.96	**0.96**	**0.99**	**0.99**	0.96	0.90	**0.99**	0.92	0.91	0.71	0.90	0.70
FF2	0.78	0.78	0.61	0.61	0.78	0.78	0.61	0.61	0.75	0.37	0.62	0.50	0.78	0.36	0.71	0.51
Feed Forward	0.87	0.87	0.80	0.80	0.87	0.87	0.80	0.80	0.86	0.63	0.81	0.71	0.84	0.53	0.80	0.60
CQ1	0.96	0.94	0.92	0.92	0.94	0.88	0.93	0.86	0.92	**0.95**	0.90	0.92	0.84	0.90	0.84	0.90
CQ2	0.94	0.94	0.92	0.92	0.93	0.94	0.93	0.91	0.94	0.94	0.92	0.92	0.92	0.93	0.93	0.92
CQ3	0.77	**0.98**	0.80	**0.99**	**0.98**	**0.98**	**0.99**	**0.99**	**0.98**	**0.96**	**0.99**	**0.96**	**0.98**	**0.96**	**0.99**	**0.97**
Content Quality	0.89	0.95	0.88	0.94	0.95	0.93	0.95	0.92	0.95	0.95	0.94	0.93	0.91	0.93	0.92	0.93
ER1	0.83	0.63	0.88	0.68	0.83	0.44	0.88	0.45	0.75	0.65	0.77	0.70	0.80	0.82	0.83	0.87
ER2	0.92	0.74	0.92	0.75	0.93	0.65	0.95	0.61	0.89	0.91	0.89	0.91	0.86	**0.94**	0.88	**0.94**
ER3	**0.96**	0.91	0.97	0.92	0.96	0.93	0.97	0.94	**0.96**	0.92	0.97	**0.93**	**0.96**	0.92	**0.97**	0.93
ER4	0.80	0.25	0.84	0.26	0.94	0.17	0.98	0.13	0.42	0.32	0.38	0.32	0.36	0.36	0.38	0.34
ER5	**0.99**	0.95	**0.99**	0.95	**0.99**	0.94	**0.99**	0.94	**0.99**	**0.97**	**0.99**	**0.97**	**0.99**	**0.95**	**0.99**	**0.95**
Errors	0.90	0.69	0.92	0.71	0.93	0.63	0.95	0.62	0.80	0.75	0.80	0.77	0.79	0.80	0.81	0.81
All criteria	0.89	0.84	0.89	0.84	0.90	0.83	0.90	0.82	0.79	0.72	0.78	0.72	0.74	0.69	0.74	0.69

Table 27: Quadratic-weighted agreement between open-weight LLM scores and human ratings for review criteria. Columns report A_{G1} (Group 1 vs. LLM) and A_{G2} (Group 2 vs. LLM) for the prompting variants shown (all, ind). Bold/underlined cells mark column-wise maxima/minima. Average rows summarize agreement within rubrics and across all criteria.

Code	Claude Opus 4-6				Gemini 3.1 Pro				GPT-5.2			
	A^{all}_{G1}	A^{ind}_{G1}	A^{all}_{G2}	A^{ind}_{G2}	A^{all}_{G1}	A^{ind}_{G1}	A^{all}_{G2}	A^{ind}_{G2}	A^{all}_{G1}	A^{ind}_{G1}	A^{all}_{G2}	A^{ind}_{G2}
SC1	0.85	0.84	0.89	0.87	0.86	0.87	0.86	0.87	0.84	0.85	0.86	0.86
SC2	0.90	0.85	0.88	0.88	0.90	0.90	0.84	0.89	0.83	0.87	0.85	0.85
Structure and Clarity	0.87	0.85	0.88	0.87	0.88	0.88	0.85	0.88	0.84	0.86	0.86	0.85
LG1	0.81	0.80	0.80	0.83	0.85	0.81	0.93	0.91	0.91	0.83	0.92	0.89
LG2	**0.98**	**0.99**	**0.99**	**0.99**	**0.98**	**0.99**	0.99	**0.99**	**0.98**	**0.98**	**0.99**	**0.98**
Language	0.90	0.90	0.90	0.91	0.92	0.90	0.96	0.95	0.94	0.91	0.95	0.94
FU1	0.87	0.87	0.87	0.83	0.78	0.83	0.76	0.77	0.83	0.85	0.80	0.79
FU2	0.87	0.84	0.85	0.82	0.78	0.80	0.75	0.76	0.83	0.84	0.79	0.79
Feed Up	0.87	0.86	0.86	0.83	0.78	0.81	0.76	0.77	0.83	0.84	0.79	0.79
FB1	0.94	0.75	**0.99**	0.75	0.94	0.75	**0.99**	0.76	0.94	0.65	**0.99**	0.64
FB2	0.93	0.75	0.94	0.74	0.94	0.90	0.95	0.96	0.93	0.28	0.92	0.22
FB3	0.78	0.75	0.74	0.71	0.74	0.60	0.71	0.62	0.62	0.42	0.65	0.47
FB4	0.84	0.85	0.96	0.97	0.84	0.83	0.96	**0.99**	0.84	0.84	0.96	0.96
FB5	0.95	0.95	0.94	0.94	0.95	0.96	0.94	0.94	0.94	0.95	0.93	0.94
FB6	0.95	**0.97**	0.96	0.98	**0.99**	**0.99**	0.96	0.97	**0.97**	0.85	0.94	0.84
Feed Back	0.90	0.84	0.92	0.85	0.90	0.84	0.92	0.87	0.88	0.66	0.90	0.68
FF1	**0.96**	0.96	**0.99**	**0.99**	0.96	0.97	**0.99**	**0.99**	0.95	**0.96**	0.98	**0.99**
FF2	0.81	0.79	0.68	0.61	0.81	0.73	0.70	0.69	0.79	0.79	0.68	0.63
Feed Forward	0.89	0.88	0.83	0.80	0.89	0.85	0.84	0.84	0.87	0.88	0.83	0.81
CQ1	0.92	0.86	0.92	0.89	0.96	0.96	0.92	0.92	0.80	0.80	0.83	0.83
CQ2	0.92	0.94	0.94	0.93	0.94	0.93	0.92	0.91	0.94	0.95	0.92	0.92
CQ3	**0.97**	**0.98**	0.99	**0.99**	**0.98**	**0.98**	**0.99**	**0.99**	**0.98**	**0.98**	**0.99**	**0.99**
Content Quality	0.94	0.92	0.95	0.93	0.96	0.96	0.94	0.94	0.91	0.91	0.92	0.91
ER1	0.86	0.84	0.91	0.85	0.84	0.84	0.89	0.88	0.75	0.79	0.79	0.83
ER2	0.85	0.84	0.85	0.82	0.93	**0.97**	0.93	0.92	0.85	0.90	0.83	0.88
ER3	0.94	0.84	0.92	0.79	0.95	0.85	0.96	0.84	0.86	0.81	0.85	0.78
ER4	0.85	0.42	0.87	0.38	0.94	0.93	0.98	0.98	0.79	0.22	0.81	0.18
ER5	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**	0.91	**0.99**	0.91	**0.99**	**0.96**	**0.99**	**0.96**
Errors	0.90	0.78	0.91	0.77	0.93	0.90	0.95	0.91	0.85	0.73	0.85	0.73
All criteria	0.90	0.85	0.90	0.84	0.90	0.88	0.91	0.88	0.87	0.79	0.88	0.78

Table 28: Quadratic-weighted agreement between closed-source LLM scores and human ratings for review criteria. Columns report A_{G1} (Group 1 vs. LLM) and A_{G2} (Group 2 vs. LLM) for the prompting variants shown (all, ind). Bold/underlined cells mark column-wise maxima/minima. Average rows summarize agreement within rubrics and across all criteria.

Domain	Opus-4.6		Gem-3.1P		GPT-5.2		OSS-120		OSS-20		L3.3-70		Qwen-80	
	All	**Ind**	**All**	**Ind**	**All**	**Ind**	**All**	**Ind**	**All**	**Ind**	**All**	**Ind**	**All**	**Ind**
Exposé criterion-based scoring (over 31 criteria)														
0-1 (#21)	0.06	0.07	0.08	0.06	0.05	0.08	0.09	0.07	0.10	0.07	0.14	0.16	0.15	0.15
0-1-2 (#10)	0.08	0.09	0.22	0.23	0.10	0.13	0.09	0.21	0.10	0.26	0.24	0.35	0.24	0.26
All (#31)	**0.07**	**0.08**	**0.13**	**0.12**	**0.07**	**0.10**	**0.09**	**0.11**	**0.10**	**0.13**	**0.17**	**0.22**	**0.18**	**0.19**
Review criterion-based scoring (over 22 criteria)														
0-1 (#10)	0.05	0.06	0.04	0.03	0.04	0.11	0.12	0.19	0.19	0.22	0.07	0.06	0.07	0.07
0-1-2 (#7)	0.07	0.12	0.12	0.10	0.24	0.19	0.09	0.10	0.08	0.09	0.11	0.11	0.21	0.21
-1-0 (#5)	0.03	0.10	0.03	0.03	0.03	0.14	0.07	0.11	0.06	0.11	0.03	0.16	0.04	0.08
All (#22)	**0.05**	**0.09**	**0.06**	**0.06**	**0.10**	**0.14**	**0.10**	**0.14**	**0.13**	**0.15**	**0.08**	**0.10**	**0.11**	**0.12**

Table 29: Weighted-mean Jensen–Shannon divergence (JSD) between human-vote and LLM-vote distributions, grouped by domain. Prompting variants: All, Ind. Lower is better. Model abbreviations: *Opus-4.6*=Claude Opus 4.6; *Gem-3.1P*=Gemini 3.1 Pro; *GPT-5.2*=GPT 5.2; *OSS-120*=GPT OSS (120B); *OSS-20*=GPT OSS (20B); *L3.3-70*=Llama 3.3 (70B); *Qwen-80*=Qwen 3 (80B).

Model	Mode	s/assessment	s/prompt	tokens/prompt	tokens/input	tokens/output
GPT OSS (20B)	All	110.20 (±27.69)	110.20 (±27.69)	8,010.67 (±1,018.78)	5,587.85 (±932.59)	2,422.82 (±629.58)
	[min–max]	[84.88–289.55]	[84.88–289.55]	[6,384.00–11,178.00]	[4,092.00–8,765.00]	[1,874.00–6,489.00]
Llama 3.3 (70B)	All	182.10 (±12.79)	182.10 (±12.79)	7,181.78 (±973.95)	5,606.69 (±947.16)	1,575.09 (±105.46)
	[min–max]	[163.18–212.11]	[163.18–212.11]	[5,594.00–10,311.00]	[4,091.00–8,839.00]	[1,399.00–1,858.00]
GPT OSS (120B)	All	258.01 (±13.74)	258.01 (±13.74)	8,227.11 (±915.79)	5,679.85 (±932.59)	2,547.25 (±148.73)
	[min–max]	[227.52–304.39]	[227.52–304.39]	[6,692.00–11,292.00]	[4,184.00–8,857.00]	[2,265.00–3,067.00]
Llama 3.3 (70B)	Ind	284.30 (±26.54)	9.17 (±1.37)	3,228.31 (±948.58)	3,162.76 (±947.61)	65.55 (±9.79)
	[min–max]	[237.56–370.23]	[5.43–14.43]	[1,656.00–6,555.00]	[1,613.00–6,490.00]	[42.00–117.00]
Qwen3 (80B)	All	430.37 (±15.13)	430.37 (±15.13)	7,473.75 (±1,046.99)	5,719.84 (±1,024.62)	1,753.91 (±64.05)
	[min–max]	[394.29–460.62]	[394.29–460.62]	[5,775.00–11,079.00]	[4,126.00–9,301.00]	[1,602.00–1,874.00]
GPT OSS (20B)	Ind	582.36 (±103.30)	18.79 (±20.85)	3,553.87 (±1,047.56)	3,162.31 (±933.03)	391.56 (±478.39)
	[min–max]	[452.01–878.16]	[7.34–357.11]	[1,821.00–11,982.00]	[1,633.00–6,431.00]	[123.00–8,091.00]
Qwen3 (80B)	Ind	862.02 (±36.97)	27.81 (±3.81)	3,359.56 (±1,027.13)	3,275.90 (±1,025.04)	83.66 (±15.20)
	[min–max]	[789.80–946.85]	[19.10–41.08]	[1,705.00–7,050.00]	[1,648.00–6,952.00]	[51.00–138.00]
GPT OSS (120B)	Ind	1133.41 (±89.76)	34.35 (±6.90)	3,383.58 (±932.15)	3,160.07 (±933.04)	223.51 (±67.47)
	[min–max]	[869.61–1398.68]	[19.20–97.15]	[1,771.00–6,902.00]	[1,627.00–6,431.00]	[100.00–854.00]

Table 30: Runtime and token usage for LLM scoring. Entries report mean (± SD) and a second row with [min–max]. Mode: All = combined prompting; Ind = single-criterion prompting. Token statistics are reported per prompt (total, input, output).

Model	Mode	s/assessment	s/prompt	tokens/prompt	tokens/input	tokens/output
GPT OSS (20B)	All	80.25 (±20.05)	80.25 (±20.05)	4,465.78 (±606.62)	2,539.31 (±402.22)	1,926.47 (±472.59)
	[min–max]	[60.56–345.62]	[60.56–345.62]	[3,641.00–10,957.00]	[2,117.00–6,330.00]	[1,449.00–8,091.00]
Llama 3.3 (70B)	All	113.76 (±8.10)	113.76 (±8.10)	3,574.05 (±434.77)	2,513.88 (±404.74)	1,060.17 (±73.19)
	[min–max]	[98.39–135.10]	[98.39–135.10]	[3,041.00–7,464.00]	[2,088.00–6,329.00]	[920.00–1,264.00]
Llama 3.3 (70B)	Ind	142.31 (±11.27)	6.47 (±0.95)	982.85 (±406.97)	924.60 (±405.24)	58.25 (±8.26)
	[min–max]	[121.03–210.90]	[4.26–11.25]	[518.00–4,845.00]	[471.00–4,774.00]	[40.00–93.00]
GPT OSS (120B)	All	194.64 (±14.13)	194.64 (±14.13)	4,485.63 (±447.10)	2,540.12 (±403.33)	1,945.51 (±146.74)
	[min–max]	[167.25–252.35]	[167.25–252.35]	[3,823.00–8,242.00]	[2,117.00–6,330.00]	[1,670.00–2,555.00]
Qwen3 (80B)	All	289.58 (±13.70)	289.58 (±13.70)	3,705.31 (±483.11)	2,567.16 (±460.99)	1,138.15 (±50.89)
	[min–max]	[254.40–359.69]	[254.40–359.69]	[3,087.00–7,550.00]	[2,066.00–6,295.00]	[1,004.00–1,361.00]
GPT OSS (20B)	Ind	343.64 (±81.03)	15.62 (±15.37)	1,316.56 (±561.36)	955.08 (±402.72)	361.47 (±363.39)
	[min–max]	[240.61–792.63]	[4.91–344.67]	[687.00–9,596.00]	[505.00–4,779.00]	[108.00–8,091.00]
Qwen3 (80B)	Ind	485.12 (±25.59)	22.05 (±2.33)	970.92 (±407.79)	902.12 (±405.52)	68.80 (±9.00)
	[min–max]	[421.38–576.16]	[15.73–40.33]	[516.00–4,822.00]	[449.00–4,740.00]	[45.00–145.00]
GPT OSS (120B)	Ind	512.82 (±88.18)	23.31 (±9.87)	1,171.90 (±420.31)	955.08 (±402.72)	216.81 (±96.63)
	[min–max]	[375.80–1048.64]	[8.89–465.17]	[634.00–5,883.00]	[505.00–4,779.00]	[85.00–4,827.00]

Table 31: Runtime and token usage for LLM scoring. Entries report mean (± SD) and a second row with [min–max]. Mode: All = combined prompting; Ind = single-criterion prompting. Token statistics are reported per prompt (total, input, output).