Developing AI-Compatible Instructional Design Datasets

Demetre - Nadiradze 22100577@ibsu.edu.ge
Computer Science/ School of Computer Science and Architecture, 3rd

Furqan - Zulva furqan_al@mhs.usk.ac.id
Informatics/Faculty of Mathematics and Natural Science, 4th

Umar - Farzan muf359@uowmail.edu.au
Computer Science/Faculty of Engineering and Applied Sciences, 3rd

Sohaila - Abdalgwad sohaila.abdalgwad@ejust.edu.eg
Computer Engineering/Faculty of Engineering and Applied Sciences, 4th

Emre - Usta emre.usta@sabanciuniv.edu
Computer Science/Faculty of Engineering and Natural Sciences, 4th

Ilkem – Dipcin
School of Languages

Abstract

This project develops a structured six-stage pipeline to transform diverse corporate training materials into pedagogically rigorous, machine-readable datasets for fine-tuning large language models. Integrating Bloom's Taxonomy and ADDIE with modern NLP methods, the workflow includes data gathering, compression, pedagogical structuring, schema formatting, model fine-tuning, and evaluation. A dual-model strategy—combining open-weight (Mistral Small 3.2) and API-based (GPT-4.1 mini) fine-tuning—demonstrated scalability, efficiency, and strong pedagogical alignment. Evaluation using a rubric on prompt adherence, completeness, clarity, and instructional soundness confirmed high-quality outputs. The research offers a research direction for scalable, standards-aligned corporate training design.

Keywords: Instructional Design, Large Language Models, Corporate Training, Fine-Tuning, Pedagogy

1 Introduction

Instructional design (ID) is defined as the systematic process through which educational experiences are crafted to facilitate effective learning. Classic frameworks—such as ADDIE (Molenda, 2003) and Merrill's Principles of Instruction (Merrill, 2002)—have traditionally been employed to ensure alignment among learning objectives, content, and assessments. Although these models provide pedagogical rigor, inherent limitations arise due to the manual efforts required for gathering, organizing, and adapting educational materials, which reduce scalability and responsiveness to learners' evolving needs (Reiser & Dempsey, 2018).

With the advent of artificial intelligence (AI), and particularly large language models (LLMs), new opportunities have emerged for enhancing the efficiency of instructional design without compromising educational quality. However, existing AI-driven tools often prioritize automation and expediency at the expense of transparency, pedagogical alignment, and educator control (Holmes et al., 2021). This tension highlights a critical gap in the literature: the need for AI systems that support, rather than supplant, human instructional expertise while maintaining technical integrity and educational accountability.

State-of-the-art frameworks have begun to address these challenges. The ARCHED (AI for Responsible, Collaborative, Human-centered Education Instructional Design) framework introduces a structured, multi-stage workflow in which educators retain primary decision-making authority. In this approach, LLMs assist by generating pedagogical options and evaluating their alignment with learning objectives—guided by Bloom's taxonomy—thereby enhancing transparency and human agency in AI-assisted design (Li et al., 2025). This framework comprises three main phases: the Learning Objective Generation System (LOGS), which creates candidate learning objectives; the Objective Analysis Engine (OAE), which evaluates those objectives based on pedagogical criteria; and an assessment phase, which helps in choosing, refining, and finalizing content. Similarly, GAIDE (Generative AI for Instructional Development and Education) embeds generative AI into curriculum development, focusing on guiding educators through prompt engineering and validation processes within a constructivist and TPCK-informed model (Dickey & Bejarano, 2024). While ARCHED and GAIDE demonstrate valuable principles for human-AI collaboration and pedagogical grounding, they also expose a persistent limitation: the insufficiency of high-quality, domain-diverse datasets and standardized data formats necessary to train and evaluate educational AI systems effectively.

The current project has been designed to address this foundational gap by focusing on two complementary objectives: (1) the creation and curation of high-quality instructional-design datasets, and (2) the conversion and alignment of collected materials into machine-readable formats—such as JSONL—compatible with modern AI training pipelines. Systematic approaches have been developed for the collection, classification, and validation of instructional content across multiple domains, incorporating ethical web scraping, annotation schemas for instructional patterns, and metrics for dataset quality assessment. Concurrently, pipelines are being engineered to clean, structure, and validate these materials, ensuring compliance with consistent schemas and facilitating alignment with AI model requirements.

By centering on the data infrastructure that underpins AI-empowered instructional design, this work makes a distinctive contribution within the emerging landscape of educational technology. While frameworks like ARCHED and GAIDE focus on human-AI interaction workflows, the emphasis here lies on establishing the data foundations necessary for those workflows to operate effectively and to fine-tune existing models for task automation and improved responses. Thus, this research advances the responsible integration of AI into instructional design by bridging the gap between visionary frameworks and the data realities required for their realization.

2 Methodology

2.1 Data Gathering and Formatting

The initial stage of this process involved collecting relevant documents, blog posts, and articles across the following topics: problem-solving, communication and time management. The targeted search was carried out using the AI-powered discovery tool Perplexity. The search specifically uses the Deep Research version to ensure comprehensive and high-quality source discovery across the internet. The selection was limited to openly accessible materials. This filtering process ensured that only well-curated and reliable sources were included in the research dataset.

Following selection, each source underwent a text extraction process to convert its content into a raw, unstructured text format. For larger sources, a data segmentation step was performed. Instead of treating an entire source as a single data point, its raw text was logically divided into smaller, coherent chunks, typically by chapter or thematic section. This approach allowed a single comprehensive source document to be methodically broken down, yielding multiple focused data samples and thereby significantly increasing the dataset's size and topical granularity.

The core stage of the process involved a structural transformation of the data, which was automated through a custom workflow built on the n8n platform and integrated with the Google Gemini API. Within this workflow, the Large Language Model functioned as an intelligent parser and reformatter. For each segmented text chunk, the LLM analyzed the unstructured input and reorganized it into a predefined three-part structure: an Objective, a Training/Learning Content section, and a Practice/Assessment component. To balance capability and efficiency, the workflow employed two models: Gemini 2.5 Pro for complex text analysis and restructuring, and Gemini 2.0 Flash for more straightforward reformatting tasks.

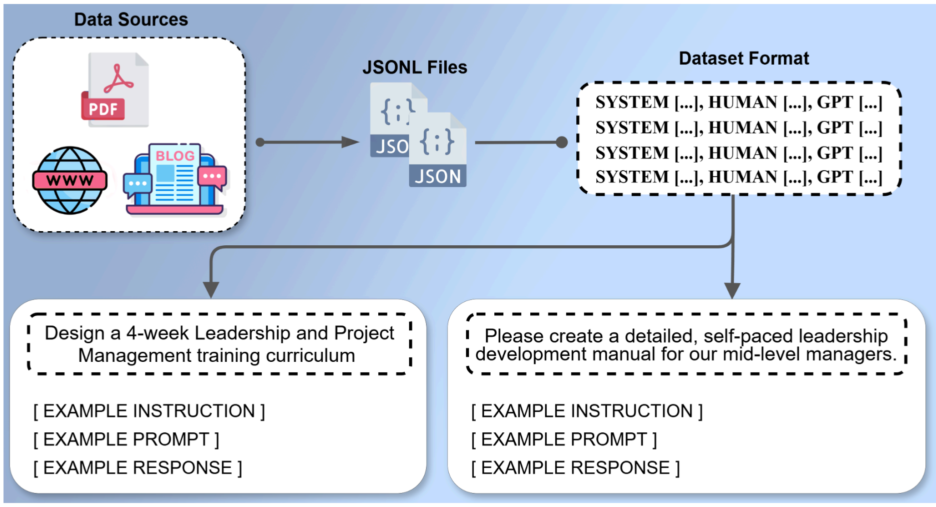

Data Sources
PDF
WWW
BLOG
JSONL Files
JSON
JSON
Dataset Format
SYSTEM [...], HUMAN [...], GPT [...]
SYSTEM [...], HUMAN [...], GPT [...]
SYSTEM [...], HUMAN [...], GPT [...]
SYSTEM [...], HUMAN [...], GPT [...]
Design a 4-week Leadership and Project Management training curriculum
[EXAMPLE INSTRUCTION]
[EXAMPLE PROMPT]
[EXAMPLE RESPONSE]
Please create a detailed, self-paced leadership development manual for our mid-level managers.
[EXAMPLE INSTRUCTION]
[EXAMPLE PROMPT]
[EXAMPLE RESPONSE]

Figure 1. Data transformation diagram

In the final stage, the structured modules were formatted to meet the technical requirements for LLM fine-tuning. The complete dataset was compiled into a JSONL file, with each line representing a single training example. A conversational structure was adopted for each JSON object, containing a conversations array that simulates a dialogue: the system turn defines the AI's persona as a professional learning assistant, the human turn contains the user prompt, and the gpt turn provides the detailed, structured learning module as the ideal response as shown in figure 1. To maintain technical integrity, all newline characters were correctly encoded and this ensured full compatibility of the dataset for the fine-tuning process.

2.2 Model Selection And Fine Tuning

We employed a comprehensive multi-stage, multi-model methodology to systematically convert heterogeneous corporate training materials into machine-readable, pedagogically sound datasets suitable for fine-tuning large language models (LLMs). The methodology was designed as a six-stage sequential process that integrates instructional design principles with state-of-the-art natural language processing techniques to ensure the production of trustworthy, high-quality corporate training content generators.

The first stage, data curation, involved the systematic assembly and digitization of training materials organized by domain expertise. This was followed by content compression, which focused on redundancy removal while maintaining instructional fidelity and pedagogical integrity. The third stage implemented pedagogical structuring, where content was organized into conventional instructional formats aligned with established educational frameworks. Schema transformation constituted the fourth stage, converting the structured content into JSONL format to ensure compatibility with fine-tuning procedures. The fifth stage involved the actual model fine-tuning process, applied to both open-weight and API-based LLMs. Finally, the sixth stage encompassed comprehensive evaluation and iterative optimization, focusing on benchmarking and optimizing model outputs for pedagogical quality and structural accuracy.

This pipeline deliberately integrates established instructional design principles, including Bloom's Taxonomy and the ADDIE framework, with contemporary NLP preprocessing methodologies. The integration ensures that fine-tuned models can serve as reliable generators of high-quality corporate training content while maintaining pedagogical rigor and practical applicability.

2.3 Model Choices

The research adopted a two-track model approach designed to maximize both performance diversity and operational flexibility. The first track utilized open-weight models, specifically the Unsloth/Mistral-Small-3.2 Instruct model, which was selected for its exceptional inference efficiency, demonstrating 2.2× faster processing speeds and 62% reduced memory consumption through Unsloth optimization. This model choice provided the additional advantage of low-level parameter control during the fine-tuning process, enabling precise customization for domain-specific requirements.

The second track employed API-hosted models, primarily OpenAI's GPT-4.1-mini, with exploratory experiments conducted using GPT-4o-mini and gpt oss. These models were chosen for their superior instruction compliance capabilities and robust zero-shot and few-shot performance, while eliminating the need for local infrastructure overhead. This hybrid approach yielded several strategic advantages that enhanced the overall research methodology.

The performance diversity achieved through multiple architecture cross-validation significantly reduced potential biases and improved the pedagogical robustness of generated outputs. From a cost-performance optimization perspective, local inference minimized long-term operational costs while API inference enabled rapid prototyping and iterative development. The infrastructure flexibility provided by this dual approach allowed for tokenization customization and management of longer sequences through local installation, while API access ensured operational dependability and scalability.

Furthermore, this multi-model strategy addressed the common challenge in corporate training content development, where theoretical abstraction often conflicts with pragmatic application requirements. The use of multiple models with different architectural foundations improved alignment to both theoretical frameworks and practical implementation needs, resulting in more balanced and comprehensive training materials.

2.4 Data Curation

The foundation dataset was constructed from a carefully curated collection of heterogeneous corporate learning materials spanning multiple domains and instructional contexts. The primary sources included proprietary leadership and management development manuals obtained from established corporations, internal HR training materials and comprehensive onboarding guides that reflected current industry practices, and transcripts from Massive Open Online Courses (MOOCs) and professional certification programs. Additionally, the dataset incorporated peer-reviewed literature in workplace learning, leadership development, and communication skills to ensure theoretical grounding and academic rigor.

The selection process was governed by three critical inclusion criteria that ensured dataset quality and relevance. Domain relevance required that all content directly addressed corporate soft skills, leadership competencies, communication effectiveness, or organizational development principles. Instructional completeness mandated the presence of clearly articulated learning objectives, comprehensive conceptual material, and well-designed assessment mechanisms. Finally, textual clarity was enforced to avoid ambiguous or poorly structured text that would increase preprocessing overhead and potentially compromise the quality of fine-tuned outputs.

2.5 Content Compression

Given the significant variation in input material length, ranging from 5 to over 300 pages per source, a systematic two-pass compression process was implemented to ensure consistency and quality. The first pass employed LLM-assisted summarization using carefully engineered prompts designed to extract critical learning objectives, key conceptual frameworks, and viable practical applications while preserving the essential pedagogical structure of the original materials.

The second pass involved comprehensive human validation conducted by domain experts who reviewed the compressed content to ensure instructional integrity and remove outdated information or figures that might compromise the contemporary relevance of the training materials. The compression process was calibrated to produce training materials of approximately 10 pages or fewer, achieving an optimal balance between token efficiency and conceptual richness that would support effective fine-tuning while maintaining comprehensive coverage of essential learning objectives.

2.6 Pedagogical Organization

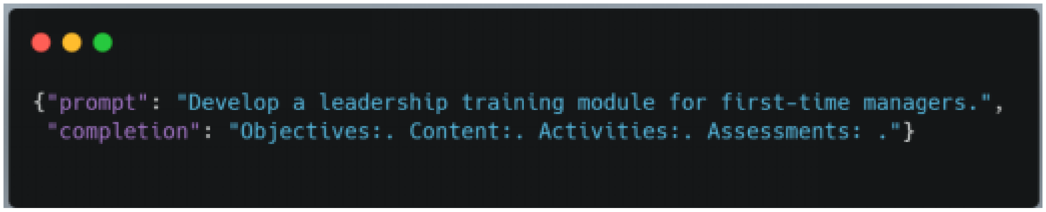
The compressed content underwent systematic restructuring into four standardized instructional sections designed to align with established pedagogical frameworks and ensure consistency across all training modules. The objectives section contained learning objectives stated in measurable, behavioral terms and aligned with appropriate levels of Bloom's Taxonomy to ensure cognitive progression and assessment validity. The training content section encompassed core conceptual material and theoretical frameworks presented in a logical, sequential manner that supported progressive skill development.

The activities section incorporated experiential learning opportunities, including role-playing exercises, problem-solving tasks, and scenario-based applications designed to bridge theoretical knowledge and practical implementation. Finally, the assessments section included comprehensive knowledge tests, project-based evaluation prompts, and scenario-based questions that measured learning achievement across multiple cognitive domains.

This structural approach provided several significant benefits for the fine-tuning process. It ensured content consistency that supported uniform learning patterns across all LLM training instances, while embedding pedagogical encoding directly within the dataset to improve the quality of model-generated lesson plans and instructional materials. The standardized structure also facilitated more effective evaluation and comparison of model outputs across different training scenarios and domains.

2.7 Schema Transformation (JSONL Encoding)

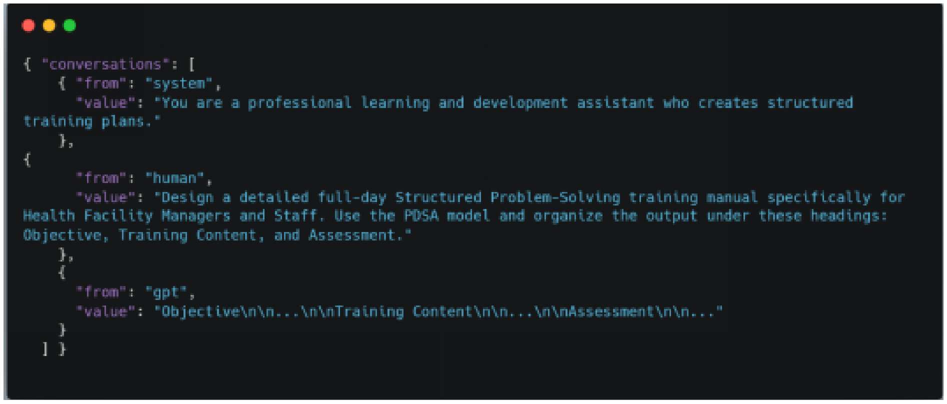
Each structured module was converted to a JSONL object for **Open AI** with:



```
{"prompt": "Develop a leadership training module for first-time managers.",
 "completion": "Objectives:. Content:. Activities:. Assessments: ."}
```

Figure 2. Prompt completion format

Each structured module was converted to a JSONL object for **Mistral** with:



```
{ "conversations": [
    { "from": "system",
      "value": "You are a professional learning and development assistant who creates structured
training plans."
    },
{
      "from": "human",
      "value": "Design a detailed full-day Structured Problem-Solving training manual specifically for
Health Facility Managers and Staff. Use the PDSA model and organize the output under these headings:
Objective, Training Content, and Assessment."
    },
    {
      "from": "gpt",
      "value": "Objective\n\n...\n\nTraining Content\n\n...\n\nAssessment\n\n..."
    }
  ] }
```

Figure 3. JSONL object

The JSONL format provided several critical advantages for the fine-tuning process. It ensured model compatibility across both Mistral fine-tuning pipelines and OpenAI's API infrastructure, while providing format uniformity that significantly reduced parsing errors and improved training stability. The structured format also facilitated more effective batch processing and quality control during the fine-tuning procedures.

2.8 Model Fine-tuning

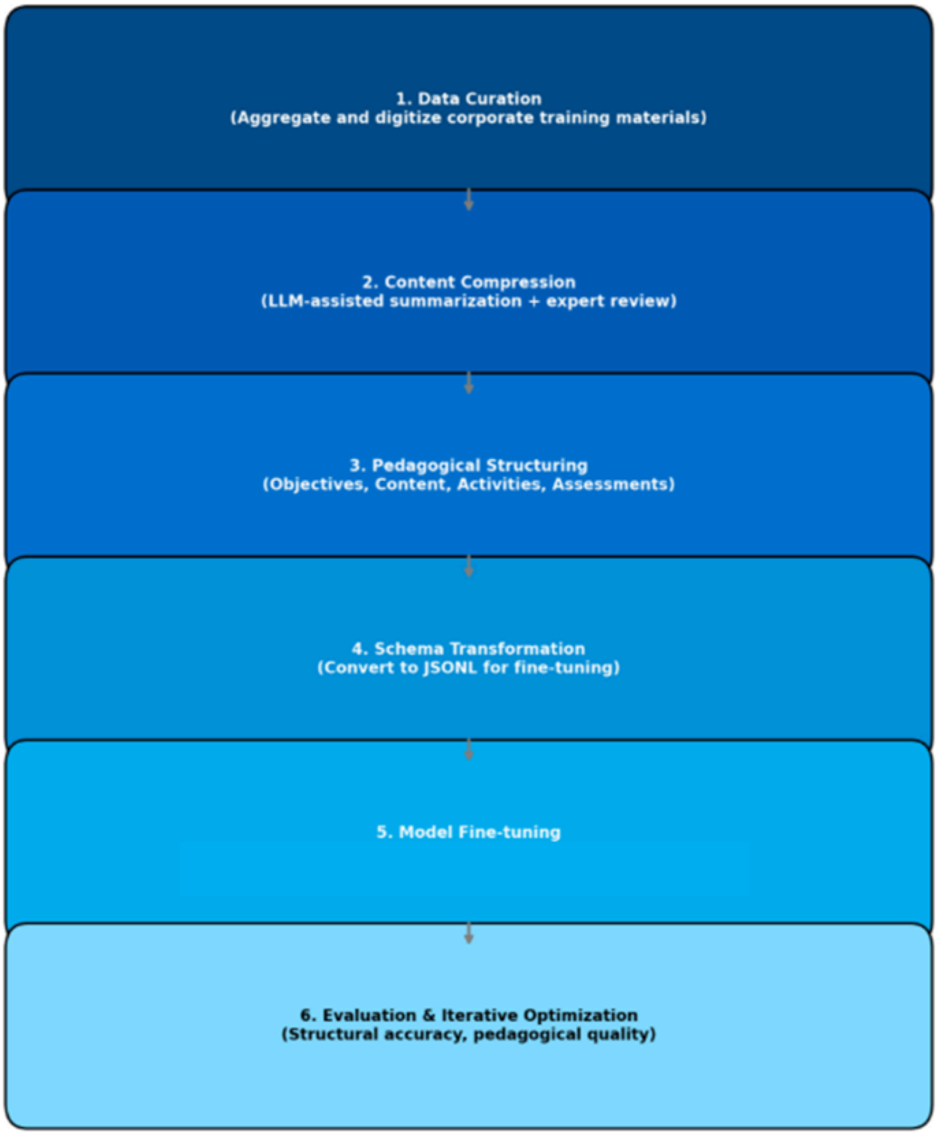
The open-weight fine-tuning process was conducted using the Unsloth/Mistral-Small-3.2 Instruct model on a cloud GPU infrastructure with 40 GB VRAM capacity. The implementation utilized the Unsloth accelerated fine-tuning framework combined with Parameter Efficient Fine-Tuning (PEFT) and Low-Rank Adaptation (LoRA) techniques to optimize computational efficiency while maintaining model performance. A SentencePiece tokenizer was trained specifically on the corpus to preserve domain-specific vocabulary and ensure accurate representation of corporate training terminology.

The hyperparameter configuration was optimized through preliminary experiments to balance training effectiveness with computational efficiency. The AdamW optimizer was employed with a learning rate of 5e-5, selected to provide stable convergence while avoiding overfitting. Training was conducted over up to 5 epochs with a batch size of 8 sequences per device and a maximum sequence length of 12288 tokens. The primary objectives focused on maximizing instructional coherence, minimizing hallucination in generated content, and improving adherence to the specified output formatting requirements.

The API fine-tuning process utilized OpenAI's GPT-4.1-mini model through the dedicated fine-tuning endpoint, employing the same JSONL corpus with appropriate length verification to ensure compliance with token constraints. The training configuration included 4 epochs with API-controlled hyperparameter optimization and a 10% validation set held out for performance monitoring. The primary objectives for API fine-tuning emphasized enhancing zero-shot instructional structuring capabilities and improving the quality of evaluation questions and assessment materials.

2.9 Summary of Workflow

The hybrid solution transformed raw, unstructured corporate training materials into a high-fidelity fine-tuning data set to teach LLMs how to produce pedagogically rigorous, fully structured training material on demand. The hybrid deployment strategy — combining Mistral-Small-3.2 Instruct locally fine-tuned and OpenAI GPT-4.1-mini fine-tuned via the API — provided both production scalability and cost-effective long-term deployment (figure 4).



1. Data Curation
(Aggregate and digitize corporate training materials)

2. Content Compression
(LLM-assisted summarization + expert review)

3. Pedagogical Structuring
(Objectives, Content, Activities, Assessments)

4. Schema Transformation
(Convert to JSONL for fine-tuning)

5. Model Fine-tuning

6. Evaluation & Iterative Optimization
(Structural accuracy, pedagogical quality)

Figure 4. Methodology workflow

3 Results

3.1 Evaluation Criteria

Model outputs underwent systematic verification using both quantitative performance criteria and qualitative evaluation standards established in consultation with supervising faculty. The evaluation framework focused on three critical dimensions that collectively assessed the pedagogical effectiveness and technical performance of the fine-tuned models.

The evaluation of the three fine-tuned models was based on a series of prompt-based tests, and the results were evaluated by project members. In order to do this, three variables were introduced: response length (measured in tokens, higher is better), response speed (measured in tokens per second, higher is better), and response quality (higher is better).

To evaluate response quality, we employ a rubric that measures how closely responses align with prompt requirements and instructional design principles. This rubric assesses four key dimensions: prompt adherence, completeness, clarity & readability, and pedagogical soundness—each scored on a 1–5 scale.

Prompt Adherence evaluates how well the response follows explicit instructions, including tone, audience, content focus, and constraints.

- 1 (Poor): Response is generic, off-topic, or disregards instructions.
- 2 (Fair): Addresses the topic but omits major details or customizations.
- 3 (Good): Relevant and mostly aligned, but misses some nuance.
- 4 (Very Good): Closely follows instructions with only minor gaps.
- 5 (Excellent): Fully tailored to the prompt with precise compliance.

Completeness measures whether all requested components (objectives, content, activities, assessments) are present and sufficiently developed.

- 1 (Poor): Major sections or concepts are missing.
- 2 (Fair): Covers only basic elements, with significant gaps.
- 3 (Good): Includes most components but lacks depth in places.
- 4 (Very Good): Covers all required elements with solid detail.
- 5 (Excellent): Comprehensive, fully developed, and thorough.

Clarity & Readability assesses how well the response is written, organized, and matched to the target audience.

- 1 (Poor): Confusing, disorganized, or inappropriate language.
- 2 (Fair): Understandable but with several clarity issues.
- 3 (Good): Generally clear and well-organized, with minor issues.
- 4 (Very Good): Highly readable, coherent, and logically structured.
- 5 (Excellent): Exceptionally clear, polished, and easy to follow.

Pedagogical Soundness evaluates whether activities and assessments logically align with stated learning objectives.

- 1 (Poor): No clear alignment; activities/assessments unrelated to objectives.
- 2 (Fair): Minimal alignment; significant redesign required.
- 3 (Good): Basic alignment present but could be stronger.
- 4 (Very Good): Strong alignment with minor refinements needed.
- 5 (Excellent): Activities and assessments fully reinforce and validate objectives.

Each evaluator independently reviewed and scored every sample against the four rubric dimensions. After all individual ratings were collected, we averaged the scores for each model.

3.2 Performance results

In order to benchmark the different models, they were fine-tuned on an identical 600-sample dataset, addressing three types of training: communication skills, time-management and problem-solving.

Measured against the best-performing model, GPT-4.1 mini, GPT-OSS regularly returned significantly lower quality results and fell behind in response speed (figure 5). Mistral Small 3.2 managed to achieve 78% of the response quality, 49% of the response length and 16% of the response speed when compared to GPT-4.1 mini. The data suggests a large performance difference between open-source, proprietary and reasoning models.

In order to measure the effect dataset size has on model responses, the Mistral Small 3.2 model was selected for additional fine-tuning. The model was fine-tuned on three 200 sample sub-datasets, each addressing one of the previously mentioned training areas. While the results differed depending on the dataset, in all cases an increase in response speed was observed. In the example of the time-management dataset (figure 7), response length was also improved.

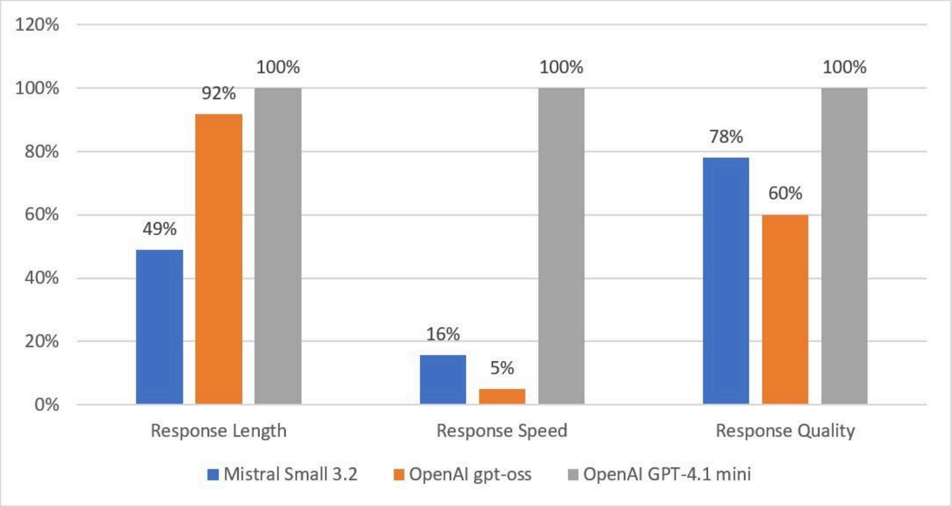


Figure 5. Comparison of various models trained on 600 samples

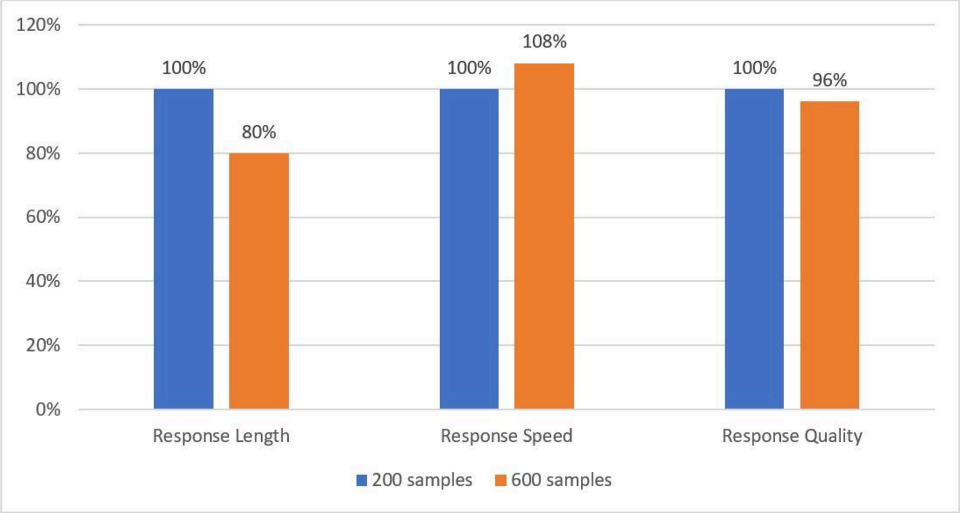


Figure 6. Comparison of communication dataset and final dataset

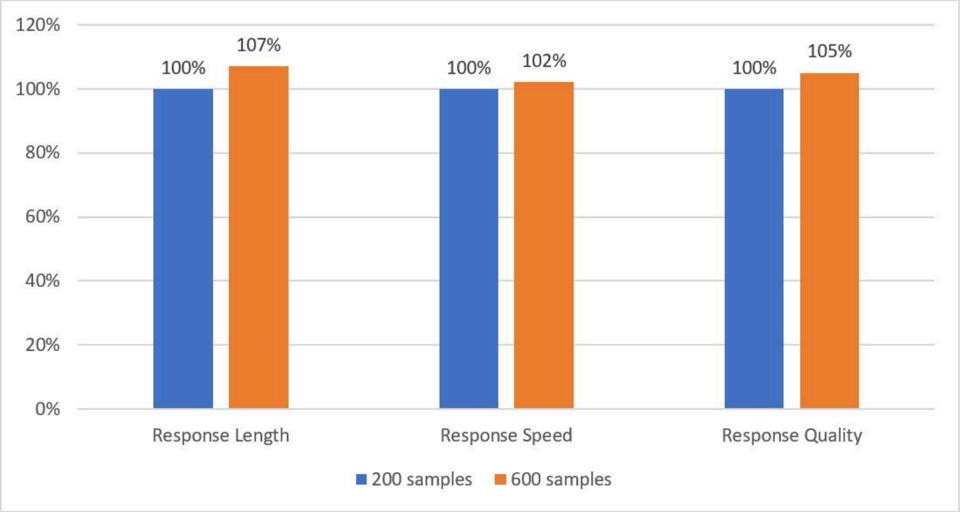


Figure 7. Comparison of time management dataset and final dataset

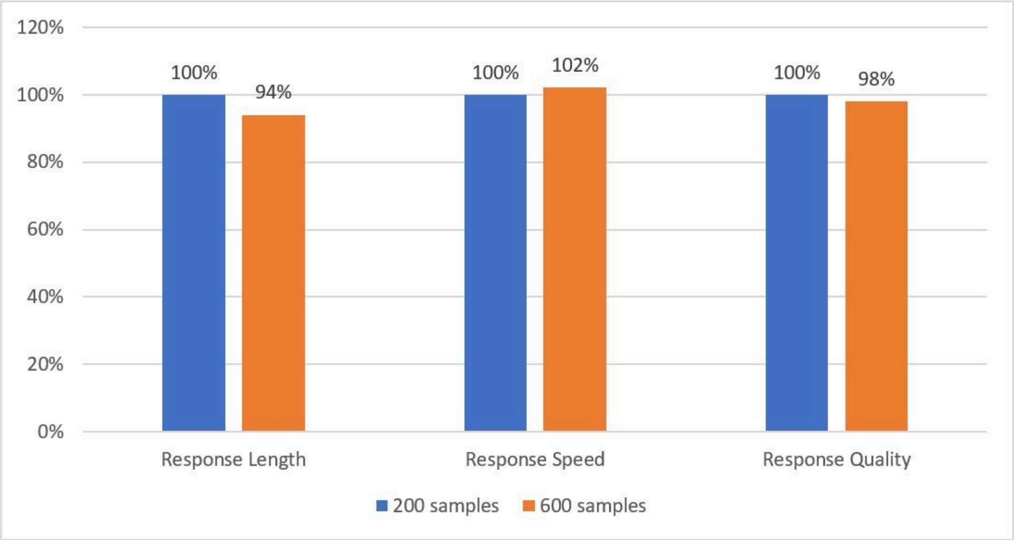


Figure 8. Comparison of problem-solving and final dataset

4 Discussion and Conclusion

The results highlight the challenges and obstacles that come with fine-tuning LLMs on large-context data. Reasoning models such as gpt oss struggle to stay on topic, and open-wright and open-source models such as Mistral-small-3.2 do not perform as well as proprietary models such as GPT-4.1-mini. However, when taking into account the high cost of fine-tuning and testing such models, going with a high-performing open-source model can be advantageous.

The testing also underscored how an increase in dataset size, even when it is not related to the domain of the prompt, can increase response speed and length. However, extensive testing is needed as these improvements were not linear and appeared to vary depending on the type of prompt and the complexity of the reasoning required.

Finally, it is important to note that the dataset used in this study was relatively limited in size. While initial trends point toward promising improvements with larger datasets, more extensive testing across diverse domains and tasks is necessary to draw conclusive insights. While the dataset was manually collected, in the future it would be necessary to utilize a web scraping strategy. Future research could explore scaling experiments with higher amounts of tokens, comparing domain-specific fine-tuning against general-purpose augmentation, and evaluating trade-offs between performance gains and resource expenditure

References

Molenda, M. (2003). *The ADDIE Model*. In *Encyclopedia of Educational Technology*. ABC-CLIO. Merrill, M. D. (2002). A pebble-in-the-pond model for instructional design. *Performance Improvement, 41*(7), 39–44.

Reiser, R. A., & Dempsey, J. V. (Eds.). (2018). *Trends and issues in instructional design and technology* (4th ed.). New York, NY: Pearson Education.

Holmes, W., Porayska-Pomsta, K., Holstein, K., Sutherland, E., Baker, T., Shum, S. B., Santos, O. C., Rodrigo, M. T., Cukurova, M., Bittencourt, I. I., & Koedinger, K. R. (2021). Ethics of AI in education: Towards a community-wide framework. *International Journal of Artificial Intelligence in Education*. https://doi.org/10.1007/s40593-021-00239-1

Li, H., Fang, Y., Zhang, S., Lee, S. M., Wang, Y., Trexler, M., & Botelho, A. F. (2025). *ARCHED: A* *human-centered framework for transparent, responsible, and collaborative AI-assisted instructional design*. arXiv. https://doi.org/10.48550/arXiv.2503.08931

Dickey, E., & Bejarano, A. (2024). *GAIDE: A Framework for Using Generative AI to Assist in Course Content Development*. *2021 IEEE Frontiers in Education Conference (FIE)*, 1–9. https://doi.org/10.1109/fie61694.2024.10893132