LLARS: Enabling Domain Expert & Developer Collaboration for LLM Prompting, Generation and Evaluation

Philipp Steigerwald[1] , **Mara Stieler**[2] , **Jennifer Burghardt**[2] , **Eric Rudolph**[1] , **Jens Albrecht**[1]
Technische Hochschule Nürnberg Georg Simon Ohm, Nürnberg, Germany
[1]Faculty of Computer Science, Centre for Artificial Intelligence (KIZ)
[2]Faculty of Social Sciences, Institute for E-Counselling
{philipp.steigerwald, mara.stieler, jennifer.burghardt, eric.rudolph, jens.albrecht}@th-nuernberg.de

Abstract

We demonstrate LLARS (LLM Assisted Research System), an open-source platform that bridges the gap between domain experts and developers for building LLM-based systems. It integrates three tightly connected modules into an end-to-end pipeline: **Collaborative Prompt Engineering** for real-time co-authoring with version control and instant LLM testing, **Batch Generation** for configurable output production across user-selected prompts × models × data with cost control and **Hybrid Evaluation** where human and LLM evaluators jointly assess outputs through diverse assessment methods, with live agreement metrics and provenance analysis to identify the best model–prompt combination for a given use case. New prompts and models are automatically available for batch generation and completed batches can be turned into evaluation scenarios with a single click. Interviews with six domain experts and three developers in online counselling confirmed that LLARS feels intuitive, saves considerable time by keeping everything in one place and makes interdisciplinary collaboration seamless.

Source code: github.com/th-nuernberg/llars

1 Introduction

Building LLM-based systems for sensitive domains [Chen *et al.*, 2024; Ling *et al.*, 2025]–online counselling [Stade *et al.*, 2024], legal analysis [Katz *et al.*, 2024], medical documentation [Van Veen *et al.*, 2024]–demands both domain expertise and technical skill [Zamfirescu-Pereira *et al.*, 2023; Schulhoff *et al.*, 2024]: practitioners must iterate prompts, select suitable models and rigorously evaluate outputs on their own data, since public benchmarks rarely transfer. Yet domain experts and developers typically work in isolation, with prompts iterated in shared documents without version control, models tested ad hoc and outputs scattered across tools. LLARS (LLM Assisted Research System) is an open-source platform that unifies these three stages (Figure 1).

It integrates three tightly connected modules. *Collaborative Prompt Engineering* lets domain experts and developers co-author prompts in a shared real-time editor with version control and instant LLM testing. *Batch Generation* produces the user-configured Cartesian product of prompts × models × data items with cost estimation and budget control. *Hybrid Evaluation* combines human and LLM evaluators in shared assessment campaigns with automated agreement statistics and provenance tracing to surface the best model–prompt pair. New prompts, models and providers are instantly available across all modules and shareable with collaborators. Completed batches become evaluation scenarios with a single click and each module supports JSON/CSV export.

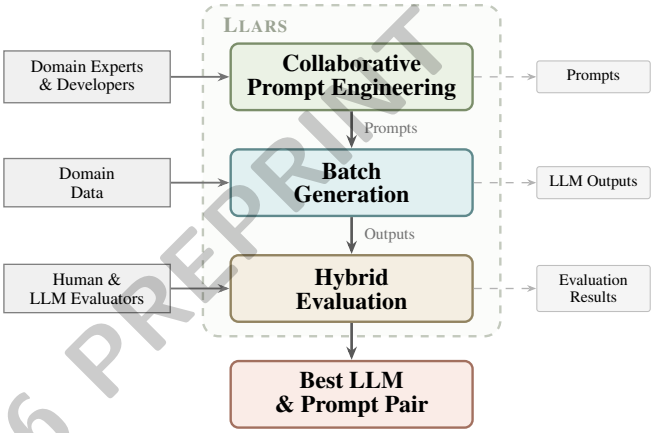

Figure 1: LLARS pipeline: domain experts and developers collaboratively develop prompts, generate outputs across LLMs and run hybrid evaluation with human and LLM evaluators. Each stage supports export and the pipeline yields a validated model–prompt combination.

Table 1 surveys twelve existing platforms along these dimensions. Agenta [Agenta AI, 2024] combines prompt versioning with LLM-as-judge scoring and A/B deployment testing but lacks real-time co-editing and structured human evaluation with inter-rater agreement. Phoenix [Arize AI, 2024], Langfuse [Langfuse, 2024] and W&B Weave [Weights & Biases, 2024] provide observability dashboards with dataset experiments but focus on tracing and debugging production systems rather than systematic batch generation or coordinated multi-evaluator campaigns. ChainForge [Arawjo *et al.*, 2024] offers a visual node-based interface for multi-model

Tool	OSS	Collab.	Batch	H-Eval	LLM-E	Stats	E2E
LLARS	●	●	●	●	●	●	●
Agenta [Agenta AI, 2024]	●	◐	◐	◐	●	○	◐
ChainForge [Arawjo *et al.*, 2024]	●	○	●	○	◐	○	◐
Phoenix [Arize AI, 2024]	●	○	◐	◐	●	○	○
Langfuse [Langfuse, 2024]	●	○	◐	◐	●	◐	○
W&B Weave [Weights & Biases, 2024]	●	○	◐	◐	●	○	○
Label Studio [HumanSignal, Inc., 2024]	●	○	○	●	◐	◐	○
Argilla [Argilla, Inc., 2024]	●	○	○	●	◐	◐	○
LangSmith [LangChain, Inc., 2024]	○	○	◐	◐	●	○	◐
Braintrust [Braintrust Data, Inc., 2024]	○	○	◐	◐	●	○	◐
Maxim [Maxim AI, 2024]	○	○	◐	◐	●	○	◐
Vellum [Vellum AI, 2024]	○	○	◐	◐	●	○	◐

Table 1: Comparison of GUI-first tools for prompt engineering, batch generation and hybrid LLM output evaluation. ● = native support; ◐ = partial; ○ = not supported. **OSS** = open-source software; **Collab.** = simultaneous multi-user prompt editing with live synchronisation; **Batch** = configurable generation over prompts × models × data with cost control; **H-Eval** = structured human evaluation with item distribution and role management; **LLM-E** = any form of automated LLM-based evaluation, including LLM-as-judge and LLM-as-evaluator; **Stats** = live analytics including inter-rater reliability and provenance analysis; **E2E** = integrated end-to-end pipeline.

comparison with template variables but omits structured human evaluation entirely. Label Studio [HumanSignal, Inc., 2024] and Argilla [Argilla, Inc., 2024] offer dedicated human annotation workflows with agreement analytics but require externally generated outputs, disconnecting evaluation from prompt development. LangSmith [LangChain, Inc., 2024], Braintrust [Braintrust Data, Inc., 2024], Maxim [Maxim AI, 2024] and Vellum [Vellum AI, 2024] provide playground-style testing with dataset-level scoring but are closed-source, lack collaborative editing and omit batch generation. LLARS unifies collaborative prompt engineering, batch generation and hybrid human–LLM evaluation with agreement analytics in a single open-source platform.

2 Platform Overview

Co-designed with developers, social science researchers and practising counsellors, LLARS runs as a containerised web application for textual data. The three modules are tightly integrated so that prompts created in the editor flow directly into batch generation and completed batches become evaluation scenarios with a single click, keeping provenance intact.

2.1 Collaborative Prompt Engineering

The collaborative prompt editor (Figure 2) lets every keystroke appear instantly for all connected users. A prompt is composed of ordered blocks–one optionally designated as system prompt, the rest concatenated into the user prompt–each maintaining its own version history that tracks insertions and deletions (visible as `+100 / −0` in Figure 2), enabling diff comparison and rollback without affecting other blocks.

Blocks may contain template variables (`{{variable_name}}`) as placeholders for external data (e.g. an entire email thread captured in a single `{{content}}` variable), keeping the surrounding instructions concise. All variables are collected in the shared Variable Palette (Figure 2, bottom-left) and annotated with sample values. A single click sends the assembled prompt

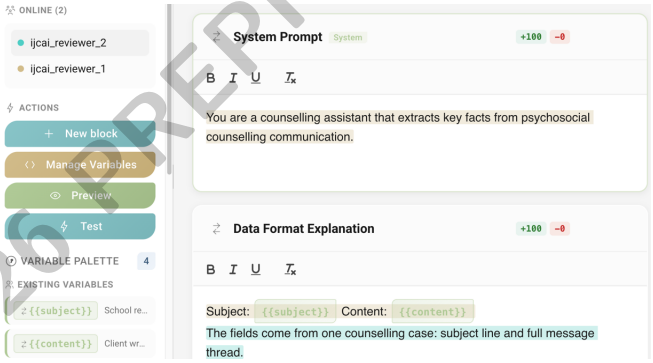

Figure 2: Collaborative prompt editor with ordered blocks and template variables inserted from the Variable Palette (bottom-left).

with sample values substituted to a selected model and streams the response back in real time. Prompts are exportable as JSON and directly available in batch generation, where template variables are filled from the uploaded data to scale a single prompt to hundreds of outputs.

2.2 Batch Generation

Batch generation extends prompt testing beyond single examples by combining any set of prompts, models and data items into their Cartesian product (prompts × models × data items). A generation matrix previews all combinations and estimates cost; optional budget caps pause jobs that exceed a threshold. For example, 50 counselling email threads with two prompts and two models produce $50 \times 2 \times 2 = 200$ outputs. Results stream in real time, each tagged with full provenance (source item, prompt version, model, parameters, tokens and cost) and exportable as CSV or JSON. Completed batches become evaluation scenarios in one click, preserving attribution end-to-end.

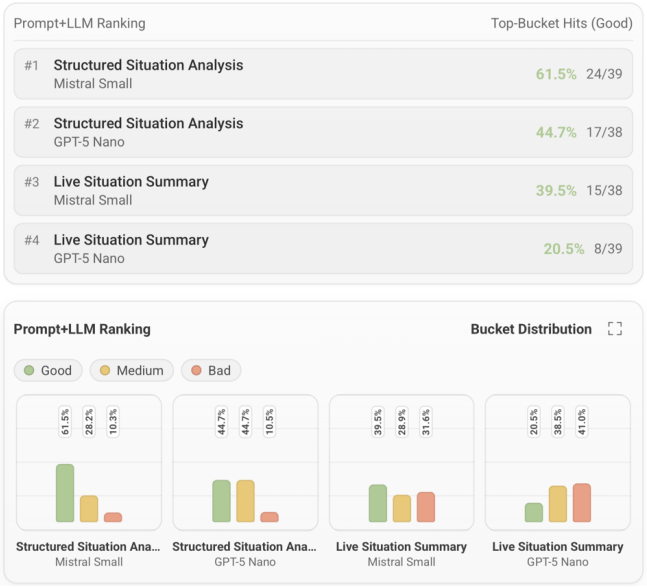

Figure 3: Provenance analysis ranking model–prompt combinations by top-bucket hit rate (top) with per-combination bucket distributions (bottom).

2.3 Hybrid Evaluation

Each evaluation is organised as a *scenario*, a self-contained campaign that defines the evaluation type, assigns evaluators and distributes items. Items are presented in randomised order and without any provenance information, so evaluators cannot tell which model or prompt produced a given output. LLARS supports multi-dimensional rating with configurable Likert scales, ranking into ordinal buckets or traditional ranking, categorical labelling, pairwise comparison, mail assessment and authenticity detection. For comparing multiple outputs per input, *bucket ranking* proved the most popular method in user feedback: items are sorted into ordinal categories and ranked within each bucket [Miller, 1956; Kiritchenko and Mohammad, 2017]. For single outputs, multi-dimensional rating is more appropriate. The *mail rating* preset, for instance, provides Likert scales tailored to assessing an LLM's reply within an email counselling thread. Scenarios are created from batch outputs with a single click or via an AI-assisted Scenario Wizard.

Owners control whether all items go to every evaluator or configurable subsets are distributed for faster throughput. LLARS treats LLM evaluators as full participants alongside humans. They receive the same items, use the same setup and their judgements enter the same agreement analysis. Scenario owners can inspect aggregated results at any time as assessments come in, including automatically computed inter-rater reliability (e.g. Krippendorff's α), filtered by human evaluators only, LLM evaluators only or both combined. Among the available analytics, *provenance analysis* leverages each item's generating model and prompt to display the top-bucket hit rate and full bucket distribution per model–prompt pair (Figure 3), surfacing the best performer.

3 Deployment and User Study

LLARS is actively used in online counselling research: the Virtual Client project [Rudolph *et al.*, 2024] uses it for prompt development of simulated client interactions, and CAIA [Steigerwald *et al.*, 2025], an AI assistant for email counselling, relies on it for prompt development and evaluation. A full end-to-end case study [Steigerwald and Albrecht, 2025] produced 253 subject lines across 11 LLMs and had five professionals and an LLM evaluator rate them (1,518 assessments), determining how small a model could be while still meeting quality requirements.

Semi-structured interviews with six domain experts and three developers confirmed that consolidating the pipeline in one platform saves considerable time and makes interdisciplinary collaboration seamless. Domain experts described the interface as intuitive: "as an evaluator you receive the task and immediately know what to do." They also noted that testing prompts directly against an LLM "developed a feeling for how to formulate the task to get the desired output." Developers valued that prompts, data and outputs remain in one place. Both groups emphasised that working in a shared workspace "finally enabled us to work together" rather than passing documents back and forth between disciplines.

4 Demo and Conclusion

Our demo[1] walks attendees through the complete workflow, from collaborative prompt authoring through batch generation to evaluation with live agreement metrics and provenance analysis, using pre-built counselling scenarios that let attendees author prompts, trigger batch runs and evaluate outputs hands-on.

LLARS unifies prompt engineering, batch generation and evaluation in a single open-source platform that closes the loop from prompt development to rigorous assessment. While currently deployed in online counselling, the platform is domain-agnostic and applicable to any text-based LLM evaluation task. Present limitations include the focus on single-turn generation rather than multi-turn conversations and the dependence of LLM-based evaluators on the underlying model's reasoning capability. Despite these constraints, interviews confirmed that centralising all stages in one workspace saves considerable time, feels intuitive and eliminates the "translation work" between disciplines. Domain experts reported being able to author prompts and conduct evaluations independently without external guidance. Beyond the current single-turn focus, future work will extend LLARS to multi-turn conversational evaluation in which each turn is assessed in context, automated calibration of LLM evaluators against human ratings to surface systematic biases in real time and an HTTP API connecting evaluation outcomes directly to model fine-tuning pipelines, enabling closed-loop iteration from prompt development through assessment to model improvement. Ultimately, LLARS bridges the gap between domain expertise and technical implementation and enables interdisciplinary work.

[1] Video: https://youtu.be/3QaKouwr4gU

References

[Agenta AI, 2024] Agenta AI. Agenta: Open-source LLMOps platform for prompt management, evaluation, and observability. https://agenta.ai, 2024. Accessed: 2026-02-01.

[Arawjo *et al.*, 2024] Ian Arawjo, Chelse Swoopes, Priyan Vaithilingam, Martin Wattenberg, and Elena Glassman. ChainForge: A visual toolkit for prompt engineering and LLM hypothesis testing. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. ACM, 2024.

[Argilla, Inc., 2024] Argilla, Inc. Argilla: Collaboration tool for AI engineers and domain experts. https://argilla.io, 2024. Accessed: 2026-01-06.

[Arize AI, 2024] Arize AI. Phoenix: Open-source AI observability and evaluation. https://phoenix.arize.com, 2024. Accessed: 2026-02-01.

[Braintrust Data, Inc., 2024] Braintrust Data, Inc. Braintrust: AI evaluation and observability platform. https://www.braintrust.dev, 2024. Accessed: 2026-02-01.

[Chen *et al.*, 2024] Zhiyu Zoey Chen, Jing Ma, Xinlu Zhang, Nan Hao, An Yan, Armineh Nourbakhsh, Xianjun Yang, Julian McAuley, Linda Petzold, and William Yang Wang. A survey on large language models for critical societal domains: Finance, healthcare, and law. *Transactions on Machine Learning Research*, 2024.

[HumanSignal, Inc., 2024] HumanSignal, Inc. Label studio: Open source data labeling platform. https://labelstud.io, 2024. Accessed: 2026-01-06.

[Katz *et al.*, 2024] Daniel Martin Katz, Michael James Bommarito, Shang Gao, and Pablo Arredondo. GPT-4 passes the bar exam. *Philosophical Transactions of the Royal Society A*, 382(2270):20230254, 2024.

[Kiritchenko and Mohammad, 2017] Svetlana Kiritchenko and Saif Mohammad. Best-worst scaling more reliable than rating scales: A case study on sentiment intensity annotation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 465–470. Association for Computational Linguistics, 2017.

[LangChain, Inc., 2024] LangChain, Inc. LangSmith: AI agent evaluation platform. https://www.langchain.com/langsmith, 2024. Accessed: 2026-02-01.

[Langfuse, 2024] Langfuse. Langfuse: Open-source LLM engineering platform. https://langfuse.com, 2024. Accessed: 2026-02-01.

[Ling *et al.*, 2025] Chen Ling, Xujiang Zhao, Jiaying Lu, Chengyuan Deng, Can Zheng, Junxiang Wang, Tanmoy Chowdhury, Yun Li, Hejie Cui, Xuchao Zhang, et al. Domain specialization as the key to make large language models disruptive: A comprehensive survey. *ACM Computing Surveys*, 58(3), 2025.

[Maxim AI, 2024] Maxim AI. Maxim: GenAI evaluation and observability platform. https://www.getmaxim.ai, 2024. Accessed: 2026-02-01.

[Miller, 1956] George A. Miller. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, 63(2):81–97, 1956.

[Rudolph *et al.*, 2024] Eric Rudolph, Natalie Engert, and Jens Albrecht. An AI-based virtual client for educational role-playing in the training of online counselors. In *Proceedings of the 16th International Conference on Computer Supported Education (CSEDU)*, pages 108–117, 2024.

[Schulhoff *et al.*, 2024] Sander Schulhoff, Michael Ilie, Nishant Balepur, Konstantine Kahadze, Amanda Liu, Chenglei Si, Yinheng Li, Aayush Gupta, HyoJung Han, Sevien Schulhoff, et al. The prompt report: A systematic survey of prompting techniques. *arXiv preprint arXiv:2406.06608*, 2024.

[Stade *et al.*, 2024] Elizabeth C. Stade, Shannon Wiltsey Stirman, Lyle H. Ungar, Cody L. Boland, H. Andrew Schwartz, David B. Yaden, Joao Sedoc, Robert J. DeRubeis, Robb Willer, and Johannes C. Eichstaedt. Large language models could change the future of behavioral healthcare: A proposal for responsible development and evaluation. *npj Mental Health Research*, 3:12, 2024.

[Steigerwald and Albrecht, 2025] Philipp Steigerwald and Jens Albrecht. Comparing large language models for automated subject line generation in e-mental health: A performance study. In *Proceedings of the 11th International Conference on Information and Communication Technologies for Ageing Well and e-Health (ICT4AWE)*, pages 70–77. SciTePress, 2025.

[Steigerwald *et al.*, 2025] Philipp Steigerwald, Nico Bienlein, Jennifer Burghardt, Mara Stieler, Robert Lehmann, and Jens Albrecht. CAIA in practice: Field evaluation of an AI-assisted support system for text-based online counselling. In *Proceedings of the IEEE International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 1476–1483. IEEE, 2025.

[Van Veen *et al.*, 2024] Dave Van Veen, Cara Van Uden, Louis Blankemeier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, Eduardo Pontes Reis, Anna Seehofnerova, et al. Adapted large language models can outperform medical experts in clinical text summarization. *Nature Medicine*, 30:1134–1142, 2024.

[Vellum AI, 2024] Vellum AI. Vellum: AI application development platform. https://www.vellum.ai, 2024. Accessed: 2026-02-01.

[Weights & Biases, 2024] Weights & Biases. Weave: Toolkit for developing AI-powered applications. https://wandb.ai/site/weave, 2024. Accessed: 2026-02-01.

[Zamfirescu-Pereira *et al.*, 2023] J.D. Zamfirescu-Pereira, Richmond Y. Wong, Bjoern Hartmann, and Qian Yang. Why johnny can't prompt: How non-AI experts try (and fail) to design LLM prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, 2023.