Agentic Pipelines Meet Retrieval-Augmented ICL: A Zero-Training Approach to Mental Health Modeling

Anson Antony[1,2], **Gautam Vijay Kumar**[3], **Annika M. Schoene**[2]

[1]Maia Medical Billing Corp, [2]Northeastern University, [3]Florida International University

Correspondence: ansonanto53@gmail.com

Abstract

We describe the Meronym Labs system for the CLPsych 2026 shared task. Our approach uses retrieval-augmented in-context learning with frozen LLMs throughout, requiring no fine-tuning. Tasks 1 and 2 use a single Qwen 3.5 27B call per post/timeline with static and dynamically retrieved examples. Task 3.1 uses a five-agent agentic pipeline including rule-based agents for change type and direction, LLM agents for dynamics extraction and summary writing, and a validator all augmented with NLI-based candidate re-ranking and iterative contradiction reduction. The system ranked 1st on Task 1.2 (RMSE 0.917) and Task 3.1 (score rank average 4.00), 3rd on Task 1.1 (F1 0.420), and 8th on Task 2 (F1 0.466).

1 Introduction

The CLPsych 2026 shared task asks participants to model dynamic mental health changes through social media timelines, building on the MIND framework (Atzil-Slonim, 2025, 2026) that conceptualises self-states as combinations of Affect, Behaviour, Cognition, and Desire (ABCD) components. The task comprises four subtasks: ABCD subelement classification (Task 1.1), self-state presence rating (Task 1.2), moments-of-change detection (Task 2), and sequence summarisation (Task 3.1). We participated in all four subtasks with a unified architecture and no*no fine-tuning*. Every system is a frozen LLM conditioned at inference time through in-context learning (ICL) and retrieval-augmented generation (RAG). This choice was motivated by three considerations. First, the training set was small (236 annotated posts, 30 timelines, 74 change sequences), making fine-tuning of billion-parameter models impractical. Second, the ABCD schema is complex (32 subelement codes across two valences and six element types), but can be fully specified in a prompt. Finally, post text is clinically sensitive, and ICL allows all inference to run locally without transmitting data to external services.

2 Related Work

Mental health NLP has a rich history of analysing social media to detect psychiatric conditions (De Choudhury et al., 2013; Coppersmith et al., 2014; Chancellor and De Choudhury, 2020), with CLPsych shared tasks progressively addressing suicidality risk (Zirikly et al., 2019), longitudinal mood change (Tsakalidis et al., 2022), and ABCD-grounded self-state dynamics (Tseriotou et al., 2025). Top-performing systems at recent editions have combined ICL with retrieval-based example selection (Uluslu et al., 2024; Antony and Schoene, 2025), a direction well-supported by broader work on RAG (Lewis et al., 2020) and ICL (Brown et al., 2020; Min et al., 2022) in clinical NLP settings (Agrawal et al., 2022). Our automated prompt-tuning loop builds on LLM-as-judge evaluation (Zheng et al., 2023) and iterative prompt optimisation (Pryzant et al., 2023; Yang et al., 2023).

3 Infrastructure

All systems run on three NVIDIA A40 GPUs (46 GB each) via a local Ollama server. Tasks 1 and 2 use *Qwen 3.5 27B* (Qwen Team, 2025), a dense 27-billion-parameter model, with thinking mode disabled (temperature 0.1, max 4,096 output tokens). Task 3.1 uses *Gemma 4 31B* (Gemma Team, 2025) with thinking mode disabled (temperature 0.2, max 900 tokens). We disabled thinking mode, because the Ollama chat API returned an empty content field and routed all output to a separate thinking field inaccessible to the pipeline. Semantic retrieval uses *Nomic Embed Text* (Nussbaum et al., 2024), a 768-dimensional embedder on the same Ollama server, over a persistent *ChromaDB* index with HNSW indexing (Malkov and

Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2026), pages 435–440

Yashunin, 2018) and cosine similarity. All inference runs locally; post text and gold labels are never transmitted to an external service. The only exception is the Task 2 prompt-tuning loop (Section 4.2), in which only aggregate error statistics (scalar F1, precision, and recall counts) and the current prompt text are sent externally. No shared-task data ever leaves local hardware. This includes post content, individual labels, and any record from which a post or annotation could be reconstructed. The external API was used purely as a prompt-rewriting tool conditioned on numeric summaries, so this usage is consistent with the CLPsych 2026 Data Access Form, which restricts the transmission, storage, and processing of the dataset itself by external services.

4 System Description

We refer the reader to Ali et al. (2026) for full definitions and evaluation metrics for each task.

4.1 Tasks 1.1 and 1.2

The schema treats subelement classification (Task 1.1) and presence rating (Task 1.2) as facets of the same prediction, so both outputs are produced in a single LLM call per post. Figure 1 shows the pipeline.

*Prompt design:*The six-part prompt contains: (1) a system message embedding the MIND framework and classification rules; (2) the full 32-code ABCD sub-element table from the task guidelines; (3) the exact 1–5 presence scale definitions; (4) *three static gold examples* chosen once at startup, covering a maximally adaptive post, a maximally maladaptive post, and a clearly mixed post; (5) *two dynamic RAG examples* retrieved per post by cosine similarity from ChromaDB; and (6) the target post with a JSON output schema. The static examples teach the model the *range* of correct outputs; the RAG examples teach it what *similar* posts have been labelled by human annotators. This combination keeps prompt length bounded while providing both distributional coverage and instance-specific guidance.

Post-validation: A deterministic layer strips any `<think>` blocks from the response, extracts the JSON output (fenced block or first balanced-brace pair), validates every sub-element code against the expected schema (e.g. adaptive Affect codes must belong to $\{1, 3, 5, 7, 9, 11, 13\}$), clamps Presence to $[1, 5]$, and drops invalid entries.

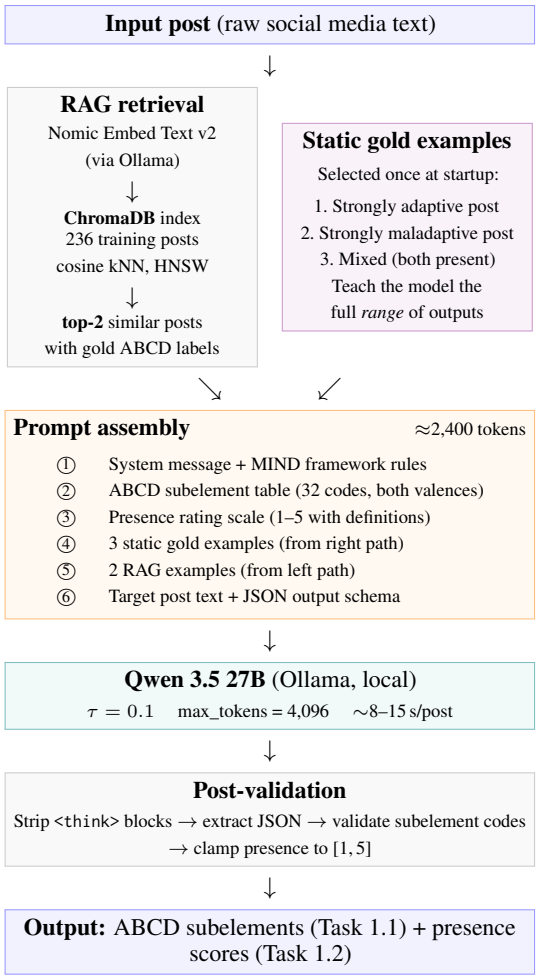

Figure 1: Overview of the Task 1 pipeline.

4.2 Task 2: Moments of Change

Our Task 2 pipeline has two stage. Stage 1 estimates per-post well-being, where gold scores are used directly in training mode; otherwise Qwen 3.5 27B assigns a GAF-rescaled 1–10 score from an anchored prompt designed to counteract the model's bias toward outputting 7. Stage 2 passes all post texts and well-being scores in a single timeline-level LLM call, returning per-post JSON labels with definitions for both change types, eleven linguistic escalation triggers, and instructions to scan every 3-, 4-, and 5-post window. The Stage 2 prompt was produced by an automated loop: eight iterations on a 50-post development sample, where aggregate F1 statistics (not post text or labels) were sent to Claude Sonnet 4 via API to rewrite the prompt, accepting only improvements. Escalation recall rose from 0.17 (rule-based) to 0.65 (LLM-first) to 0.92 (auto-tuned); Table 1 shows the full progression.

Variant	Sw F1	Esc F1	Comb.	Esc R
v1: rule + LLM verify	0.507	0.241	0.394	0.17
v2: LLM-first, hybrid switch	0.593	0.583	0.512	0.65
v2 + auto-tune (i7)	**0.743**	**0.654**	**0.699**	**0.92**
Submitted (full test set)	0.423	0.597	0.466	0.71

Table 1: Task 2 ablation. Rows 1–3: 50-post development sample. Row 4: official test-set submission.

4.3 Task 3.1: Sequence Summarisation

We made three Codabench submissions for Task 3.1, each building on the previous, with Gemma 4 31B as the generator throughout.

Submission 1 (CS 0.7464) Five specialist modules operate in sequence and two rule-based agents deterministically derive change type (Switch vs. Escalation) and direction (improvement vs. deterioration), offloading structural reasoning from the LLM. A *DynamicsExtractor* call returns structured JSON covering the central theme, temporal phases, and ABCD dynamics. A *SummaryWriter* composes a $\leq$350-word summary conditioned on these outputs and three ChromaDB-retrieved gold exemplars, whiel a rule-based *Validator* checks word count, ABCD tag coverage ($\geq$0.85), and explicit naming of change identity and direction, routing failures to a targeted reviser. Our analysis of the training-to-test gap identified contradicting sentences as the dominant failure mode, motivating Submission 2.

Submission 2 (CS 0.7874) We added two components, where the four candidate summaries generated at temperatures $\{0.1, 0.3, 0.5, 0.7\}$ are reranked by DeBERTa-v3-large (He et al., 2023) NLI scores against gold exemplars. A *CT-Reducer* then rewrites sentences whose contradiction score exceeds 0.4 using targeted strategies (verb-swap, hedging, universal-removal), reducing CT by 0.13 relative to Submission 1.

Submission 3 (CS 0.8007, best) Here, our analysis revealed two failure modes, where 12% of sentences exceeded contradiction 0.5 but escaped the 0.4 threshold, and five of the ten worst sentences were formulaic openers scored as contradicting gold summaries. We addressed this in two additional passes address by re-writing regex-detected openers that are anchored in post evidence, and the CT-Reducer threshold is lowered from 0.4 to 0.25.

4.4 Task 3.2: Recurrent Dynamic Signatures

Our approach uses a map-reduce pipeline over the 74 training gold summaries with Gemma 4 31B. The summaries are partitioned into improvement and deterioration groups via keyword detection with an LLM tiebreak, chunked into groups of eight, and recurring ABCD patterns extracted per chunk (*map* phase). Chunk-level patterns are then synthesised into one signature per direction (*reduce* phase), constrained to $\leq$90 words. The submitted version adds multi-candidate generation ($N = 5$), a hedging pass mirroring the CT-Reducer from Task 3.1, and LLM-as-judge reranking.

5 Results

Tables 1, 2, and 3 present the full results. Table 1 shows the Task 2 ablation, Table 2 reports official scores for Tasks 1 and 2, and Table 3 shows the Task 3.1 progression across submissions. The gap between adaptive (0.32) and maladaptive (0.52) subelement F1 on Task 1.1 is consistent with prior findings on sentiment asymmetry in mental health posts. For Task 2, the post-level macro F1 (0.510) exceeds the timeline-level F1 (0.421) by a larger margin than most competing teams, suggesting the single-timeline-prompt design loses structural coherence when predictions are aggregated. The largest single gain on Task 3.1 came from NLI reranking and CT-Reducer combined (+0.041 CS, $-$0.128 CT, Sub 1 to Sub 2).

6 Discussion

The Task 3.1 rank 1 result is notable because our system did not lead on any single metric, but achieved the best score rank average by being consistently strong across all four. Teams that led on individual metrics (e.g. Aurevia, 1st on CS but 11th on ROUGE-L and BERTScore) fell off on others, suggesting that optimising for a single metric is insufficient for this task. For Task 2, the largest performance jump came from replacing the rule-based escalation detector with a linguistically grounded prompt, as the signal is present in word choice and

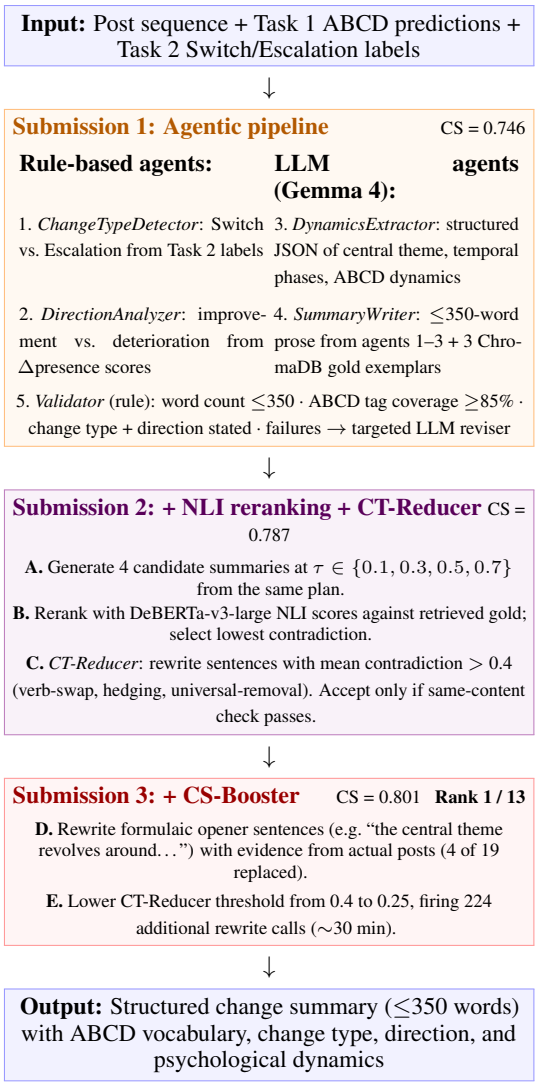

Figure 2: Overview of Task 3.1 pipeline.

tone rather than numeric well-being thresholds. For Task 3.1, each submission improved by targeting a specific failure mode from prior outputs, with each CS and CT gain accompanied by a small ROUGE-L regression as rewriting moved text away from gold summary phrasing. Post-deadline scaling experiments with DeepSeek-R1 70B suggest that generator size alone does not drive summary quality; gains came instead from post-hoc NLI reranking and iterative contradiction reduction (Appendix A).

7 Conclusion

We presented a retrieval-augmented in-context learning system for the CLPsych 2026 shared task that achieves competitive results across all four submitted subtasks without any fine-tuning. The system ranked 1st on Task 1.2 (presence rating) and Task 3.1 (sequence summarisation), 3rd on Task 1.1

Task 1.1: Subelement Classification (Rank 3 / 17)	
Avg. Subelement Macro F1 (ranking)	0.4197
Adaptive Subelement F1	0.3228
Maladaptive Subelement F1	0.5166
Avg. Element Presence Macro F1	0.6383
Task 1.2: Presence Rating (Rank 1 / 17)	
Avg. RMSE (ranking)	0.9167
Adaptive RMSE	1.0000
Maladaptive RMSE	0.8333
Quadratic Weighted κ	0.6774
Spearman ρ	0.6973
MAE	0.6667
Task 2: Moments of Change (Rank 8 / 18)	
Combined Post+Timeline Macro F1 (ranking)	0.4655
Post-level Macro F1	0.5098
Switch F1	0.4231
Escalation F1	0.5965
Timeline-level Macro F1	0.4212
Baselines (combined macro F1)	
TempoFormer	0.5721
Llama-3.1-8B + Task 1.1 input	0.3648
Llama-3.1-8B zero-shot	0.2722

Table 2: Tasks 1 and 2 official results with valence breakdown and baseline comparisons.

(subelement classification), and 8th on Task 2 (moments of change). Key design choices included combining static diversity examples with dynamic retrieval for structured prediction, automated LLM-driven prompt optimisation for change detection, and iterative NLI-guided contradiction reduction for summarisation. These results demonstrate that frozen LLMs with well-designed prompts and retrieval can match or exceed fine-tuned approaches in low-resource clinical NLP settings.

Limitations

The Task 2 auto-tuned prompt produced a score of 0.466 on the official submission but 0.427 on a later resubmission of the same predictions, indicating instability the leaderboard position alone does not reflect. The adaptive presence RMSE on Task 1.2 was 1.00 compared to 0.83 for maladaptive, suggesting the prompt examples do not adequately cover the range of adaptive state presentations. Task 3.2 yielded weak results, with the map-reduce approach failing to capture cross-sequence patterns for improvement trajectories in particular.

Ethics Statement

This work analyses social media posts from individuals experiencing mental health difficul-

System	CS↑	CT↓	RL↑
Development set (74 sequences)			
Single-call RAG	0.763	0.744	**0.374**
Agentic v1	0.772	0.726	0.297
Official test set (19 sequences)			
Sub 1: Agentic	0.746	0.801	0.293
Sub 2: +NLI +CT-Red	0.787	0.673	0.282
Sub 3: +CS-Boost	**0.801**	**0.659**	0.266
Official ranking (Rank 1 / 13, Score Rank Avg = 4.00)			
CS (consistency)	0.801	Rank 3 / 13	
CT (contradiction)	0.659	Rank 3 / 13	
ROUGE-L recall	0.266	Rank 6 / 13	
BERTScore recall	0.345	Rank 4 / 13	

Table 3: Task 3.1 ablation across submissions and official per-metric ranking.

ties. All data was provided through the official CLPsych 2026 shared task under institutional data use agreements. All inference ran on local hardware; post text and gold labels were never transmitted to external services. The sole exception was the Task 2 prompt-tuning loop, where only aggregate error statistics and prompt text (not post content, individual labels, or any reconstructable record) were sent to an external API for prompt rewriting. We confirm that this usage complies with the CLPsych 2026 Data Access Form: the shared-task dataset was never transmitted to, stored by, or processed by any external service, and the API received only scalar performance summaries that disclose nothing about any individual post or annotation. We recognise that automated mental health assessment systems can carry multiple risks, including but note limited to misclassification and should not be used as standalone diagnostic tools. The systems described here are research prototypes evaluated in a shared task setting. To support reproducibility, the full prompts used across all subtasks are available from the corresponding author upon reasonable request. These prompts contain no shared-task data.

References

Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David Sontag. 2022. Large language models are few-shot clinical information extractors. In *Proceedings of the 2022 conference on empirical methods in natural language processing*, pages 1998–2022.

Iqra Ali, Talia Tseriotou, Guy Dvir, Callum Chan, Yuxiang Zhou, Juan Antonio Lossio-Ventura, Ayal Klein, Aya Shamir, Dan Sayda, Anthony Hills, Aya Zirikly, Diana Inkpen, Dana Atzil-Slonim, and Maria Liakata. 2026. Overview of the clpsych 2026 shared task: Capturing and characterizing mental health changes through social media timeline dynamics. In *Proceedings of the 11th Workshop on Computational Linguistics and Clinical Psychology*. Association for Computational Linguistics.

Anson Antony and Annika M. Schoene. 2025. Retrieval-enhanced mental health assessment: Capturing self-state dynamics from social media using in-context learning. In *Proceedings of CLPsych 2025*, pages 268–278.

Dana Atzil-Slonim. 2025. Multimodal intrapersonal and interpersonal dynamics (MIND): A transtheoretical coding manual.

Dana Atzil-Slonim. 2026. Leveraging theoretical and technological innovations to study the mechanisms that underlie therapeutic change in psychotherapy. In Louis G. Castonguay, Dana Atzil-Slonim, Michael Barkham, and Wolfgang Lutz, editors, *Practice-Based Evidence in the Psychological Therapies: Toward Policy Implications for Research, Training, and Clinical Guidelines*. Oxford University Press, New York.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, and 1 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.

Stevie Chancellor and Munmun De Choudhury. 2020. Methods in predictive techniques for mental health status on social media: A critical review. *npj Digital Medicine*, 3(1):43.

Glen Coppersmith, Mark Dredze, and Craig Harman. 2014. Quantifying mental health signals in twitter. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology*, pages 51–60.

Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting depression via social media. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 7, pages 128–137.

Gemma Team. 2025. Gemma 3 technical report. *arXiv preprint*.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *International Conference on Learning Representations*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474.

Yury A. Malkov and Dmitry A. Yashunin. 2018. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(4):824–836.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064.

Zach Nussbaum, John X. Morris, Brandon Duderstadt, and Andriy Mulyar. 2024. Nomic embed: Training a reproducible long context text embedder. *arXiv preprint arXiv:2402.01613*.

Reid Pryzant, Dan Iter, Jerry Li, Yin Lee, Chenguang Zhu, and Michael Zeng. 2023. Automatic prompt optimization with "gradient descent" and beam search. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 7957–7968.

Qwen Team. 2025. Qwen3 technical report. *arXiv preprint*.

Adam Tsakalidis, Jenny Chim, Iman Munire Bilal, Ayah Zirikly, Dana Atzil-Slonim, Federico Nanni, Philip Resnik, Manas Gaur, Kaushik Roy, Becky Inkster, Jeff Leintz, and Maria Liakata. 2022. Overview of the CLPsych 2022 shared task: Capturing moments of change in longitudinal user posts. In *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology*, pages 184–198, Seattle, USA. Association for Computational Linguistics.

Talia Tseriotou, Jenny Chim, Ayal Klein, Aya Shamir, Guy Dvir, Iqra Ali, Cian Kennedy, Guneet Singh Kohli, Anthony Hills, Ayah Zirikly, Dana Atzil-Slonim, and Maria Liakata. 2025. Overview of the CLPsych 2025 shared task: Capturing mental health dynamics from social media timelines. In *Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2025)*, pages 193–217, Albuquerque, New Mexico. Association for Computational Linguistics.

Ahmet Yavuz Uluslu, Andrianos Michail, and Simon Clematide. 2024. Utilizing large language models to identify evidence of suicidality risk through analysis of emotionally charged posts. In *Proceedings of CLPsych 2024*, pages 264–269.

Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V Le, Denny Zhou, and Xinyun Chen. 2023. Large language models as optimizers. In *The Twelfth International Conference on Learning Representations*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, and 1 others. 2023. Judging LLM-as-a-Judge with MT-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36.

Ayah Zirikly, Philip Resnik, Özlem Uzuner, and Kristy Hollingshead. 2019. CLPsych 2019 shared task: Predicting the degree of suicide risk in Reddit posts. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 24–33.

A DeepSeek-R1 70B Model Comparison

Table 4 compares the Task 3.1 agentic pipeline (Submission 1 architecture) using Gemma 4 31B (submitted) vs. DeepSeek-R1 70B (post-deadline experiment) on the full 74-sequence training set. Metrics are from a local proxy evaluator; official CS/CT/BERTScore were not computed for the DeepSeek run.

Metric	Gemma 4 31B	DeepSeek 70B
Summary quality (proxy)		
ROUGE-L	**0.285**	0.277
Unigram F1	**0.523**	0.513
Mean word count	262.4	238.4
Structural compliance		
ABCD tag coverage	1.000	1.000
Identity correct	100%	100%
Direction stated	98.6%	**100%**
Over 350 words	0%	0%
Pipeline reliability		
Tag repair needed	—	27 / 74
JSON parse failures	—	4 / 74
Infrastructure		
Hardware	3× A40	2× L40S
Runtime (74 seq.)	∼**38 min**	∼5.5 hours
CPU offload	No	24%

Table 4: Gemma 4 31B vs. DeepSeek-R1 70B on the Task 3.1 agentic pipeline (training set, 74 sequences). The 70B model does not improve proxy metrics despite 8× runtime.