Response to Reviewers

Manuscript: *Metatuning: An Empirical Study of Judge-Guided Prompt Refinement and Its Boundary Co*

January 4, 2026

Overview of Major Changes

We thank the editors and reviewers for their constructive feedback. In this revision, we aligned the paper's narrative with the empirical evidence and made targeted clarifications requested by both reviewers. The key updates are:

- **Title update:** The title now explicitly reflects the empirical focus on judge-guided prompt refinement and boundary conditions.
- **Narrative pivot to boundary conditions:** The Abstract and Conclusion now frame the paper as an empirical study of a prompt-refinement technique and its limits, rather than a broad claim about a new neurosymbolic paradigm.
- **Reframed Sections 1–2:** Sections 1–2 now emphasize an operational framing of symbolic feedback and judge-guided prompt refinement, and avoid unsupported claims about internal grounding/alignment without probing evidence.
- **Stronger analysis of negative results:** Section 5.2 now analyzes why metatuning can be non-additive (or harmful) when combined with CoT/self-reflection, including a *context saturation* hypothesis.
- **Judge vs. reasoning trace:** We clarify that the judge currently corrects final answers, not intermediate CoT traces, and we identify trace-level critique as a key future direction.
- **Metatuning vs. in-context learning (ICL):** Section 5.3 now explicitly distinguishes metatuning from standard ICL by emphasizing error-driven example selection (learned prompt artifact) rather than static/retrieved demonstrations.
- **Related work expansion:** We added brief discussion and citations situating metatuning within automated prompt optimization/meta-prompting literature.
- **Editorial fixes:** We filled in the paper organization paragraph and corrected the typos/formatting issues highlighted by the reviewers.

Response to Reviewer 1 (Major Revision)

1. **“Section 5.2 results are surprising; please analyze more (including model differences).”**

   Response: We agree that the cross-model behavior is an important empirical signal. We expanded the analysis in Section 5.2 to explicitly discuss non-additivity with CoT/self-reflection and to motivate why model architecture/attention behavior may lead to different outcomes.

   Action taken in manuscript: Added an “Analysis” paragraph in Section 5.2 discussing *context saturation* and model-dependent interactions between injected metatuning demonstrations and long CoT traces.

2. **“A third model comparison would strengthen the metatuning + CoT comparison.”**

   Response: We agree that adding a third model would strengthen the empirical story. However, running an additional large-scale model was outside the scope of this revision timeline. We therefore (i) strengthened the analysis of the observed boundary condition, and (ii) explicitly flag multi-model replication as future work.

   Action taken in manuscript: We treat the observed behavior as a boundary condition and explicitly motivate multi-model replication as future work (see Section 5.2 analysis and the Conclusion’s “future work” directions).

3. **“Judge does not use CoT/self-reflection trace; missed opportunity.”**

   Response: We agree. In our current implementation, the judge corrects the final answer rather than the intermediate reasoning trace. This can make CoT and metatuning redundant, since the feedback is not targeted at the reasoning process that CoT exposes.

   Action taken in manuscript: We added an explicit note in Section 5.2 and reinforced in the Conclusion that trace-level critique/correction is a key direction for next-generation symbolic feedback systems.

4. **“CLEVRER metatuning looks indistinguishable from ICL; please clarify the distinction.”**

   Response: We clarified that while the inference-time mechanism resembles ICL, metatuning differs in how examples are constructed. In standard ICL, examples are typically static or retrieved by similarity. In metatuning, examples are derived from an error-driven training sweep: the prompt demonstrations are selected/constructed because the candidate model previously failed on them and the judge provided corrected solutions/feedback.

   Action taken in manuscript: Added an explicit “Metatuning vs. ICL” paragraph in the CLEVRER section describing error-driven example selection as a learned prompt artifact (Section 5.3).

5. **“Formatting / English issues: missing organization paragraph; typos (“difficiulity”, “step1”); missing period.”**

   Response: We fixed all listed issues and performed an additional pass for similar errors.

   Action taken in manuscript: Filled in the paper organization paragraph in the Introduction and corrected the listed typos and punctuation issues.

Response to Reviewer 2 (Minor Revision)

1. **“Reframe as boundary conditions / limits; avoid “new neurosymbolic paradigm” framing.”**

   Response: We agree and adopted this as the central revision theme. The paper is now positioned as an empirical study of iterative prompt refinement (metatuning) and its limits.

   Action taken in manuscript: Rewrote the Abstract and Conclusion to foreground the boundary conditions and negative results as key contributions; reduced speculative framing throughout Sections 1–2.

2. **“Model-grounded symbolic framing is philosophical; no internal probing/grounding evidence.”**

   Response: We agree that the prior framing overreached given our evidence. We now explicitly use “model grounding” in a limited, operational sense (representations mediating prompt steering) and avoid implying empirical claims about representation alignment.

   Action taken in manuscript: Revised Sections 1–2 to soften claims and clarify the intended scope; we no longer present internal “vector alignment”/grounding as an empirically supported finding. We also identify representation probing as future work.

3. **“Method remains heuristic; no formal convergence; soften/remove unproven claims (e.g., “cycle ensures...” and data-efficiency).”**

   Response: We agree. We revised the text to avoid guarantee language and to make the lack of formal convergence explicit. We also reframed data efficiency claims as potential *sample efficiency within the context window* for targeted error correction, rather than global training efficiency.

   Action taken in manuscript: Rephrased “ensures” to “aims” and added an explicit sentence noting no formal convergence guarantees; softened data-efficiency language accordingly.

4. **“Strengthen literature comparison: in-context learning, meta-prompt optimization, automated prompt tuning.”**

   Response: We agree and added discussion and citations connecting metatuning to prompt optimization and iterative self-improvement approaches.

   Action taken in manuscript: Added brief related-work positioning and citations in the Introduction and Section 4 (including work on automatic prompt engineering, LLMs-as-optimizers, and iterative refinement with self-feedback).

5. **“Align introduction with empirical results; negative findings should be framed as valuable.”**

   Response: We agree. We now explicitly present the negative results as evidence of boundary conditions (non-additivity with CoT and non-generalization to CLEVRER), and we aligned the paper’s narrative to prepare the reader for those outcomes.

   Action taken in manuscript: Updated framing across Abstract, Introduction, Section 5 analysis, and Conclusion to emphasize scientific value of identifying limits rather than positioning the method as universally effective.

Remaining Scope / Future Work (Explicitly Acknowledged)

- **Additional model replication:** Evaluate metatuning + CoT across a broader set of candidate models to validate model-dependent effects.
- **Trace-level symbolic feedback:** Extend the judge to critique and correct intermediate reasoning traces (CoT), not only final answers.
- **Representation analysis:** Add mechanistic/representation probing to test (or refute) stronger claims about grounding/alignment.
- **Cost/efficiency evaluation:** Quantify compute and annotation/judge-cost tradeoffs compared to alternative prompt-optimization baselines.