

Meta-Review Fairness Analysis

<https://gemini.google.com/app/1fad43e8f5bc881f>

User prompt: In the following feedback for Computational Linguistics, read the two reviews, and then look at the metareview. Question: does the metareview sound like a fair characterization of what the two reviewers wrote? Meta-review: While both reviewers acknowledge the merit of the work, they raise serious and complementary concerns about this manuscript. Reviewer 1 raises the limitations of the current experimental setup in terms of the number of models used (currently one), and emphasizes that the findings cannot be fully trusted unless replicated using model variations with different sizes and architectures. Also all tested number systems use a base of 10, which again limits the scope of the work. Reviewer 2 on the other hand raises fundamental questions about the formulation of the study and its relevance to natural languages in general, and questions the impact of the findings and the insight they generate. -- We appreciate your submitting your work to Computational Linguistics and wish you the best in your future research. Sincerely, Professor Wei Lu Nanyang Technological University, Singapore Editor-in-Chief, Computational Linguistics editor@cljournal.org on behalf of Afra Alishahi Action Editor Computational Linguistics Reviewer 1: Recommendation: Revisions Required 1. Bottom-line Decision My recommendation for this paper is: Revise and resubmit for review: Needs more than minor changes If you selected "Revise and resubmit for review" please indicate what kind of revisions are required: - Further Work Required 2. What is this paper about? This is a follow-up to a 2022 paper from one of the authors where they argued that some problems such as text normalization and specifically number reading can't be solved with purely neural models as they suffer from catastrophic errors. The new paper revisits that question and tests two Claude models on a set of artificial number systems of different complexities. They find that Sonnet performs poorly, but Opus was able to learn the number systems near-perfectly from few-shot prompts. 3. Strengths: The difference between Sonnet and Opus is striking. The paper is quite timely given the recent release of Opus 4.8. Designing artificial number reading systems is a neat idea to avoid train/eval set contamination. Critically revisiting their own work after 4 years is something that doesn't happen often, but it should be encouraged as it sets both papers in a long-term context that is often overlooked. 4. Weaknesses: I get that this paper merely addresses the "feasibility" question, so finding a single successful model (Opus 4.8) is technically enough to answer it. Still, extending the experimental setup to more model versions, model families, and different sizes would make the results less anecdotal and more useful — is Opus 4.8 just a lucky shot, or are all models bigger or newer than Opus 4.8 fine? All tested number systems use a base of 10. Have you considered a) other naturally occurring bases, or b) bases that are not used in any human language to understand the nature of Opus' generalization capabilities when more than 10 single digit words are used? 5. Substantive Revisions Complete this section if either "Revise and Resubmit" or "Reject" has been recommended. Substantive Revisions to be Required: Test a broader set of models Substantive Revisions to be Encouraged: Test number systems with different bases 6. Minor Revisions Minor Revisions to be Required: Minor Revisions to be Encouraged: 7. Typographic Errors 8. Revision time How much time do you think the authors might need to make the required changes you have requested? 1 month Some other time (please specify): Reviewer 2: Recommendation: Revisions Required 1. Bottom-line Decision My recommendation for this paper is: Accept with minor revisions: Only minor points to be addressed 2. What is this paper about? I read this piece with interest, as it is not merely a scientific article but a reflection on the capabilities of language models to solve problems that, *a priori*, seem uninteresting to the vast majority of the NLP community. In particular, the authors previously argued—and to some extent demonstrated—that language models could not solve certain text normalization problems, specifically the issue of unrecoverable errors arising from verbalizations that are fluent yet semantically incorrect. 3. Strengths: In this new piece, utilizing more advanced models, the authors present several instances of the problem of naming (reading) numerals (in Arabic notation) in an artificial language; as I understand it, these languages possess an associated grammar that could potentially be learned. Here, the authors report the success of certain new models on the proposed languages. 4. Weaknesses: A question that immediately comes to

mind is whether the authors conducted experiments aimed at generating a grammar for the proposed languages (using traditional techniques) and exposing the rule-based model solely to a set of examples. A question that immediately comes to mind is whether the authors conducted experiments aimed at generating a grammar for the proposed languages (using traditional techniques) and exposing the rule-based model solely to a set of examples. The second question concerns the universe of discourse: what evidence demonstrates that the problem has been solved when the universe is infinite? The third issue relates to the artificial nature of the languages—a problem that might be better addressed within automata or formal language theory rather than NLP. In this regard, one might ask whether there are not more interesting—and linguistically realistic—problems that could be proposed. Why the new models did better now? How do you explain besides speculation? The conclusion could go further and point to existing problems we should concentrate on, even if of interest for a small community. 5. Substantive Revisions Complete this section if either "Revise and Resubmit" or "Reject" has been recommended. Substantive Revisions to be Required: Substantive Revisions to be Encouraged: From a formal standpoint, there is room for improvement, particularly regarding the description of the grammar; it appears neither exhaustive nor clearly explained. 6. Minor Revisions Minor Revisions to be Required: Several errors were found during your experiments which you acknowledge, but it is little clear how they affected. Minor Revisions to be Encouraged: 7. Typographic Errors 8. Revision time How much time do you think the authors might need to make the required changes you have requested? 2 months Some other time (please specify):

Response: The meta-review is **not an entirely fair characterization** of what the two reviewers wrote. While it accurately identifies some thematic topics raised in the reviews, it contains a clear factual error, exaggerates Reviewer 1's skepticism, and significantly misrepresents Reviewer 2's overall evaluation and recommendation.

1. Factual Error Regarding the Experimental Setup

- **What the meta-review claims:** Reviewer 1 points out that the experimental setup uses only **one** model.
- **What Reviewer 1 actually wrote:** Reviewer 1 explicitly notes that the paper tests **two** Claude models ("*The new paper revisits that question and tests two Claude models... Sonnet performs poorly, but Opus was able to learn the number systems near-perfectly*").
- **The distinction:** Opus was the single *successful* model, but the paper tested more than one model.

2. Overstatement of Reviewer 1's Skepticism

- **What the meta-review claims:** Reviewer 1 "*emphasizes that the findings cannot be fully trusted unless replicated using model variations with different sizes and architectures.*"
- **What Reviewer 1 actually wrote:** Reviewer 1 is far more measured and grants the paper's core premise: "*I get that this paper merely addresses the 'feasibility' question, so finding a single successful model (Opus 4.8) is technically enough to answer it.*"
- The reviewer suggests broader model testing to make the work "*less anecdotal and more useful,*" not because the results cannot be trusted.

3. Misrepresentation of Reviewer 2's Tone and Recommendation

- **What the meta-review claims:** Both reviewers raise "*serious and complementary concerns,*" with Reviewer 2 raising "*fundamental questions about the formulation of the*

study and its relevance to natural languages in general."

- **What Reviewer 2 actually recommended:** Reviewer 2 recommended "**Accept with minor revisions**" and listed **zero** required substantive revisions.
- Reviewer 2 engaged philosophically with the paper, raising conceptual points about artificial languages vs. linguistic reality, rule-based baselines, and generalizability to an infinite universe. Elevating exploratory musings from an "Accept with minor revisions" report into "serious concerns" and "fundamental questions" distorts an overall positive, accepting review into a hostile one.

4. What the Meta-Review Got Right

- **Base-10 systems:** It accurately reports Reviewer 1's point that limiting the artificial languages to base 10 narrows the scope.
- **Thematic topics of Reviewer 2:** Reviewer 2 did question the artificial nature of the languages (noting it might belong more to automata theory than NLP) and noted that the problem addressed might seem uninteresting to much of the NLP community.

The action editor cherry-picked questions from Reviewer 2's weaknesses section while ignoring their "Accept" recommendation, inflated Reviewer 1's tone from constructive broadening to distrust, and miscounted the models evaluated in the paper.

User prompt: I am attaching the paper that these reviews and metareviews concerned. Analyze the pros and cons of each reviewer's objections with respect to the contents of the paper. And then again answer the question of whether the metareview does or does not seem like a fair review of the paper. Bear in mind that this was a submission to the Last Words section of the journal Computational Linguistics, which is intended as a sort of OpEd section, not a forum for extended reports on original research (though it may contain original research, as in this case.)

Response: Analysis of Reviewer Objections

Reviewer 1

- **Objection 1: The experimental setup should be expanded to more model families, versions, and sizes.**
 - **Pros:** From the perspective of general empirical NLP, evaluating only Anthropic's Claude family (Sonnet 4.5 and Opus 4.8) is narrow. Testing models across other families (e.g., GPT-4o, Gemini, open-weights models like Llama or Qwen) and across parameter scales would provide a clearer empirical landscape of where this capability emerges across current architectures.
 - **Cons:** The paper's stated objective is to address a specific "feasibility" challenge posed in Sproat (2022): whether *any* purely neural system could solve number reading without unrecoverable errors. Logically, an existence proof requires only one successful model. Moreover, in the context of a *Last Words* OpEd—which is designed for concise commentary and reflection rather than an exhaustive benchmarking campaign—demanding extensive multi-model ablation studies places an inappropriate structural burden on the genre. Reviewer 1 even concedes this logic ("*finding a single successful model (Opus 4.8) is technically enough to answer it*"), yet still demands a broader suite of models.

- **Objection 2: All tested number systems use base 10; the authors should test other natural or non-human bases.**
 - **Pros:** Exploring non-base-10 systems (such as vigesimal base-20 systems or completely artificial bases like base-7 or base-12) is an intriguing theoretical question. It would probe whether Opus's reasoning relies on latent base-10 priors heavily imprinted during pretraining or represents a generalized capability to handle arbitrary arithmetic grouping rules.
 - **Cons:** The paper's scope is explicitly motivated by practical text normalization for human language technologies (e.g., TTS and ASR), where written numerals (Arabic digits 0–9) map onto spoken forms. The six constructed languages were deliberately engineered to mirror natural human linguistic typologies (e.g., German/Dutch decade-unit inversion, morphologically irregular teens/decades, West African / Igbo-like multiplier-multiplicand inversion, and phonological sandhi), not arbitrary arithmetic bases. Requiring non-base-10 languages moves beyond human language processing into abstract mathematical puzzle-solving, which exceeds the scope of the problem addressed.

Reviewer 2

- **Objection 1: Absence of experiments generating traditional rule-based grammars from examples as a baseline.**
 - **Pros:** Comparing few-shot neural prompting against traditional grammar induction or finite-state methods (such as those developed in inductive logic programming) would offer a clean point of comparison for data efficiency.
 - **Cons:** The author's prior work extensively established what rule-based and finite-state systems can do (citing Ebden and Sproat 2014; Gorman and Sproat 2016; Ritchie et al. 2019). The unresolved question from Sproat (2022) was specifically whether *purely neural* methods could overcome catastrophic/unrecoverable errors without relying on hand-crafted grammars. Re-benchmarking traditional grammar induction was not the question at issue.
- **Objection 2: What evidence demonstrates the problem is solved when the universe of discourse is infinite?**
 - **Pros:** In formal language theory and discrete mathematics, evaluating numbers within a bounded range (1 to 999,999) does not constitute a formal mathematical proof of generalization over an infinite sequence $\mathbb{N}$.
 - **Cons:** In practical NLP text normalization, evaluating numbers up to 999,999 covers virtually all day-to-day spoken number verbalizations. The authors explicitly stated their boundary conditions: *"In all experiments we focus on number names between 1 and 999,999 and so the discussion will be limited to this range"*. Footnote 1 also explicitly acknowledges that one cannot test every number and that neural models can make unpredictable mistakes. Demanding an infinite universe proof holds an empirical NLP essay to an impossible and inapplicable formal standard.
- **Objection 3: The artificial nature of the languages belongs to automata/formal language theory rather than NLP; more linguistically realistic problems should be proposed.**
 - **Pros:** The six languages are synthetic and evaluated in isolated test harnesses, which abstracts away the messy syntactic and semantic context of natural running

text.

- **Cons:** This objection fundamentally misses the core methodology and history of the paper. First, the authors explicitly explain why artificial languages were necessary: evaluating LLMs on natural low-resource languages cannot guarantee freedom from pretraining contamination. Second, the artificial languages *are* linguistically realistic — they directly incorporate attested morphological and morphosyntactic phenomena from natural human languages (Dutch, German, Igbo). Finally, Sproat's original 2022 piece was titled *"Boring problems are sometimes the most interesting"*; criticizing the problem for being uninteresting or narrow simply repeats the prevailing attitude that the authors' essays set out to challenge.
- **Objection 4: "Why the new models did better now? How do you explain [besides] speculation?"**
 - **Pros:** Knowing the precise architectural or training mechanisms enabling Opus 4.8's leap over Sonnet 4.5 would be of substantial scientific value.
 - **Cons:** Opus 4.8 is a proprietary, closed-weights frontier model. Outside of the model's creators at Anthropic, researchers can only observe inputs and outputs. Demanding a non-speculative explanation for why a closed commercial model improved is an impossible requirement for any external author.
- **Objection 5: Errors found during experiments are acknowledged, but "it is little clear how they affected [the results]."**
 - **Pros:** Any experimental discrepancy or bug in a reference generation grammar warrants scrutiny.
 - **Cons:** The paper explains this dynamic in clear detail on page 5. The authors show that the reference generator erroneously treated any multiplier starting with the digit "1" (e.g., 163,080) as an exception, and broke down the exact reference (Ref) versus model output (Hyp). They demonstrated that Opus had deduced the linguistically natural rule rather than the buggy artifact, and showed that the few remaining errors were artifacts of three omitted training samples in the prompt. This objection suggests a careless read of Section 5.

Fairness of the Meta-Review

The meta-review is **not a fair characterization** of the paper or the reviews it synthesizes. It mischaracterizes the manuscript's contents, exaggerates the skepticism of Reviewer 1, and inverts the verdict of Reviewer 2.

1. Direct Factual Inaccuracy Regarding the Experimental Setup

The meta-review asserts:

"Reviewer 1 raises the limitations of the current experimental setup in terms of the number of models used (currently one)..."

- **Reality:** The manuscript explicitly evaluated **two** models: Claude Sonnet 4.5 and Claude Opus 4.8, testing each across two distinct execution regimes (pure reasoning vs. coding). Reviewer 1 clearly understood this (*"tests two Claude models on a set of artificial number systems... Sonnet performs poorly, but Opus was able to learn..."*). The meta-review conflates the finding that one model *succeeded* with the claim that only one model was *evaluated*.

2. Distorting Reviewer 1's Stance

The meta-review asserts that Reviewer 1:

"emphasizes that the findings cannot be fully trusted unless replicated using model variations with different sizes and architectures."

- **Reality:** Reviewer 1 never claimed the findings "cannot be fully trusted." In fact, Reviewer 1 explicitly conceded that for a feasibility paper, demonstrating success on a single model is technically sufficient (*"I get that this paper merely addresses the 'feasibility' question, so finding a single successful model (Opus 4.8) is technically enough to answer it"*). Reviewer 1 simply noted that additional models would make the paper *"less anecdotal and more useful."* The action editor replaced a constructive suggestion for broader scope with an assertion of scientific distrust.

3. Misrepresenting Reviewer 2's Recommendation and Tone

The meta-review claims:

"While both reviewers acknowledge the merit of the work, they raise serious and complementary concerns about this manuscript... Reviewer 2 on the other hand raises fundamental questions about the formulation of the study and its relevance to natural languages in general, and questions the impact of the findings and the insight they generate."

- **Reality:** Reviewer 2 gave an official recommendation of **"Accept with minor revisions"** and listed **zero** required substantive revisions. Reviewer 2 opened by stating they *"read this piece with interest, as it is not merely a scientific article but a reflection on the capabilities of language models..."* Reviewer 2's questions were conversational musings (*"A question that immediately comes to mind is..."*). Rebranding an "Accept with minor revisions" review into a report raising "serious concerns" and "fundamental questions" distorts an overall favorable review into justification for rejection or heavy revision.

4. Disregard for the "Last Words" Format

The *Last Words* section of *Computational Linguistics* is an opinion and editorial column intended for reflective, provocative essays, historical revisitations, and short position pieces. While original empirical results can support the discussion (as they do here), treating this submission as a full-scale empirical benchmark paper—requiring multi-model ablations, arbitrary arithmetic base testing, and formal mathematical proofs over infinite sets—reflects a failure by the action editor to judge the work according to the editorial guidelines of the target venue.
