[TACL] Editor Decision for Submission 6025: (d)

Alexander Clark via Open Journal Systems <office@transacl.org>

Fri 3/8/2024 13:01

To:Claire Stevenson <C.E.Stevenson@uva.nl>;Mathilde ter Veen <h.m.terveen@uva.nl>;Rochelle Choenni <r.m.v.k.choenni@uva.nl>;Han van der Maas <H.L.J.vanderMaas@uva.nl>; Katia Shutova <e.shutova@uva.nl>

Dear Claire Stevenson, Mathilde ter Veen, Rochelle Choenni, Han van der Maas, Ekaterina Shutova:

Thank you for submitting paper 6025, "Do large language models solve verbal analogies like children do?", to TACL. However, I regret to inform you that the paper will not be accepted, with no possibility of resubmission to TACL for a period of 1 year.

The reviewers expressed concerns with this paper in terms of the framing , presentation and the technical development; I think correcting these flaws would require a major revision that would in effect be a new paper, so I am accordingly rejecting this paper.
I hope that these detailed reviews will nonetheless be helpful in revising this paper for submission either here or elsewhere.

The detailed reviews are below. I hope you will find their feedback useful in your future research.

I am sorry that I cannot give you better news, but we do appreciate your submitting your work to TACL, and wish you the best in your future research.

Sincerely yours, Alexander Clark, CLASP,Gothenburg University, alexsclark@gmail.com

---

Reviewer A:
Recommendation: Decline Submission

CLARITY: For the reasonably well-prepared reader, is it clear what was done and why? Is the paper well-written and well-structured?

4. Understandable by most readers.

INNOVATIVENESS: How original is the approach? Does this paper break new ground in topic, methodology, or content? How exciting and innovative is the research it describes?

Note that a paper can score high for innovativeness even if its impact will be limited.

2. Pedestrian: Obvious, or a minor improvement on familiar techniques.

SOUNDNESS/CORRECTNESS: First, is the technical approach sound and well-chosen? Second, can one trust the claims of the paper -- are they supported by proper experiments and are the results of the experiments correctly interpreted?

2. Troublesome. There are some ideas worth salvaging here, but the work should really have been done or evaluated differently.

RELATED WORK: Does the submission make clear where the presented system sits with respect to existing literature? Are the references adequate?

Note that the existing literature includes preprints, but in the case of preprints:

- Authors should be informed of but not penalized for missing very recent and/or not widely known work.
- If a refereed version exists, authors should cite it in addition to or instead of the preprint.

4. Mostly solid bibliography and comparison, but there are a few additional references that should be included. Discussion of benefits and limitations is acceptable but not enlightening.

SUBSTANCE: Does this paper have enough substance (in terms of the amount of work), or would it benefit from more ideas or analysis?

Note that papers or preprints appearing less than three months before a paper is submitted to TACL are considered contemporaneous with the submission. This relieves authors from the obligation to make detailed comparisons that require additional experiments and/or in-depth analysis, although authors should still cite and discuss contemporaneous work to the degree feasible.

1. Seems thin. Not enough ideas here for a full-length paper.

IMPACT OF IDEAS OR RESULTS: How significant is the work described? If the ideas are novel, will they also be useful or inspirational? If the results are sound, are they also important? Does the paper bring new insights into the nature of the problem?

2. Marginally interesting. May or may not be cited.

REPLICABILITY: Will members of the ACL community be able to reproduce or verify the results in this paper?

4. They could mostly reproduce the results, but there may be some variation because of sample variance or minor variations in their interpretation of the protocol or method.

IMPACT OF PROMISED SOFTWARE: If the authors state (in anonymous fashion) that their software will be available, what is the expected impact of the software package?

2. Documentary: The new software will be useful to study or replicate the reported research, although for other purposes it may have limited interest or limited usability. (Still a positive rating)

IMPACT OF PROMISED DATASET(S): If the authors state (in anonymous fashion) that datasets will be released, how valuable will they be to others?

2. Documentary: The new datasets will be useful to study or replicate the reported research, although for other purposes they may have limited interest or limited usability. (Still a positive rating)

TACL-WORTHY AS IS? In answering, think over all your scores above. If a paper has some weaknesses, but you really got a lot out of it, feel free to recommend it. If a paper is solid but you could live without it, let us know that you're ambivalent.

Reviewers: after you save this review form, you'll have to make a confidential recommendation to the editors via pull-down menu as to: *what degree of revision would be needed to make the submission eventually TACL-worthy?*

2. Leaning against: I'd rather not see it appear in TACL.

Detailed Comments for the Authors

Reviewers, please draft your comments on your own filesystem and then copy the results into the text-entry box. You will thus have a saved copy in case of system glitches.

This paper explores lexical analogies in LLMs to study whether models solve them similarly to human children. The paper provides a lengthy background survey on analogy in the language acquisition and NLP literatures, collects Dutch data from school aged children, and uses that data to evaluate Dutch LLMs on analogy. Ultimately, the models are shown to use associative reasoning more than older children but their accuracy is similar to younger children.

While I think the topic and dataset seem interesting, the paper could have gone much deeper, and I have serious reservations about the framing.

Depth:
The paper mainly uses a few categorical accuracy measures to assess whether the model behaves like children. Beyond looking for associative distractors, no real error analysis is provided of whether there are patterns in the preferred distractors for each item and how those compare with children. One short paragraph is devoted to this as Section 7.5, but I had hoped that this would be a much bigger part of the paper. Similarly, the models are presumed to select the highest probability item, but there is no analysis of the actual probability associated with the highest likelihood option or of the probability margin between the most likely item and the distractors to establish the reliability of a given selection. I understand that such analyses would not be directly comparable with children but it would help get at the extent to which models are processing the analogies in a reasonable way. Twice on the first page the paper talks about studying the process by which models reason about analogy (Lines 69-71 and 96), but the paper doesn't really touch on this beyond associativity and accuracy measures binned by item category. Each experiment and its result was presented in just 1-2 short paragraphs each. Combined with the fact that the first experiment is only described on Page 7, the paper felt very light on analyses. Much of the first four pages can probably be removed to make room for deeper error analyses examining the response patterns of models or comparing error patterns in both models and humans.

Technical Issues:
The conceptual distance and distractor salience measures were computed using word2vec and fasttext models, but it's not clear that the contextual models actually being evaluated in the paper share the similarity judgments of these baseline models. Therefore, it's possible that the analysis fails to control for distractor salience in each of the models. Why wouldn't salience be computed on a model-by-model basis (yielding a GPT-3 salience, an XLM-V salience, etc)? Then the mean salience across all models could be used for the children.

It also wasn't clear to me why this analysis was conducted over binned factors rather than treating salience and conceptual distance as continuous factors.

The association-corrected analyses were only conducted on the "best performing models" of XLM-V and GPT-3, but GPT-3 is not a good stand-in for general language modeling behavior since GPT-3 received explicit supervised training by adult humans through RLHF which likely included solving analogies, so analyses of it likely actually overestimate the capability of these models. This should probably be discussed in the paper.

Framing:
The paper is framed in a confusing and possibly misleading way. One of the clearest examples of this was in Figure 2, which plots mean child accuracy by year (ignoring sub year increments), connected by lines. Then models are placed on this line where the child age:accuracy line coincides with the model accuracy, and the misleading caption "At what age level does each LLM perform?" is used. Any model with an accuracy between 30 and 60 percent can therefore be placed on this line even if the response pattern differs entirely from that of the associated age's response pattern. We cannot conclude that the model's underlying process is similar to a child's from these accuracy comparisons. Much later in the text, the authors acknowledge that only 30-40% errors were shared between models and humans, but a very likely outcome would be for Figure 1 to be shared on social media without this nuance. More frustratingly, by placing model accuracies on such a plot, the figure implies some sort of developmental progression on the part of models similar to children. Instead, I think it would be more accurate for the models to each be treated as samples from the population of models and get compiled into a single mean accuracy dot with error bars to show model accuracy, similar to how each dot on the human chart is made up of several samples from the human population. Or maybe the chart could group models by class. These model accuracy boxes could be placed next to the child accuracy plot rather than on the language development line itself. A similar criticism could be leveled against Figure 6.

I couldn't actually tell what conclusion the paper intended to convey. The overt paper conclusion in the final sentence describes the results as evidence _against_ emergent analogical reasoning, but many of the research questions seemed framed optimistically for the models. For example, RQ1 asked whether recent LLMs solve analogy similar to human 12 year olds (which is then equated with adult-like performance). Three models obtained accuracies close 11 year olds, so the paper claims that models do process analogy similar to older children or adults. Shouldn't the conclusion instead be that no model achieved 12 year old performance or, if the goal is shifted to 11 year old performance (such shifting would need to be explicitly highlighted in the paper), why is the conclusion not that 5/8 models failed to achieve such performance? As is, I think readers will leave this paper thinking that it shows evidence _for_ emergent analogical reasoning (the opposite of the conclusion sentence) due to the framing of the rest of the paper.

Minor:
Are there error bars on the human results for Figures 1 and 6?

Figure 3 and Line 743 mention that GPT-3 follows adult-like patterns but adult accuracies weren't included in the paper for comparison.

Lines 820-823 seem to be a drafting error.

REVIEWER CONFIDENCE

3. Pretty sure, but there's a chance I missed something. Although I have a good feel for this area in general, I did not carefully check the paper's details, e.g., the math, experimental design, or novelty.

Reviewer D:
Recommendation: Resubmit for Review

CLARITY: For the reasonably well-prepared reader, is it clear what was done and why? Is the paper well-written and well-structured?

4. Understandable by most readers.

INNOVATIVENESS: How original is the approach? Does this paper break new ground in topic, methodology, or content? How exciting and innovative is the research it describes?

Note that a paper can score high for innovativeness even if its impact will be limited.

3. Respectable: A nice research contribution that represents a notable extension of prior approaches or methodologies.

SOUNDNESS/CORRECTNESS: First, is the technical approach sound and well-chosen? Second, can one trust the claims of the paper -- are they supported by proper experiments and are the results of the experiments correctly interpreted?

5. The approach is very apt, and the claims are convincingly supported.

RELATED WORK: Does the submission make clear where the presented system sits with respect to existing literature? Are the references adequate?

Note that the existing literature includes preprints, but in the case of preprints:

- Authors should be informed of but not penalized for missing very recent and/or not widely known work.
- If a refereed version exists, authors should cite it in addition to or instead of the preprint.

5. Precise and complete comparison with related work. Benefits and limitations are fully described and supported.

SUBSTANCE: Does this paper have enough substance (in terms of the amount of work), or would it benefit from more ideas or analysis?

Note that papers or preprints appearing less than three months before a paper is submitted to TACL are considered contemporaneous with the submission. This relieves authors from the obligation to make detailed comparisons that require additional experiments and/or in-depth analysis, although authors should still cite and discuss contemporaneous work to the degree feasible.

4. Represents an appropriate amount of work for a publication in this journal. (most submissions)

IMPACT OF IDEAS OR RESULTS: How significant is the work described? If the ideas are novel, will they also be useful or inspirational? If the results are sound, are they also important? Does the paper bring new insights into the nature of the problem?

3. Interesting but not too influential. The work will be cited, but mainly for comparison or as a source of minor contributions.

REPLICABILITY: Will members of the ACL community be able to reproduce or verify the results in this paper?

3. They could reproduce the results with some difficulty. The settings of parameters are underspecified or subjectively determined, and/or the training/evaluation data are not widely available.

IMPACT OF PROMISED SOFTWARE: If the authors state (in anonymous fashion) that their software will be available, what is the expected impact of the software package?

1. No usable software will be released.

IMPACT OF PROMISED DATASET(S): If the authors state (in anonymous fashion) that datasets will be released, how valuable will they be to others?

1. No usable software will be released.

TACL-WORTHY AS IS? In answering, think over all your scores above. If a paper has some weaknesses, but you really got a lot out of it, feel free to recommend it. If a paper is solid but you could live without it, let us know that you're ambivalent.

Reviewers: after you save this review form, you'll have to make a confidential recommendation to the editors via pull-down menu as to: *what degree of revision would be needed to make the submission eventually TACL-worthy?*

3. Ambivalent: OK but does not seem up to the standards of TACL.

Detailed Comments for the Authors

Reviewers, please draft your comments on your own filesystem and then copy the results into the text-entry box. You will thus have a saved copy in case of system glitches.

The paper evaluates a couple of LLMs on the task of analogy solving. The basis for the experiments is a new dataset of analogies solved by children, collected from an online educational platform in the Netherlands. This makes it possible to compare LLM performance to the performance of children of different ages. The evaluation is designed around three main hypotheses. The results show that the best-performing LLMs (e.g. GPT-3) achieve the performance of older children, that their performance is sensitive to similar factors as children's performance, and that at least some of the analogies seem to be solved with associative reasoning, rather than real analogical reasoning.

Strengths:

The paper is well-written and very clear. I really appreciated the very systematic structure that centers around the 3 main research questions.

The dataset collected and presented in the paper is very rich, and a great basis for studying analogies (e.g., it includes fine-grained data on item difficulty ratings, errors from many participants, analogy relations, ...).

Analogies have been an important phenomenon in the evaluation of embedding models, for a long time. It is definitely interesting to see how recent models perform on this task and to do more fine-grained analyses of factors that affect the performance (like analogy relation).

Weaknesses:

One of the main differences to existing studies on analogies in LLMs is that the dataset used here has analogy data from children, and not from adults. Yet, in the end, I was not sure what the comparison with children data really tells us, since the best performing LLM is expected to perform on par with older children in the first place.

The most interesting part of the paper is the last part of the evaluation (Sec. 7), which asks whether LLMs may use associative rather than analogical reasoning to solve analogies. In this part, though, I found the analysis procedure a bit disappointing, and not entirely clear. The main idea is to manipulate the analogy template (A:B::C:D), e.g. by filtering items where LLMs can predict the correct D, when only seeing C (C:D). When these items are filtered, LLM performance is lower across the board, and it drops a bit more than it does for children. However, I am not convinced that this is enough evidence to conclude that LLMs only use associative reasoning. In fact, many of the remaining difficult items can still be solved correctly. Sec 7.3 and 7.4 present another manipulation that is supposed to further support the claim, but I could not follow it completely (plus, sec 7.4. is looking at a very small number of items)

Overall, the main idea of the paper and many of the analyses are pretty similar to two existing papers on analogy solving in LLMs: Ushio et al. 2021, and Webb et al. 2023. The results also seem to largely confirm these existing studies, except the fact that the authors contradict Webb et al.'s assumption that analogical reasoning "emerges" in LLMs. However, the claim that LLMs rely on associative rather than analogical reasoning when solving analogies is not substantiated enough, to really count as a solid contribution of this paper.

REVIEWER CONFIDENCE

4. Quite sure. I tried to check the important points carefully. It's unlikely, though conceivable, that I missed something that should affect my ratings.

Reviewer G:
Recommendation: Resubmit for Review

CLARITY: For the reasonably well-prepared reader, is it clear what was done and why? Is the paper well-written and well-structured?

3. Mostly understandable to me (a qualified reviewer) with some effort.

INNOVATIVENESS: How original is the approach? Does this paper break new ground in topic, methodology, or content? How exciting and innovative is the research it describes?

Note that a paper can score high for innovativeness even if its impact will be limited.

4. Creative: An intriguing problem, technique, or approach that is substantially different from previous research.

SOUNDNESS/CORRECTNESS: First, is the technical approach sound and well-chosen? Second, can one trust the claims of the paper -- are they supported by proper experiments and are the results of the experiments correctly interpreted?

2. Troublesome. There are some ideas worth salvaging here, but the work should really have been done or evaluated differently.

RELATED WORK: Does the submission make clear where the presented system sits with respect to existing literature? Are the references adequate?

Note that the existing literature includes preprints, but in the case of preprints:

- Authors should be informed of but not penalized for missing very recent and/or not widely known work.
- If a refereed version exists, authors should cite it in addition to or instead of the preprint.

4. Mostly solid bibliography and comparison, but there are a few additional references that should be included. Discussion of benefits and limitations is acceptable but not enlightening.

SUBSTANCE: Does this paper have enough substance (in terms of the amount of work), or would it benefit from more ideas or analysis?

Note that papers or preprints appearing less than three months before a paper is submitted to TACL are considered contemporaneous with the submission. This relieves authors from the obligation to make detailed comparisons that require additional experiments and/or in-depth analysis, although authors should still cite and discuss contemporaneous work to the degree feasible.

2. Work in progress. There are enough good ideas, but perhaps not enough results yet.

IMPACT OF IDEAS OR RESULTS: How significant is the work described? If the ideas are novel, will they also be useful or inspirational? If the results are sound, are they also important? Does the paper bring new insights into the nature of the problem?

4. Some of the ideas or results will substantially help other people's ongoing research.

REPLICABILITY: Will members of the ACL community be able to reproduce or verify the results in this paper?

4. They could mostly reproduce the results, but there may be some variation because of sample variance or minor variations in their interpretation of the protocol or method.

IMPACT OF PROMISED SOFTWARE: If the authors state (in anonymous fashion) that their software will be available, what is the expected impact of the software package?

1. No usable software will be released.

IMPACT OF PROMISED DATASET(S): If the authors state (in anonymous fashion) that datasets will be released, how valuable will they be to others?

4. Useful: I would recommend the new datasets to other researchers or developers for their ongoing work.

TACL-WORTHY AS IS? In answering, think over all your scores above. If a paper has some weaknesses, but you really got a lot out of it, feel free to recommend it. If a paper is solid but you could live without it, let us know that you're ambivalent.

Reviewers: after you save this review form, you'll have to make a confidential recommendation to the editors via pull-down menu as to: *what degree of revision would be needed to make the submission eventually TACL-worthy?*

2. Leaning against: I'd rather not see it appear in TACL.

Detailed Comments for the Authors

Reviewers, please draft your comments on your own filesystem and then copy the results into the text-entry box. You will thus have a saved copy in case of system glitches.

This paper compares how language models solve analogies vs. how children solve analogies, using data from children in the Netherlands. The authors perform various analyses, e.g., comparing the performance accuracies between different categories of analogies, and analyzing the effect of distractors. Overall, while I believe this paper has quite some potential - the data and the questions are super interesting- the paper requires a major revision, since various analysis steps were not clearly motivated, and details were missing.

Strengths:

- The paper is well written. I appreciated that the authors are clear about the different data processing steps (e.g. item selection

criteria).
- The dataset is very interesting (it is only released upon request), and allows some intriguing comparisons between how humans solve analogies vs. LLMs.
- The research questions are interesting and well motivated.

Weaknesses:

The analyses can be more rigorous, are sometimes missing details, and some steps should be better motivated. See some more detailed comments below.

- It was not clear to me how the performance of LLMs was exactly compared to children, as the analyses primarily focus on performance numbers (accuracy). Was item difficulty taken into account at all? E.g. being 90% accurate on easy items means something different than a 90% score on difficult items. The children saw items matched to their abilities. How did you take this into account when comparing against LLMs? It would also be useful to have an idea of the how performance varies within the children (e.g. Figure 2).
- The used methodology to solve analogies using LLMs can be motivated more clearly. In particular, I was wondering how your approach compared to the other studies you cited that also considered similar models (e.g. Ushio et al.), since your approach is limited to single-token words.
- Line 493: I did not understand step (1). Doesn't this favor XLM-V? How did you handle cases where tokens were not in the vocabulary of other models? Related to this, I don't think you can say that tokens that are not in a model's vocabulary are not known (950). A model might have seen a token multiple times, but yet the token might be split up in the vocabulary.
- Section 6: It was not clear how this analysis was carried out. How exactly did you fit a logistic regression model? How well did it fit the data?
- I had really trouble following 7.3 and 7.4; I read these sections multiple times. E.g. 856: why was this expected? (H3b was not mentioned earlier). Furthermore, performance dropped slightly, yet you say 'improved' in 864. it is in short not immediately clear to me what this result tells you about how LLMs solve analogies.
- RQ3: It seems to me that your results show that both associative and relational reasoning seems to play a role, since even in the most difficult experiment performance was better than chance. So I wonder if the question is phrased too strong (associative processes -or- analogical reasoning) and also whether your conclusion 935-937 is phrased too strongly.

Small comments:
- 131: 'to lead to..'
- Linzen 2016 should be cited (https://aclanthology.org/W16-2503/), as it also raises attention to problems with analogy tasks with word embeddings. The proposed baselines in this paper are relevant, especially the Only-B baseline is the same as your C? baseline

- The formatting in the reference list can be checked, e.g., there are some capitalization issues (e.g., Ushio et al's paper title)
- Figures 3-5 are hard to read in grayscale, children vs XLM-V are hard to distinguish.
- Paragraph starting at 330: This paragraph was a bit hard to follow, and could be explained more clearly. For example, I did not immediately understand what Ushio et al. did based on the description (e.g., what are the head and tail in this context). I also didn't fully understand how 338-349 followed from that.
- 853: misses a)
- 820-823: ('Of the x% items …') Are some numbers missing here?
- 284: does this contradict with contribution (1) in line 93?
- 395 footnote 1: How does your conclusion follow from GPT-4's performance?
- Data collection: Can you clarify the time period that this data covers?
- Did you inspect the distributions of the cosine similarity scores, to see whether the threshold you chose made sense? How cosine scores are distributed can vary widely between models.
- 355: cite the Vries and Nissim instead of Radford et al. for D-GPT-2
- A very recent paper (EMNLP 2023, so I understand that the authors missed it) that is relevant " Can Language Models Learn Analogical Reasoning? Investigating Training Objectives and Comparisons to Human Performance" https://aclanthology.org/2023.emnlp-main.1022.pdf
- Figure 2 was hard to understand. For example, it was not immediately clear why the 7-year olds outperform the baselines. Are they the squares?

REVIEWER CONFIDENCE

3. Pretty sure, but there's a chance I missed something. Although I have a good feel for this area in general, I did not carefully check the paper's details, e.g., the math, experimental design, or novelty.
