Quality Assessment of Open-Source Language Model Generated Sentence Pairs in Rotorcraft Aviation Domain

Emma L. McDaniel and Alicia I. Ruvinsky

Department of Computer Science
Georgia State University

US Army Corps Engineer Research
Development Center

INTRODUCTION

- Rotorcraft decision support system
- Mining incident and accident reports for information deficiency landscape
- To data-mine, it is necessary to fine-tune for semantic similarity [1]
- To fine-tune using contrastive learning, sentence pairs are required, not in dataset
- Endeavor to generate sentences pairs using open-source InstMixtral (8x7B SMoE) [2, 3] requires assessment due to hallucinations

METHODOLOGY

Evaluated generated sentences for 50 samples from our dataset across 7 prompts totaling 350 sentences.

Quality Assessment Metrics:

1) Stemmed word [4] similarity (BoW & Location [5])
2) Part of speech [6] similarity (BoW & Location)
3) Readability Scores [7] (Flesch-Kincaid Grade Level Score and Flesch Readability Ease)
4) Cosine Similarity [8] by embedding sentences using MPNet [9] (Semantic similarity)
5) Percentage of extra generated text

Table 1: Seven prompts utilized to generate sentences

#	Prompt
1	Rewrite this sentence:
2	Rewrite this sentence to be more informal:
3	Rewrite this sentence to be more formal:
4	Reorganize this sentence:
5	Rewrite this sentence to be more concise:
6	Rewrite this in another way:
7	Rewrite this sentence using different vocabulary:

RESULTS

Table 2: Median scores calculated across prompts (excluding Flesch-Kincaid, Flesh, and Extra %). "Average" is the mean of median scores per prompt. "*OS*" stands for original sentence score in readability metrics.

Metrics	*Average*	*Prompt 1*	*Prompt 2*	*Prompt 3*	*Prompt 4*	*Prompt 5*	*Prompt 6*	*Prompt 7*
Stem Union %	63.05%	71.42%	48.53%	64.17%	91.67%	71.43%	60.77%	33.33%
Stem Lev Dist %	80.67%	78.47%	81.82%	82.09%	73.03%	63.64%	89.17%	96.43%
POS Union %	99.20%	100%	100%	100%	100%	94.38%	100%	100%
POS Lev Dist %	57.02%	72.25%	50%	50%	64.06%	50%	62.83%	50%
Flesch-Kincaid	*OS:12.14*	13.61	7.93	16.3	13.01	10.57	13.61	14.65
Flesch	*OS:45.2*	38.08	71.4	25.34	40.87	49.75	39.37	29.14
Cosine Sim	0.836	0.869	0.782	0.823	0.907	0.833	0.859	0.777
Extra %	12%	8%	28%	18%	8%	14%	2%	6%

- Stemmed Union: Average across prompts is ~63%. Prompt 4 has the highest (91.67%); Prompt 7 has the lowest (33.33%)
- Stemmed Levenshtein Distance: Average across prompts is high (80.67%); Prompt 5 has the most similarity of word order (63.64%); Prompt 7 the most (96.43%)
- Part of Speech Union: high across all prompts due to low number of tags
- Part of Speech Levenshtein Distance: ordering is on average 57%
- Flesch-Kincaid Grade Level and Flesch Reading Ease: Prompt 2 had the lowest grade level score and the highest readability ease
- Cosine Similarity: Embeddings to original sentences produced low similarities despite perceived semantic closeness observed in manual evaluation
- Extra Sentences/Text: 12% had extraneous results in the prompt, with Prompt 2 having the highest at 28%

REFERENCES

[1] Emma L. McDaniel and Alicia I. Ruvinsky "Building Holistic Situation Awareness through Large Language Models." In proceedings of *Cognitive Situation Management 2023* (CogSIMA '23); 16-20 October 2023 in Philadelphia, Pennsylvania. *Proceedings to be published.*

[2] INSTMixtral. Mixtral-8x7B-Instruct-v0.1. Version: 4.36.2, Docs: https://huggingface.co/mistralai/Mixtral-8x7B-Instruct-v0.1

[3] Jiang, A. Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., Chaplot, D. S., Casas, D. d. l., Hanna, E. B., Bressand, F., et al. (2024). Mixtral of Experts. arXiv preprint arXiv:2401.04088.

[4] NLTK. Natural Language Toolkit. Version: 3.7.2, Docs: https://www.nltk.org/api/nltk.stem.html

[5] Levenshtein Python C Extension Module. Version: 0.24.0, Docs: https://github.com/rapidfuzz/Levenshtein

[6] spaCy. Version: 3.7.2, Docs: https://spacy.io/usage/linguistic-features

[7] Readability Metrics. Version: 1.4.5, Docs: https://pypi.org/project/py-readability-metrics/

[8] Scikit-Learn. Version: 1.4.0, Docs: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.pairwise.cosine_ similarity.html

[9] MPNet. sentence-transformers/all-mpnet-base-v2. Version: 4.36.2, Docs: https://huggingface.co/sentence-transformers/all-mpnet-base-v2

CONCLUSION

Prompt 7 generates the most *different* sentences and with manual evaluations, the sentences still have similar semantic meanings.

Extra text generation occurs the most with Prompt 2.

Cosine Similarity between original sentence and generated sentences did not result in meaningful comparisons for the two sentences.

Future work: utilize different semantic similarity metrics and other generative LLMs.

REPRODUCIBILITY

The code and dataset utilized in this work are available on our open science framework project repository: https://osf.io/9rtfy/

ACKNOWLEDGEMENTS

This research was supported in part by an appointment to the Department of Defense (DOD) Research Participation Program administered by the Oak Ridge Institute for Science and Education (ORISE) through an interagency agreement between the U.S. Department of Energy (DOE) and the DOD. ORISE is managed by ORAU under DOE contract number DE-SC0014664. All opinions expressed in this paper are the author's and do not necessarily reflect the policies and views of DOD, DOE, or ORAU/ORISE.