3^{rd} International Conference of Science of Science & Innovation, July 1-3, 2024
National Academy of Sciences, Washington, DC, USA

Can LLMs Discern Evidence for Scientific Hypotheses? Case Studies in the Social Sciences

Sai Koneru, Jian Wu, Sarah Rajtmajer

Keywords: Large Language Models, Scientific Hypothesis Evidencing, Natural Language Understanding

Extended Abstract

Hypothesis formulation and testing are central to empirical research. However, with exponential increase in the number of scientific articles published annually, manual aggregation and synthesis of evidence related to a given hypothesis is a challenge. Scholarly databases fail to aggregate, compare, contrast, and contextualize existing studies in a way that allows comprehensive review of the relevant literature. Work in the areas of natural language processing (NLP) and natural language understanding (NLU) has emerged to address various challenges related to synthesizing scientific findings. Automated approaches for *fact-checking* [3], for example, have received significant attention in the context of misinformation to assess the accuracy of a factual claim based on a literature [5]. What remains a gap, however, are methods to determine whether a research question is addressed within a paper based on its abstract, and if so, whether the corresponding hypothesis is supported or refuted by the work. Automatically identifying published work in support or refute of particular hypotheses will facilitate building connections between publications beyond citations and aggregating scientific contributions to automatically and dynamically evaluate hypotheses with strong and weak evidence.

In this work accepted at **LREC-COLING 2024**, our contributions are as follows. First, we propose the **scientific hypothesis evidencing** (SHE) task which is defined as the identification of the association between a given declarative hypothesis and a relevant abstract. This association can be labeled either *entailment*, *contradiction*, or *inconclusive*.

Second, we curate a novel Collaborative Reviews (CoRe) dataset for the task using community-driven annotations of studies in the social sciences. **Our CoRe dataset** is built from 12 different open-source collaborative literature reviews actively curated and maintained by domain experts and focused on specific questions in the social and behavioral sciences. The dataset contains 69 unique hypotheses tested across 602 different scientific articles. The findings are aligned to 3 labels leading to a total of 638 triplets containing abstract, hypothesis, and label. We split the dataset into to training (70%), development (15%), and held-out test (15%) sets.

Finally, **we evaluate state of the art NLU models** on the SHE task. Specifically, we evaluated two families of NLP methods on the task using our dataset: transfer learning models; LLMs. In the case of transfer learning models, we evaluate sentence pair classifiers based on pre-trained embeddings and Natural Language Inference models. For the sentence pair classification, concatenated hypothesis and abstract embeddings are used as input to the model, which contains three successive fully-connected layers followed by a three-way softmax layer. We evaluate the performance of two pre-trained embedding models: *longformer* [1]; and OpenAI's *text-embedding-ada-002*. In case of Natural Language Inference models, we use an abstract as the premise and determine whether it entails a given hypothesis. Among models proposed for the NLI task, we evaluate the Enhanced Sequential Inference Model (ESIM) [2] and Multi-Task Deep Neural Network (MT-DNN) [4].

We tested two LLMs, namely OpenAI's ChatGPT and Google's PaLM 2, and experimented with five prompts used in prior work. All are *prefix* prompts, i.e., prompt text comes entirely

Type	Model	Setting	Accuracy	macro F1
Sentence pair classification	Longformer	Supervised on CoRe	65.60%	0.558
	text-embedding-ada-002	Supervised on CoRe	**70.31%**	**0.615**
Transfer learning using NLI models	MT-DNN	Fine-tuned on CoRe	67.97%	0.523
		Fine-tuned on SNLI	42.97%	0.342
	ESIM	Supervised on CoRe	64.84%	0.489
		Supervised on SNLI	39.84%	0.335
LLM	ChatGPT	Zero-shot w/o ensemble	47.22%*	0.414*
		Few-shot w/o ensemble	59.85%*	0.517*
		Zero-shot with ensemble	53.94%	0.500
		Few-shot with ensemble	66.57%	0.576
	PaLM 2	zero-shot w/o ensemble	59.78%*	0.504*
		Few-shot w/o ensemble	69.78%*†	0.583*†
		Zero-shot with ensemble	62.87%	0.536
		Few-shot with ensemble	76.40%	0.678*†

* Mean of responses across all temperatures, prompt templates, and iterations † Incomplete responses

Table 1: Results summarizing the performance of models on the held-out set under different settings.

before model-generated text. Depending on the prompt template, we requested LLMs return one of three sets of labels: *(true, false, neutral)*; *(yes, no, maybe)*; *(entail, contradict, neutral)*. We tested the models in a zero-shot setting, retrieval-augmented few-shot, and using prompt ensembling with majority voting to ensemble the outputs of our five individual prompts. Table 1 summarizes model performance on the test set. Reported metrics are averaged across experimental settings. The sentence pair classification model using *text-embedding-ada-002* embeddings yielded the best performance achieving a macro-F1-score of 0.615, followed by the pre-trained gpt-3.5-turbo model with prompt ensembling in the few-shot setting.

The observation that all models achieve macro-F1-scores less than 0.65 demonstrates that SHE is a challenging task for current NLU and that LLMs do not seem to perform better than traditional language models and transfer learning models. Our study quantitatively showcases the limited reasoning capability of state of the art LLMs and suggests there is still a ways to go before LLMs are readily usable for discerning evidence of scientific hypotheses, at least in the social sciences. Our dataset has been shared with the research community.[1]

References

[1] I. Beltagy, M. E. Peters, and A. Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020.

[2] Q. Chen, X. Zhu, Z.-H. Ling, S. Wei, H. Jiang, and D. Inkpen. Enhanced LSTM for natural language inference. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2017.

[3] Z. Guo, M. Schlichtkrull, and A. Vlachos. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10:178–206, 2022.

[4] X. Liu, P. He, W. Chen, and J. Gao. Multi-task deep neural networks for natural language understanding, 2019.

[5] J. Vladika and F. Matthes. Scientific fact-checking: A survey of resources and approaches. *arXiv preprint arXiv:2305.16859*, 2023.

[1] `https://github.com/Sai90000/ScientificHypothesisEvidencing.git`