Can Large Language Models Discern Evidence for Scientific Hypotheses? Case Studies in the Social Sciences

Sai Koneru[1], **Jian Wu**[2], **Sarah Rajtmajer**[1]
[1] Pennsylvania State University, State College, PA
[2] Old Dominion University, Norfolk, VA
{sdk96, smr48}@psu.edu, j1wu@odu.edu

Abstract

Hypothesis formulation and testing are central to empirical research. A strong hypothesis is a best guess based on existing evidence and informed by a comprehensive view of relevant literature. 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. Our work explores the ability of current large language models (LLMs) to discern evidence in support or refute of specific hypotheses based on the text of scientific abstracts. We share a novel dataset for the task of *scientific hypothesis evidencing* using community-driven annotations of studies in the social sciences. We compare the performance of LLMs to several state of the art methods and highlight opportunities for future research in this area. Our dataset is shared with the research community: `https://github.com/Sai90000/ScientificHypothesisEvidencing.git`

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

1. Introduction

Translating scholarly research findings into actionable, evidence-based impacts relies on iterative refinement for robust understanding of a given phenomenon across multiple studies, contexts, etc. The sequential approach to scientific interrogation is also at the heart of null hypothesis significance testing. Namely, a hypothesis is an informed theory, or an *educated guess*, based on available information and prior findings (Wald, 1992). As such, synthesis and understanding of current literature is essential to study planning and to efficient research more broadly. Yet, scholarly databases fail to aggregate, compare, contrast, and contextualize existing studies in a way that allows comprehensive review of the relevant literature in service to a targeted research question. In part, this is because the sheer volume of published work is difficult to navigate and the narrative format through which most empirical work is reported was not envisioned with machine readability in mind.

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* (Guo et al., 2022), for example, have received significant attention in the context of misinformation and disinformation. This task aims to assess the accuracy of a factual claim based on a literature (Vladika and Matthes, 2023). 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. In this work, we propose this task as *scientific hypothesis evidencing* (SHE).

Notably, grassroots efforts to usefully assemble the literature have popped up, e.g., in the form of shared Google docs, contributed to by authors of related work and socialized primarily via Twitter (Haidt and Bail, 2022). In these documents, authors synthesize existing work that tests closely-related hypotheses or a similar research question, e.g., *Does social media cause political polarization?* Chapters within these collaborative documents add additional structure, highlighting studies with similar outcomes or similar experimental settings.

In this work, we study whether and to what extent state-of-the-art NLU and large language models can supplant manual expert-driven collaborative meta-analyses, or parts thereof, in service to discerning hypotheses and primary findings from scientific abstracts. We focus on the social sciences due to the availability of high quality datasets annotated by domain experts. Our work makes the following primary contributions:

1. We propose the SHE task–the identification of evidence from the abstract of a scientific publication in support or refute of a hypothesis;
2. We build and share a benchmark dataset for SHE using expert-annotated collaborative literature reviews;
3. Using this dataset, we evaluate performance of state of the art transfer learning and large language models (LLMs) for the SHE task.

Our findings suggest that this task is challenging for current NLU and that LLMs do not seem to per-

LREC-COLING 2024, pages 2787–2797
20-25 May, 2024.

Research question (from the review). Is there an association between social media use and bad mental health outcomes?
Abstract. Although studies have shown that increases in the frequency of social media use may be associated with increases in depressive symptoms of individuals with depression, the current study aimed to identify specific social media behaviors related to major depressive disorder (MDD). Millennials (N = 504) who actively use Facebook, Twitter, Instagram, and/or Snapchat participated in an online survey assessing major depression and specific social media behaviors. Univariate and multivariate analyses were conducted to identify specific social media behaviors associated with the presence of MDD. The results identified five key social media factors associated with MDD. Individuals who were more likely to compare themselves to others better off than they were (p = 0.005), those who indicated that they would be more bothered by being tagged in unflattering pictures (p = 0.011), and those less likely to post pictures of themselves along with other people (p = 0.015) were more likely to meet the criteria for MDD. Participants following 300 + Twitter accounts were less likely to have MDD (p = 0.041), and those with higher scores on the Social Media Addiction scale were significantly more likely to meet the criteria for MDD (p = 0.031). Participating in negative social media behaviors is associated with a higher likelihood of having MDD. Research and clinical implications are considered.
Hypothesis (declarative). There is an association between social media use and bad mental health outcomes.
Label. Entail

Table 1: An example from the training dataset containing a paper's *abstract*, a *hypothesis* of interest, and corresponding *label* identifying the relationships between the hypothesis and the abstract.

form better than traditional language models and transfer learning models. We offer perspectives and suggestions for the path forward.

2. Related Work

The task of *scientific claim verification* is treated either: (1) as a natural language inference (NLI) problem using deep neural networks trained on human-annotated datasets (Khot et al., 2018; Wadden et al., 2022); or, (2) as a classification problem using a joint claim-evidence representation (Oshikawa et al., 2020). In support of these efforts, multiple *claim verification* datasets have been proposed as benchmarks for the community, e.g., for topics in biomedical sciences (Wadden et al., 2020, 2022), public health (Kotonya and Toni, 2020; Sarrouti et al., 2021; Saakyan et al., 2021) and environment (Diggelmann et al., 2020). Examples of the NLI datasets include the Stanford Natural Language Inference (SNLI) dataset (Bowman et al., 2015a) and the Allen AI's SciTail dataset (Khot et al., 2018). SNLI contains about 550,000 premise-hypothesis pairs. The premises were derived from image captions and hypotheses were created by crowdworkers. SNLI was the first NLI corpus to see encouraging results from neural networks. The SciTail dataset contains 27,000 premise-hypothesis pairs created from multiple-choice science exams and web sentences. Examples of the claim-evidence representations include the SciFact dataset (Wadden et al., 2020) and its extension – SciFact-Open (Wadden et al., 2022). SciFact contained about 1.4K scientific claims and a search corpus of about 5K abstracts that provided either supporting or refuting evidence for each claim. The claims in SciFact-open were extracted from the citation context of papers in biomedical sciences including 279 claims verified against a search corpus of 500K abstracts.

QUESTION: IS THERE AN ASSOCIATION BETWEEN SOCIAL MEDIA USE AND BAD MENTAL HEALTH OUTCOMES?

1. STUDIES INDICATING AN ASSOCIATION
   1. Study 1
   2. Study 2
2. STUDIES INDICATING LITTLE OR NO ASSOCIATION
   1. Study 3
   2. Study 4
3. STUDIES SHOWING MIXED EVIDENCE
   1. Study 5
   2. Study 6

Figure 1: Exemplar collaborative review document structure for one question.

LLMs are trained on large datasets sourced from the internet representing a wide spectrum of both general and domain knowledge. They have shown remarkable performance across a range of NLU tasks such as reading comprehension, and question answering (Liang et al., 2022). The SHE task offers a distinctive opportunity to assess these models in the context of scientific research domain expertise, thereby enabling reasoning abilities compared to those of human experts.

The SHE problem formulation is distinct from the hypothesis and premise pairs encountered in conventional NLI tasks (Bowman et al., 2015a). The language used in scientific publications contains domain-specific terminology which is different from the the premise-hypothesis pairs in general scientific domains (e.g., SNLI (Bowman et al., 2015a)). Furthermore, abstracts from scientific articles contain numerical data that is often not present in traditional NLP datasets.

3. Problem Definition

Scientific hypothesis evidencing (SHE) 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*. The complexity of the task arises from contextual reasoning. For example, in Table 1, identifying the relationship between the abstract and hypothesis provided requires the model to reason that *depressive disorder* is a *bad mental health outcome*, *usage of Twitter, Facebook, Instagram, Snapchat* is *use of social media* leading to *higher likelihood of major depressive disorder* and hence the relationship is *entailment*. In SHE, the hypotheses or research questions are typically expressed at a higher level of abstraction than the evidence provided within the abstract. In this work, we assume that the hypothesis in each hypothesis-abstract pair is addressed by the paper in question and focus on identification of relations between hypotheses and abstracts. Identifying evidence about an arbitrary hypothesis from a literature database is a bigger challenge, usually involving an information retrieval component. This would require a larger corpus of labeled documents as ground truth (Wadden et al., 2022; Pradeep et al., 2021).

4. Dataset

Our Collaborative Reviews (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 (Haidt and Twenge, 2019, 2023; Haidt and Bail, 2022; Haidt and Zach, Ongoing(a); Haidt et al., Ongoing(b); Haidt and Zach, Ongoing(b),O; Haidt et al., Ongoing(a), 2023b,a; Haidt and Zach, Ongoing(c), 2019, 2023). The majority of these reviews were started in 2019 to map important studies within social and behavioral sciences and were maintained using Google docs. These documents are openly available for public viewing and academic researchers in relevant domains can request edit access to make changes. Each review categorizes articles based on a set of research questions related to the topic and the outcomes of each study. Any discrepancies in the classification are resolved by the lead authors alongside a domain expert[1].

Figure 1 gives a schematic illustration of a block of reviews about social media and mental health (Haidt and Twenge, 2023). Most articles are peer-reviewed scientific papers, but several reviews also contain blog posts, news articles, books, and other reports. Our CoRe dataset includes only scientific publications.

Raw data were compiled using all reviews available on July 1, 2023. Research questions, labels, and Digital Object Identifiers (DOIs) were extracted from the reviews through automatic parsing of the document text. DOIs not readily available within the reviews were manually extracted from the publication links provided in the review. We then queried Semantic Scholar (S2) using DOIs to collect article titles and abstracts. In cases where S2 did not have coverage for certain articles, we queried CrossRef (CR). For articles outside the coverage of both S2 and CR, titles and abstracts were collected manually from the publication webpage. Research questions were converted to declarative statements in order to match the structure of hypotheses in the NLI task. Study outputs were manually mapped into one of three classes: *entailment*; *contradiction*; or, *inconclusive*. The curated dataset contains *(hypothesis, abstract, label)* triplets where the *abstract* contains the evidence required to test the *hypothesis* and predict the *label*.

Table 2 provides a list of topics covered by the 12 collaborative reviews and an overview of key statistics. The dataset contains 69 distinct hypotheses tested across 602 scientific articles and findings aligned to our 3 labels. In total, the dataset contains 638 triplets because a fraction of articles address multiple hypotheses. The *entailment* class has greatest representation within the dataset; 61.6% of triplets represent articles that contain evidence in support of the corresponding hypothesis. The *contradiction* class makes up 25.7% of the triplets. The remaining 12.7% are in the *inconclusive* class with a mixed evidence. The distribution of articles across topics is imbalanced, with some reviews containing substantially more literature than others.

Table 3 presents a comparison between the CoRe dataset and the SNLI, and SciFact datasets. Notably, as the CoRe, SciFact datasets use abstracts of scientific publications as premises, their lengths are longer compared to that in SNLI dataset. The mean hypotheses, premise lengths in CoRe dataset are similar to SciFact dataset however, their domains are different. For training and evaluating the models, we shuffled the dataset and split it into to training (70%), development (15%), and held-out test (15%) datasets.

5. Methods

We evaluate two families of methods on the SHE task using the CoRe dataset: transfer learning models and zero- and few-shot LLMs. In the case of transfer learning models, we evaluate sentence pair classifiers based on pre-trained embeddings and Natural Language Inference models.

[1] Further detail about these reviews can be found at https://jonathanhaidt.com/reviews/

Topic	Hyp.	Art.	Tri.	Ent.	Cont.	Inc.	Train	Dev.	Test
Adolescent mood disorders	4	34	37	36	0	1	27	10	0
Adolescent mental illness crisis	8	40	40	35	0	5	25	8	7
Changes in cognitive ability	1	12	13	11	0	2	10	3	0
Digital gambling and mental health	1	3	3	0	3	0	2	1	0
Free play and mental health	5	36	37	23	13	1	17	8	12
Online communities and adolescent health	2	3	3	1	0	2	1	1	1
Phone free schools	5	37	38	8	26	4	21	8	9
Porn use and adolescent health	6	47	47	14	24	9	30	10	7
Social media and mental health	14	222	232	178	48	6	142	38	52
Social media and political dysfunction	9	144	152	67	40	45	82	35	35
Video game use and adolescent health	9	30	32	18	9	5	23	4	5
Gen Z Phone-Based Childhood	2	3	3	2	0	1	2	1	0
Total	69	602	637	393	163	81			
Train	59	370	382	243	92	47			
Dev.	46	127	127	79	28	20			
Test	35	126	128	71	43	14			

Table 2: Statistical overview of the CoRe dataset showing number of hypotheses, articles, triplets along with the distribution of labels across various topics within the dataset. Columns *Train, Dev., Test* correspond to the number of triplets within each respective split.
Hyp.=Hypotheses; Art.=Articles; Tri.=Triplets; Ent.=Entail; Cont.=Contradict; Inc.=Inconclusive; Dev.=Development

Dataset	Hyp.	Pre.	Size	Domain
CoRe	10	194	637	Social Sciences
SciFact	12	194	1,409	Medicine/Biology
SNLI	7	12	570,152	Non Scientific

Table 3: Comparison of average number of words in hypotheses, premises, instance counts, and domains.*Hyp.=Hypotheses; Pre.=Premises*

5.1. Sentence pair classification based on pre-trained embeddings

To investigate embedding models' performance on SHE, we adopt the sentence pair classification framework outlined in (Bowman et al., 2015b). 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 (see Figure 2).

We evaluate the performance of two pre-trained embedding models: *longformer* (Beltagy et al., 2020) and *text-embedding-ada-002* (Greene et al., 2022). Longformer is a transformer-based text encoder model developed to process information at the document-level, therefore eliminating the need for chunking long input text sequences. It uses a combination of local windowed attention and global attention to create a sparse attention matrix (vs. a full attention matrix) making attention more efficient. Longformer supports sequences of length up to 4,096 and produces embeddings of size 768.

Text-embedding-ada-002 is a transformer decoder language model developed by OpenAI and at the time of its release (December 2022) was shown to achieve state-of-the-art performance on tasks such as text search and sentence similarity (Greene et al., 2022). It is capable of embedding sequences of length up to 8,192 and generates 1,536-dimensional embedding vectors.

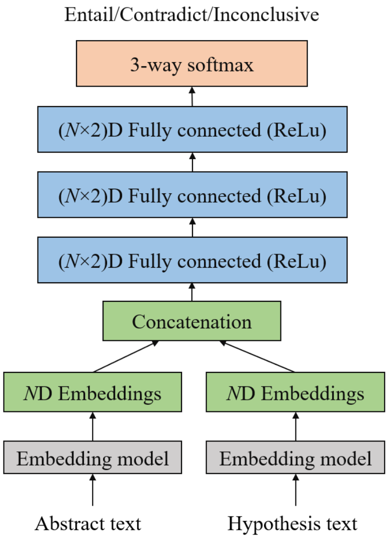

Entail/Contradict/Inconclusive
3-way softmax
(N×2)D Fully connected (ReLu)
(N×2)D Fully connected (ReLu)
(N×2)D Fully connected (ReLu)
Concatenation
ND Embeddings
ND Embeddings
Embedding model
Embedding model
Abstract text
Hypothesis text

Figure 2: Sentence pair classification based on pre-trained embeddings for concatenated hypothesis-abstract pairs

5.2. Transfer learning using Natural Language Inference models

In this approach, we treat SHE task as an NLI task. Specifically, we use an abstract as the premise

CoRe	SNLI	P1	P2, P3, P5	P4
Entail	Entail	true	Yes	e
Contradict	Contradict	false	No	c
Inconclusive	Neutral	neutral	Maybe	n

Table 4: Label map across datasets and prompts.

and determine whether it entails a given hypothesis. Among models proposed for the NLI task, we evaluate the Enhanced Sequential Inference Model (ESIM) (Chen et al., 2017) and Multi-Task Deep Neural Network (MT-DNN) (Liu et al., 2019).

ESIM is a supervised learning model that uses bidirectional Long Short Term Memory (biLSTM) layers to encode hypothesis and premise for inference (Chen et al., 2017). It has achieved high performance on NLI tasks and a reported accuracy of 88.6% on the SNLI dataset. The model uses a 840B token version of GloVe embeddings (Pennington et al., 2014) for word representations.

MT-DNN is a model aiming at learning robust representations across NLU tasks, such as text summarization, NLI and question answering (Liu et al., 2019). MT-DNN achieved the new state-of-the-art performance on the SNLI and the SciTail datasets. We use the MT-DNN model built on the pre-trained *bert-base-uncased* model (Devlin et al., 2018) and fine-tuned it over 5 epochs for the task.

5.3. Large language models

We tested two LLMs, ChatGPT and PaLM 2 (Anil et al., 2023), on our test split. For ChatGPT model, we used the API version of *gpt-3.5-turbo* which offers a faster and significantly less expensive model than OpenAI's other GPT-3.5, GPT-4 models. From PaLM 2, we used the generative model *text-bison-001* (Vertex AI, 2023a) an LLM fine-tuned to follow natural language instructions on a variety of language tasks, e.g., information extraction, problem solving, text edition, and data extraction (Vertex AI, 2023b). We explored these models' performance in zero-shot and few-shot settings. Models were prompted with the abstract and the hypothesis embedded into predefined templates. Prompts contained specific instructions to generate a single output label.

Prompt engineering

Prompt engineering refers to the task of finding the best prompt for an LLM in support of given task (Liu et al., 2023b). We experiment with five prompts used in prior work. All are *prefix* prompts, i.e., prompt text comes entirely before model-generated text. Prompt templates and their sources are summarized in Table 5. Depending on the prompt template, we requested LLMs return one of three sets of labels: *(true, false, neutral)*; *(yes, no, maybe)*; *(entail, contradict, neutral)*. Table 4 maps each label to our canonicalized label set. Because prompts were queried without providing any training data, we refer this method zero-shot learning.

Prompt ensembling

In the context of LLMs, prompt ensembling refers to using several individual prompts at inference (Liu et al., 2023b). Ensembling has shown better performance than fine-tuned models by harnessing their complementary strengths (Li et al., 2023). Here, we use a majority voting strategy to ensemble the outputs of our five individual prompts.

Few Shot Learning

In the few-shot learning (FSL) setting, LLMs are provided with several examples that demonstrate how the model should respond to the prompt (Brown et al., 2020). Studies show that the examples chosen for FSL have a significant impact on the performance of LLMs (Kang et al., 2023). We used a semantic search method to select nine samples from the training dataset to provide examples for each hypothesis-abstract pair in the held-out dataset.[2]

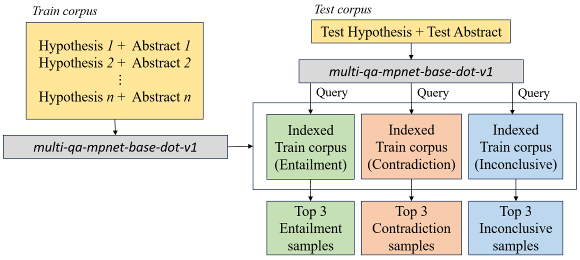

Train corpus
Hypothesis 1 + Abstract 1
Hypothesis 2 + Abstract 2
⋮
Hypothesis n + Abstract n
multi-qa-mpnet-base-dot-v1
Test corpus
Test Hypothesis + Test Abstract
multi-qa-mpnet-base-dot-v1
Query
Query
Query
Indexed Train corpus (Entailment)
Indexed Train corpus (Contradiction)
Indexed Train corpus (Inconclusive)
Top 3 Entailment samples
Top 3 Contradiction samples
Top 3 Inconclusive samples

Figure 3: Semantic search-based sample selection for few-shot learning.

To do so, we incorporated a pre-trained transformer encoder model, specifically the Huggingface implementation of *multi-qa-mpnet-base-dot-v1*[3] which was designed for semantic search. As shown in Figure 3, we first split the training set into three subsets each having a different label. For each instance in the test dataset, we calculate cosine similarity between this instance against each concatenated hypothesis-abstract vectors in the training set and selected the top three pairs in each subset. This results in 9 hypothesis-abstract pairs used for FSL. For each instance in the held-out set, we calculated cosine similarity between the con-

[2] Number of training samples was constrained by the LLM prompt length limitations.

[3] https://huggingface.co/sentence-transformers/multi-qa-mpnet-base-dot-v1

Id	Source	Template
P1	(Basmov et al., 2023)	You are given a pair of texts. Say about this pair: given Text 1, is Text 2 true, false or neutral (you can't tell if it's true or false)? Reply in one word. Text 1: Abstract Text 2: Hypothesis
P2	(Luo et al., 2023)*	Decide if the following summary is consistent with the corresponding article. Note that consistency means all information in the summary is supported by the article. Article: Abstract Summary: Hypothesis Answer (yes or no or maybe):
P3	(Cheng et al., 2023)	Here is a premise: "Abstract." Here is a hypothesis: "Hypothesis." Is it possible to conclude that if the premise is true, then so is the hypothesis? Yes, No, or Maybe?
P4	(Liu et al., 2023a)	Instructions: You will be presented with a premise and a hypothesis about that premise. You need to decide whether the hypothesis is entailed by the premise by choosing one of the following answers: 'e': The hypothesis follows logically from the information contained in the premise. 'c': The hypothesis is logically false from the information contained in the premise. 'n': It is not possible to determine whether the hypothesis is true or false without further information. Read the passage of information thoroughly and select the correct answer from the three answer labels. Read the premise thoroughly to ensure you know what the premise entails. Premise: Abstract Hypothesis: Hypothesis
P5	(Sun et al., 2023)*	The task is to identify whether the premise entails the hypothesis. Please respond with "yes" or "no" or "maybe" Premise: Abstract Hypothesis: Hypothesis Answer:

* Added a third label *maybe*

Table 5: Overview of the different prompts used for testing LLMs and their sources. Since prompts P2, P5 have only two labels *yes, no*, a third label *maybe* was added.

catenated hypothesis-abstract pair and examples in the training corpora.

6. Experiments

We evaluate each model's ability to discern the relationship between a given abstract and a hypothesis, written as a declarative statement in the CoRe dataset. Our approach aligns with methodologies widely used in the literature to evaluate performance on NLI datasets, e.g., SNLI (Bowman et al., 2015b), MNLI (Williams et al., 2017), where models are presented with a premise and a declarative hypothesis asked to classify their relationship. Performance of all models is measured by macro-F1-score, calculated as the average of F1-scores over all three class labels, and accuracy is calculated as the fraction of correct predictions.

For sentence pair classification based on embeddings, all layers utilize the ReLU activation function. Hyperparameters such as learning rate and regularization parameter, were set based on Bayesian hyperparmeter tuning with an objective to maximize macro-F1-score on the test data (Akiba et al., 2019). For training ESIM and fine-tuning MT-DNN on the SNLI dataset, we adopted default hyperparameters, as recommended in respective papers.

We evaluate LLMs in zero-shot and few-shot settings. The temperature parameter controls the creativity of the text generated by the LLMs. Lower temperatures result in more consistent output and higher temperatures result in more creative, diverse responses. We compare the performance of LLMs with temperature settings from 0 to 1, by increments of 0.25. We query each LLM using the same prompt 5 times in each temperature setting.

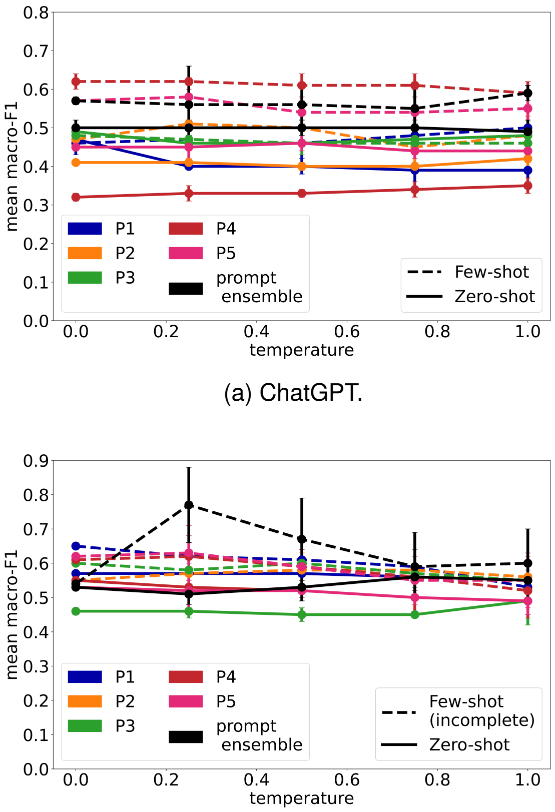
To test prompt ensembling, we query each LLM using the set of five prompts and determine the final classification label by majority voting aggregation. Prompt ensembling was tested in both zero-shot and few-shot settings across different temperatures. Similarly to the single prompt evaluation, we tested the prompt ensembling under five independent runs for each temperature configuration.

7. Results

Table 6 summarizes model performance on the test set. Here, we focus on comparing different types of models, so we report metrics averaged across all settings. The observation that all models achieve macro-F1-scores less than 0.65 demonstrates that SHE is a challenging task. **The sentence pair classifier 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.

7.1. Natural Language Inference models

As anticipated, ESIM and MT-DNN models when trained or fine-tuned on the SNLI dataset respectively, exhibited significantly lower performance compared to the model when trained or fine-tuned on the CoRe dataset. This can be attributed to the differences in the characteristics of hypotheses and premises in the two datasets. For instance in CoRe, each hypothesis has an average length of 10 words compared to 7 in case of SNLI. Average context length is 194 in CoRe compared to a 13-word premise in SNLI. Furthermore, levels of abstraction of hypotheses within the datasets vary. In CoRe, the evidence required to identify the abstract-hypothesis relationship is latent within the premise. Additionally, around 2,500 words from the CoRe dataset vocabulary are missing from the SNLI dataset. This underscores the need for domain specific datasets for fine-tuning.



(a) ChatGPT.

(b) PaLM 2.

Figure 4: Average macro-F1 of LLMs with different prompt templates and temperature settings.

7.2. Performance of LLMs

Figure 4 summarizes the results for LLMs tested under different settings. Although not directly trained on the CoRe dataset, LLMs were able to comprehend the evidence within scientific abstracts and relate them to hypotheses. This can be attributed to their large-scale training. In the zero-shot setting, LLMs generally achieved macro-F1 of about0.5, which is comparable with transfer learning models fine-tuned on the CoRe dataset. Additionally, in the zero-shot setting across all prompt styles, PaLM 2 consistently outperformed ChatGPT.

We observed that when conducting FSL with PaLM 2, the model may output *null* output and this occurrences are unpredictable at different temperatures and iterations. This occured over different temperatures and different iterations. The number of such instances ranged from 48 (37.5%) to 64 (50%) for an evaluation under a certain setting; overall, 20 (15.6%) of responses yielded *null* outputs consistently across all prompt styles. We will investigate this anomaly in the future.

While variations in performance were observed across individual runs, for a given prompt, the temperature setting did not have a major influence on the average performance of LLMs. This observation remained consistent across the zero-shot, few-shot, and prompt ensembling settings.

Effect of prompt template

As expected, the choice of prompt style had clear influence on LLM performance. Different performance metrics for models tested with different prompts are summarized in Table 7. In the zero-shot setting, ChatGPT recorded the lowest mean macro-f1-score of 0.337 when prompted with P4 whereas it achieved highest mean macro-f1-score of 0.479 when using prompt P3. A similar trend is observed in the few-shot setting where the lowest performance was recorded when using prompt P3 with a mean macro-f1-score of 0.466 while the highest performance was reported for P4 with a mean macro-f1-score of 0.609. There was no single prompt that had consistent high performance across all temperatures, models, and settings. This indicates that ensembling over multiple prompt templates is more reliable than using a single prompt.

Unequal performance gains with FSL

FSL in general enhances the performance of LLMs across prompt styles, although performance gains are unequal. In case of ChatGPT, prompt P4 showed the greatest improvement with mean macro-f1-score increasing from 0.337 in zero-shot to 0.609 in the few-shot setting. Conversely, the performance slightly declined with P3, from 0.479 in zero-shot to 0.466 in the few-shot setting. Prompt ensembling with ChatGPT in the zero-shot setting achieves performance comparable to the few-shot setting.

8. Conclusion

We have introduced the Scientific Hypothesis Evidencing task and a novel dataset for this task. Our goal is to determine whether a paper, based on its

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 Fine-tuned on SNLI	67.97% 42.97%	0.523 0.342
	ESIM	Supervised on CoRe Supervised on SNLI	64.84% 39.84%	0.489 0.335
LLM	ChatGPT	Zero-shot w/o ensemble Few-shot w/o ensemble Zero-shot with ensemble Few-shot with ensemble	47.22%* 59.85%* 53.94% 66.57%	0.414* 0.517* 0.500 0.576
	PaLM 2	zero-shot w/o ensemble Few-shot w/o ensemble Zero-shot with ensemble Few-shot with ensemble	59.78%* 69.78%*† 62.87% 76.40%	0.504* 0.583*† 0.536 0.678*†

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

† Incomplete responses

Table 6: Results summarizing the performance of models on the held-out set under different settings.

Model	Prompt	F1	P	R
PaLM 2 Zero-shot	*P1*	$0.56_{(0.01)}$	$0.60_{(0.01)}$	$0.59_{(0.01)}$
	P2	$0.46_{(0.02)}$	$0.50_{(0.03)}$	$0.51_{(0.01)}$
	P3	$0.46_{(0.02)}$	$0.50_{(0.03)}$	$0.51_{(0.01)}$
	P4	$0.52_{(0.02)}$	$0.60_{(0.06)}$	$0.59_{(0.04)}$
	P5	$0.51_{(0.01)}$	$0.53_{(0.01)}$	$0.53_{(0.01)}$
PaLM 2 Few-shot*	*P1*	$0.60_{(0.04)}$	$0.65_{(0.03)}$	$0.68_{(0.06)}$
	P2	$0.57_{(0.01)}$	$0.58_{(0.04)}$	$0.58_{(0.01)}$
	P3	$0.58_{(0.02)}$	$0.61_{(0.02)}$	$0.64_{(0.03)}$
	P4	$0.58_{(0.04)}$	$0.58_{(0.04)}$	$0.59_{(0.04)}$
	P5	$0.59_{(0.04)}$	$0.63_{(0.07)}$	$0.59_{(0.03)}$
ChatGPT Zero-shot	*P1*	$0.41_{(0.03)}$	$0.54_{(0.032)}$	$0.50_{(0.02)}$
	P2	$0.41_{(0.01)}$	$0.44_{(0.03)}$	$0.47_{(0.00)}$
	P3	$0.47_{(0.01)}$	$0.58_{(0.02)}$	$0.53_{(0.02)}$
	P4	$0.34_{(0.01)}$	$0.61_{(0.03)}$	$0.44_{(0.01)}$
	P5	$0.45_{(0.01)}$	$0.51_{(0.01)}$	$0.50_{(0.01)}$
ChatGPT Few-shot	*P1*	$0.47_{(0.02)}$	$0.47_{(0.03)}$	$0.50_{(0.01)}$
	P2	$0.49_{(0.02)}$	$0.52_{(0.02)}$	$0.53_{(0.03)}$
	P3	$0.4_{(0.01)}$	$0.56_{(0.013)}$	$0.53_{(0.01)}$
	P4	$0.61_{(0.01)}$	$0.63_{(0.01)}$	$0.64_{(0.01)}$
	P5	$0.56_{(0.03)}$	$0.57_{(0.02)}$	$0.58_{(0.02)}$

* Partial results due to *null* responses

Table 7: Comparison of performance metrics of LLMs across various prompt templates on the CoRe dataset. The metrics are averaged for different temperature settings across all the runs. Subscripts indicate standard deviation over 5 runs. *Note: The macro averaged precision and recall metrics are skewed due to class imbalance.*

abstract, offers evidence in support or refute of a given hypothesis. This goal broadly underlies all of meta-analysis. It supports efforts to highlight inconsistencies and gaps in existing literature, motivate next studies, and support evidence-based decision-making and policy. Given the wide availability of abstracts, e.g., in scholarly data repositories, methods which can successfully operate on abstracts as opposed to full text are preferable.

Our CoRe dataset exhibits imbalance in class distribution and topical coverage. Imbalance in label classes is likely to be pervasive given publication bias where positive outcomes are more likely to be published than negative or inconclusive results. This imbalance is likely to affect the performance of models requiring training or fine-tuning, e.g., sentence pair classification, ESIM, and MT-DNN models. However, the impact on the performance of LLMs and pre-trained models, which are used solely for inference, is expected to be lesser.

Our findings suggest that hypothesis evidencing is challenging task for current NLU models, including state-of-the-art LLMs trained on a diverse set of data. Notably, sentence pair classification using embedding-based models outperforms LLMs in our experiments. Yet, these models are known to be less generalizable than LLMs.

Looking ahead, Open AI has recently introduced fine tuning functionality for gpt-3.5-turbo. Future work should investigate the performance of LLMs following fine tuning on the data. This will indicate whether the higher performance of embedding-based models is result of exposure to the complete training dataset vs. fewer examples provided in FSL setting. In addition, for a more robust assessment, future work should explore five-fold cross-validation. And, given the limitations associated with human-generated discrete prompts, future work should explore automatic prompt tuning.

We expect that, in providing the research community with an initial benchmark dataset, our work will catalyze some of these next steps. Future work should continue to build out datasets like ours, keeping in mind ways to ameliorate class imbalance and diversify represented topics. Furthermore, future work should prioritize moving beyond binary or ternary labels towards appropriately capturing nuances in hypothesis-evidence relationships. Notably, the assembly of these datasets serves multiple aims, bringing together domain experts around important questions in their own area, highlighting robustness and reliability amongst claims, and of course moving forward NLP and NLU.

Ethics Statement

The data we compiled were originally contributed by volunteering social scientists. Although we did not seek their consensus, we cite each review as an individual reference and acknowledge all contributors for their efforts on this open list, which not only benefits social scientists but also computer and information scientists. In addition, because the edit request needed to be approved, we assumed the contributors were all qualified scientists or researchers, so the quality of the data was ensured.

9. Bibliographical References

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2623–2631.

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.

Victoria Basmov, Yoav Goldberg, and Reut Tsarfaty. 2023. Chatgpt and simple linguistic inferences: Blind spots and blinds. *arXiv preprint arXiv:2305.14785*.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015a. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015b. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv: 1508.05326*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2017. 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.

Daixuan Cheng, Shaohan Huang, Junyu Bi, Yuefeng Zhan, Jianfeng Liu, Yujing Wang, Hao Sun, Furu Wei, Denvy Deng, and Qi Zhang. 2023. Uprise: Universal prompt retrieval for improving zero-shot evaluation. *arXiv preprint arXiv:2303.08518*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Thomas Diggelmann, Jordan Boyd-Graber, Jannis Bulian, Massimiliano Ciaramita, and Markus Leippold. 2020. Climate-fever: A dataset for verification of real-world climate claims. *arXiv preprint arXiv:2012.00614*.

Ryan Greene, Ted Sanders, Lilian Weng, and Arvind Neelakantan. 2022. New and improved embedding model.

Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10:178–206.

Jonathan Haidt and Christopher Bail. 2022. Social media and political dysfunction: A collaborative review. *Unpublished manuscript, New York University*.

Jonathan Haidt, George Eli, and Rausch Zach. 2023a. Changes in cognitive ability among children and teens: A review. *Unpublished manuscript, New York University*.

Jonathan Haidt, George Eli, and Rausch Zach. 2023b. Online communities and adolescent health: A collaborative review. *Unpublished manuscript, New York University*.

Jonathan Haidt, George Eli, and Rausch Zach. Ongoing(a). Gen z on social media and the effects of a phone-based childhood: A collaborative review. *Unpublished manuscript, New York University*.

Jonathan Haidt, Yejin Park, and Bentov Yaneev. Ongoing(b). Free play and mental health: A collaborative review. *Unpublished manuscript, New York University*.

Jonathan Haidt and Jean Twenge. 2019. Is there an increase in adolescent mood disorders, self-harm, and suicide since 2010 in the usa and uk? a review. *Unpublished manuscript, New York University, New York City, New York*.

Jonathan Haidt and Jean Twenge. 2023. Social media and mental health: A collaborative review. *Unpublished manuscript, New York University. Retrieved from: tinyurl.com/SocialMediaMentalHealthReview*.

Jonathan Haidt and Rausch Zach. 2019. Adolescent mood disorders since 2010: A collaborative review. *Unpublished manuscript, New York University*.

Jonathan Haidt and Rausch Zach. 2023. The effects of phone-free schools: A collaborative review. *Unpublished manuscript, New York University*.

Jonathan Haidt and Rausch Zach. Ongoing(a). Alternative hypotheses to the adolescent mental illness crisis: A collaborative review. *Unpublished manuscript, New York University*.

Jonathan Haidt and Rausch Zach. Ongoing(b). Digital gambling and adolescent health: A collaborative review. *Unpublished manuscript, New York University*.

Jonathan Haidt and Rausch Zach. Ongoing(c). Porn use and adolescent health: A collaborative review. *Unpublished manuscript, New York University*.

Jonathan Haidt and Rausch Zach. Ongoing(d). Video game use and adolescent mental health: A collaborative review. *Unpublished manuscript, New York University*.

Sungmin Kang, Juyeon Yoon, and Shin Yoo. 2023. Large language models are few-shot testers: Exploring llm-based general bug reproduction. In *2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE)*, pages 2312–2323. IEEE.

Tushar Khot, Ashish Sabharwal, and Peter Clark. 2018. Scitail: A textual entailment dataset from science question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Neema Kotonya and Francesca Toni. 2020. Explainable automated fact-checking for public health claims. *arXiv preprint arXiv:2010.09926*.

Yifei Li, Zeqi Lin, Shizhuo Zhang, Qiang Fu, Bei Chen, Jian-Guang Lou, and Weizhu Chen. 2023. Making language models better reasoners with step-aware verifier. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5315–5333.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.

Hanmeng Liu, Ruoxi Ning, Zhiyang Teng, Jian Liu, Qiji Zhou, and Yue Zhang. 2023a. Evaluating the logical reasoning ability of chatgpt and gpt-4. *arXiv preprint arXiv:2304.03439*.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023b. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019. Multi-task deep neural networks for natural language understanding.

Zheheng Luo, Qianqian Xie, and Sophia Ananiadou. 2023. Chatgpt as a factual inconsistency evaluator for text summarization.

Ray Oshikawa, Jing Qian, and William Yang Wang. 2020. A survey on natural language processing for fake news detection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6086–6093, Marseille, France. European Language Resources Association.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Ronak Pradeep, Xueguang Ma, Rodrigo Nogueira, and Jimmy Lin. 2021. Scientific claim verification with VerT5erini. In *Proceedings of the 12th International Workshop on Health Text Mining* *and Information Analysis*, pages 94–103, online. Association for Computational Linguistics.

Arkadiy Saakyan, Tuhin Chakrabarty, and Smaranda Muresan. 2021. Covid-fact: Fact extraction and verification of real-world claims on covid-19 pandemic. *arXiv preprint arXiv:2106.03794*.

Mourad Sarrouti, Asma Ben Abacha, Yassine M'rabet, and Dina Demner-Fushman. 2021. Evidence-based fact-checking of health-related claims. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3499–3512.

Xiaofei Sun, Linfeng Dong, Xiaoya Li, Zhen Wan, Shuhe Wang, Tianwei Zhang, Jiwei Li, Fei Cheng, Lingjuan Lyu, Fei Wu, et al. 2023. Pushing the limits of chatgpt on nlp tasks. *arXiv preprint arXiv:2306.09719*.

Vertex AI. 2023a. Palm 2. https://cloud.google.com/vertex-ai/docs/generative-ai/model-reference/text#model_versions.

Vertex AI. 2023b. Palm api. https://developers.generativeai.google/models/language.

Juraj Vladika and Florian Matthes. 2023. Scientific fact-checking: A survey of resources and approaches. *arXiv preprint arXiv:2305.16859*.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550, Online. Association for Computational Linguistics.

David Wadden, Kyle Lo, Bailey Kuehl, Arman Cohan, Iz Beltagy, Lucy Lu Wang, and Hannaneh Hajishirzi. 2022. Scifact-open: Towards open-domain scientific claim verification. *arXiv preprint arXiv:2210.13777*.

Abraham Wald. 1992. Sequential tests of statistical hypotheses. In *Breakthroughs in statistics: Foundations and basic theory*, pages 256–298. Springer.

Adina Williams, Nikita Nangia, and Samuel R Bowman. 2017. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*.