# From Philosophy to NLU: Evolving Definitions of Research Hypotheses

Jian Wu[1,*], Sarah Rajtmajer[2,*]

[1] *Old Dominion University*
[2] *The Pennsylvania State University*

### Abstract

Over the past decades, alongside advancements in natural language processing, significant attention has been paid to training models to automatically extract, understand, test, and generate hypotheses in open and scientific domains. However, interpretations of the term *hypothesis* for various natural language understanding (NLU) tasks have migrated from traditional definitions in the natural, social, and formal sciences. Even within NLU, we observe differences defining hypotheses across literature. In this paper, we overview and delineate various definitions of hypothesis. Especially, we discern the nuances of definitions across recently published NLU tasks. We highlight the importance of well-structured and well-defined hypotheses, particularly as we move toward a machine-interpretable scholarly record.

### Keywords

natural language processing, natural language understanding, natural language inference, hypothesis extraction, hypothesis generation, scientific claim verification, large language models

## 1. Introduction

The word “hypothesis” has been used variably with different meanings, over decades and centuries, across the social, natural, and formal sciences—from its conceptual roots in ancient Greek philosophy [1, 2, 3] to the development of hypothesis testing as statistical methods [4, 5] and subsequent evolution of the term in different fields reflecting their unique questions and approaches [6]. Of course, ambiguity around language and variable use of terminology is pervasive within and outside of science. Language has always been an impoverished tool for representation and expression of complex ideas [7, 8, 9]. In many cases though, this is not a problem. Members of a particular community develop shared understanding of the meaning of a given term in context, and this allows them to communicate effectively toward collective goals.

We argue that ambiguity and variability around the definition of hypotheses, which was once acceptable—even perhaps productive—is now a critical concern in light of natural language processing (NLP) and natural language understanding (NLU) tasks requiring quantitative operationalization of hypotheses, in particular, hypothesis extraction (detection/identification), verification, and generation. Emerging technical work in these fields often do not include explicit definitions of hypotheses, claims, or evidence, instead relying *de facto* on benchmark datasets to provide implicit definitions.

Following, we survey the definitions and operationalizations of hypotheses, focusing on research hypotheses engaged in the hypothesis mining literature. Research hypotheses are hypotheses designed for systematic investigation within a research framework. In this paper, we do not distinguish between research hypothesis and scientific hypothesis. In principle, these two terms have different scopes, but in practice, they are often used interchangeably in modern hypothesis mining papers. The related tasks are particularly in the areas of natural language inference (NLI), hypothesis extraction, scientific hypothesis evidencing and scientific claim verification, and scientific hypothesis generation. We highlight important

*Sci-K'25: International Workshop on Scientific Knowledge: Representation, Discovery, and Assessment, November 2-6, 2025, Nara, Japan*

[*] The two authors made equal contributions to this paper.

j1wu@odu.edu (J. Wu); smr48@psu.edu (S. Rajtmajer)
https://www.cs.odu.edu/~jwu/ (J. Wu); https://www.rajtmajerlab.net/ (S. Rajtmajer)
0000-0003-0173-4463 (J. Wu); 0000-0002-1464-0848 (S. Rajtmajer)

differences and discuss the challenges these differences impose on knowledge assembly and aggregation. Our work is motivated by the vision of a computable scholarly record—a verifiable and extensible knowledge base synthesizing computationally and data-enabled discoveries [10]. This vision, we suggest, will be enabled by machine-readable hypotheses well-structured in predictable formats.

## 2. Background

### 2.1. Conceptual origins

The word *hypothesis* derives from Greek and means, literally, *a putting under* or supposition. Ancient Greek philosophers used the term to describe a foundational assumption upon which to build out further reasoning. Plato uses the term in several of his dialogues with this intention–namely, as a claim accepted temporarily in order to explore its implications. Of particular interest to Plato was whether a hypothesis could support consistent and coherent conclusions [11, 12]. Aristotle also engaged with the term. Aristotle viewed hypotheses more cautiously, being skeptical of relying on hypotheses without empirical verification. He delineated tentative assumptions from axioms or first principles, insisting that scientific knowledge must be based on demonstrated causes, not just assumed premises [13].

During the scientific revolution of the 16th and 17th centuries, the concept of hypothesis evolved significantly. Galileo and Newton began using hypotheses as formal components of the scientific method, emphasizing the importance of testing through observation and experiment [14, 15]. René Descartes also contributed to this shift, promoting skepticism and the formulation of testable propositions [16, 17].

By the 19th and 20th centuries, the hypothesis had become a central pillar in science. Prominent philosophers of science Karl Popper and Thomas Kuhn centered hypotheses within the scientific process, both agreeing that hypotheses must engage with empirical data in some way, i.e., should be testable, observable [18, 19]. However, Kuhn and Popper's views on scientific advancement differed in important ways and their respective views on the role of hypotheses reflected these differences. Kuhn, a historian, viewed periods of science through the lens of *paradigms* and hypotheses as statements that operates within a paradigm (vs. free-floating assumptions to be directly tested in isolation) [20]. Popper, on the other hand, highlighted the asymmetry between verification and *falsification*—hypotheses cannot be proven true, only proven false. Central to Popper's thinking is his assertion that confirmations for a theory are easy to find if we look for them. Confirming evidence should only count when it is the result of a genuine test of the theory (i.e., we conclude that the theory withstood an attempt to disconfirm it) [3]. In essence, Popper argued that Kuhn's hypotheses risked self-fulfillment. Popperian theories underlie the current open science movement triggered by the replication crisis, including efforts promoting development of strong, *testable* hypotheses, and preregistration to delineate exploratory vs. confirmatory findings [21, 22].

### 2.2. Modern forms

In the last decades, there has been a growing interest in hypothesis mining in scientific literature, mostly in the fields of NLP and NLU, but also in interdisciplinary fields between social science and AI. The exact definitions of hypotheses involved are not always provided in the context of research problems, and the specific forms and expressions vary across papers. Here, we categorize modern hypotheses into several types, which may deviate from traditional definitions, e.g., [23].

**Ideas as hypotheses.** As defined in Kuhn & Hawkins [1], an *idea* is a realization or hypothesis that can challenge and shift paradigms within a scientific community. Several recent papers about hypothesis generation adopted this conceptualization and treated ideas as hypotheses [24, 25]. In Kumar et al. [24], the authors build a dataset containing *Future Research Ideas (FRIs)* and then consider the generated FRIs to be hypotheses. The structure of these ideas includes: premises; a traditional research hypothesis; and its context. In Wang et al. [25], the authors do not discern the term ideas from hypotheses and use them interchangeably in certain contexts, but the ground truth data shows that the

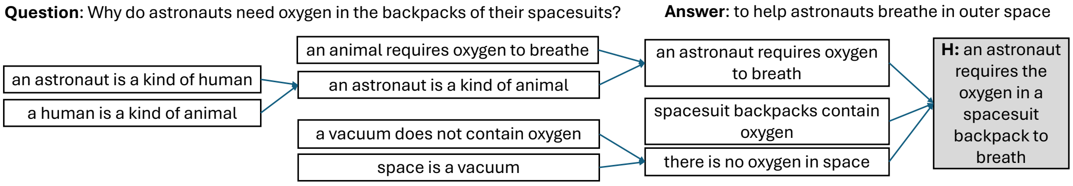
Question: Why do astronauts need oxygen in the backpacks of their spacesuits?
Answer: to help astronauts breathe in outer space
an astronaut is a kind of human
a human is a kind of animal
an animal requires oxygen to breathe
an astronaut is a kind of animal
a vacuum does not contain oxygen
space is a vacuum
an astronaut requires oxygen to breath
spacesuit backpacks contain oxygen
there is no oxygen in space
H: an astronaut requires the oxygen in a spacesuit backpack to breath

**Figure 1:** An example in the EntailmentBank dataset, demonstrating the entailment tree structure to support the hypothesis (in a shaded box) generated based on the question and answer. The figure is adopted from [35].

generated content contains preliminary and broad notations intended to inspire further investigation, which is aligned with the concept of ideas. An example is shown in Table 1.

**Claims as hypotheses.** The classical definition of *claim* is the conclusion or assertion that you want your audience to accept [26]. Adopting this definition, in scientific literature, a claim can be defined as a specific assertion reported as a finding of the paper. A paper can make more than one claim, and a claim may contain one or multiple sentences. One definition of hypothesis is a claim that has not been tested [27]. In Alipourfard et al. [28], authors label a *claim trace* for each paper in their corpora, and each trace contains four claims. Hypotheses and evidence are treated as two types of claims. In the recent SciHyp dataset [29] developed for hypothesis detection and classification in Computer Science papers, many hypotheses in the ground truth are claims manually extracted throughout the full text. This ambiguity also occurs in scientific hypothesis evidencing (SHE; [30]) and scientific claim verification (SCV; [31]) tasks. Both tasks aim to discern the relationship (or stance) between a hypothesis (in SHE, or a claim in SCV) and a candidate piece of evidence.

**Hypothesis-proposals.** In recent works about hypothesis generation, models are built to generate not only a hypothesis but also a series of related sections such as its background, justification, and test procedures, resulting in a *proposal*-style document [32]. The hypothesis-proposal increases the transparency of hypothesis generation and provides a guide for testing. However, the specific format/sections of the proposal differ by model. An example in [32] is shown in Table 1.

**Formal expressions.** In early work, a research hypothesis is broken down into three dimensions, namely contexts, variables, and relationships [16]. Each hypothesis is associated with a target variable and a set of independent variables, and relationships refer to the interactions between a given set of variables under a given context that produces the hypothesis. A hypothesis is then naturally expressed with a *semantic tree* in which the nodes represent variables and the edges represent relationships.

In more recent works in NLU, papers have expressed research hypotheses in various ways, depending on the focal tasks. For example, in hypothesis generation tasks, the generated hypothesis may be composed of multiple declarative statements, in which one serves as the main hypothesis and the others provide additional context or details (see[33] and [25] in Table 1). In the hypothesis evidencing task, hypotheses can be written as questions [30], which can be converted into hypotheses in declarative form. A research hypothesis can also be decomposed as a question and an answer, e.g., the SciTail dataset [34]. Their entailment relation can be further explained using an *entailment tree* showing how the hypothesis follows from the text corpus [35] (Figure 1).

## 3. Scientific hypothesis-related tasks and datasets

### 3.1. Natural language inference

Natural language inference (NLI, i.e., recognizing textual entailment (RTE)) involves assessing whether a given textual premise entails or implies a given hypothesis [36, 37]. Most NLI datasets, such as SNLI [38] and RTE-6 [39], are in open domains [38]. SciTail is one of the few datasets built for scientific NLI [34]. Hypotheses are expressed in single declarative sentences (see Table 1).

Another scientific NLI dataset is EntailmentBank, built for multi-step scientific inference (Figure 1). The task is to generate an entailment tree given a hypothesis. The tree shows a hierarchical supportive

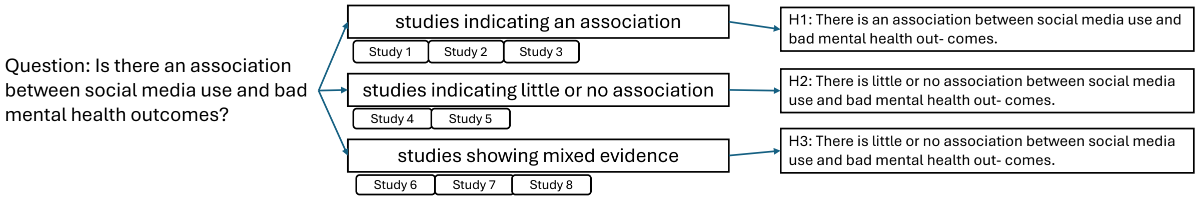
Question: Is there an association between social media use and bad mental health outcomes?
studies indicating an association
Study 1
Study 2
Study 3
studies indicating little or no association
Study 4
Study 5
studies showing mixed evidence
Study 6
Study 7
Study 8
H1: There is an association between social media use and bad mental health out- comes.
H2: There is little or no association between social media use and bad mental health out- comes.
H3: There is little or no association between social media use and bad mental health out- comes.

**Figure 2:** The collaborative review document structure for an example question and the hypotheses derived from the (question, answer) pairs. [30].

structure of claims toward a hypothesis. Other scientific NLI datasets include MediNLI [40] and BioNLI [41] in the medical and biomedical domains, and e-SNLI-VE [38] for visual scientific inference.

## 3.2. Hypothesis and claim extraction

Here, the goal is to automatically identify hypotheses from a scientific document. Because hypotheses can be viewed as claims prior to testing, the input, output, and methods for extracting hypotheses and claims are similar. The input document can be an abstract, e.g., [42, 43] or a full paper [28]. In White et al. [44], authors propose and apply a schema for annotating sentences in full text of scientific articles into 9 types: hypothesis; goal; motivation; background; method; experiment; result; observation; and conclusion [44]. In the dataset used for training DeepCause, a model for hypothesis extraction, selected claims identified from the full text are labeled as hypotheses [23].

Claim extraction may benefit from structured abstracts which contain, e.g., a dedicated *Findings* section (see Lancet [45]). Yet still, the specific sections differ across journals. For example, in [42], *findings* or *proposed items* are labeled as claims and one abstract may contain multiple claims (Table 1).

It is worth noting that papers in computer science and several other domains often claim findings or contributions without explicitly stating hypotheses, e.g., [25]. We consider these findings or contributions *pseudo-hypotheses*.

## 3.3. Scientific hypothesis evidencing and scientific claim verification

Scientific hypothesis evidencing (SHE) [30] is the task of automatically identifying evidence from scientific publications in support of or refutation of a given hypothesis. This task is similar to another task called Scientific claim verification (SCV), both reflecting a model's reasoning capability. The main difference is that in SHE, the hypotheses are usually high-level research questions (Figure 2). In SCV datasets, the hypotheses are usually lower-level claims in specific contexts. However, certain cases are in between. Both tasks can be divided into two subtasks: identifying evidence candidates from a corpus of documents; and discerning the relationship between the hypothesis (claim) and an evidence candidate. Most research focuses on the second subtask, in which the relationships are classified into exclusive categories, namely *SUPPORT*, *REFUTE*, and *NEI* (not enough information) or their variants.

In Koneru et al. [30] authors build a dataset for the task of SHE using community-driven annotations of studies in social sciences. The input is a hypothesis and an abstract (i.e., candidate evidence), and the output is a label indicating whether the abstract entails, contradicts, or is inconclusive to the hypothesis. In Wadden et al. [31], authors build an SCV dataset, the input of which is a (claim, abstract) pair (see an example in Table 1). In the Covid-Fact dataset [46], claims are obtained by filtering titles of social media posts. While, in the HealthVer dataset, claims are manually extracted from questions and snippets returned by search engines [47].

DiscoveryBench [48] is a benchmark designed for a task called *data-driven discovery*. Similar to SCV, the goal is to verify a hypothesis, originally expressed in the form of a research question. Nevertheless, instead of using abstracts as evidence, the verification is grounded in data. In an example task, a model is given a dataset in the form of a spreadsheet and a research question. The model is expected to generate a stepwise workflow that tries to answer the question using the given data. The final output, treated as a hypothesis, is decomposed into a semantic tree containing (context, variable, relationship) and

**Table 1**
Examples of input and output for selected hypothesis-related tasks. Task names are abbreviated. Gen=Hypothesis generation; Ext=Hypothesis (Claim) extraction; SCV=Scientific claim verification; DDD=Data-driven discovery.

| Tasks | Input | Output | Examples |
|---|---|---|---|
| NLI | (Premise, Hypothesis) (SciTail [34]) | Label: Entail, Neural, Contradiction | Input premise: Beats are the periodic and repeating fluctuations heard in the intensity of a sound when two sound waves of very similar frequencies interfere with one another.<br>*Input Hypothesis: When waves of two different frequencies interfere, beating occurs.*<br>*Output label: entail* |
| Ext | Full text or an abstract [42] | Hypotheses or claims | Input: The abstract of a paper titled *A Morphological Hessian Based Approach for Retinal Blood Vessels Segmentation and Denoising Using Region Based Otsu Thresholding* [49]<br>*Output claim: We proposed a less computational unsupervised automated technique with promising results for detection of retinal vasculature by using morphological hessian based approach and region based Otsu thresholding.* |
| SCV | (claim, abstract) (SciFact [31]) | Label: Support, Refute, NEI | Input claim: The coronavirus cannot thrive in warmer climates.<br>Input abstract: ...most outbreaks display a pattern of clustering in relatively cool and dry areas...This is because the environment can mediate human-to-human transmission of SARS-CoV-2, and unsuitable climates can cause the virus to destabilize quickly...<br>*Output label: Support* |
| DDD | Goal and data (DiscoveryBench [48]) | Workflow answer and decomposed hypothesis | Input goal: How did urban land use affect the invasion of different types of introduced plants in Catalonia?<br>Input data: a relational table of habitat type vs. plant type<br>*Output workflow: A workflow to answer the question.*<br>*Output answer (hypothesis): Urban land use increased invasion by agriforest plants over gardening introduced ones in Catalonia.*<br>*Output (context, variable, relationship)=(urban habitat type, gardening, unintentional, reduced)* |
| Gen | Keywords (SciAgents [33]) | Proposal | Input keywords: heat transfer performance, soft lithography, etc.<br>*Output hypothesis: We hypothesize that integrating biomimetic materials with microfluidic chips will significantly enhance their heat transfer performance and biocompatibility, making them ideal for advanced biomedical applications. Specifically, we propose that the lamellar structure of biomaterials, inspired by keratin scales, can be engineered into microfluidic chips using soft lithography techniques to improve their mechanical behavior and heat transfer efficiency under cyclic loading conditions.*<br>*Output other sections: outcome, mechanisms, design principles, unexpected properties, comparison, and novelty, and their expanded versions.* |
| Gen | Data [50] | Hypothesis | Input: Tweet pairs in the Tweet Popularity dataset [51].<br>*Output: Tweets with named entities like people, places, or organizations tend to get more retweets by being more specific.* |
| Gen | Background (SciMon [25]) | Ideas | Input seed term: diverse relational edge embedding<br>Input background: the task of converting a natural lanuage question into an executable sql query , known as text - to -sql, is an important branch of semantic parsing . the state - of - the -art graph -based encoder has been successfully used in this task but does not model the question syntax well.<br>*Output: We propose a novel graph-based encoder that uses a diverse relational edge embeddings to model the question syntax.* |

compared against the gold standard hypothesis (see an example in Table 1). SciClaimHunt is a dataset recently built for SCV, in which a small amount of claims are manually extracted from the discussion and conclusion sections of research papers in computer science. Most claims are generated by LLMs. Of note, a fraction of claims are not self-contained and require reference to the context of the source paper.

### 3.4. Scientific hypothesis generation

The goal of scientific hypothesis generation is to automatically create new, testable scientific hypotheses or research ideas that identify novel relationships, phenomena, or gaps in existing knowledge [24]. Recent advancements in LLMs, e.g., Llama [52] and GPT [53], offer promise. Here too, existing literature is inconsistent with respect to task formulation, i.e., *input* and *output* (see examples in Table 1).

We identify four types of input for the hypothesis generation tasks:
(1) *keywords*, concise descriptions of the topics or central concepts of the model, such as in SciMon [25] and SciAgents [33];
(2) *goals*, brief discourse outlining research goals. In AI co-scientist [32], goals can be a request to propose a novel hypothesis, suggest special requirements, or ask a question. In Si et al. [54], LLMs are provided "topics", such as "novel prompting methods that can better quantify uncertainty or calibrate the confidence of large language models", which serve as goals. In Pu et al. [55], an input is described as an "objective", which is equivalent to a goal;
(3) *data*, i.e., a dataset, based on which the model is requested to generate a hypothesis, e.g., [50];
(4) *background*, context, rationale, or theoretical foundation of a hypothesis. For example, in the SciMon framework [25], the input includes seed terms, including concepts and keywords, and background context, which contains problems, motivations, or focus points. The Mamba framework [56] uses the same ground truth as SciMon [25]. The MOOSE framework [57] uses background and inspiration derived from the raw web corpus as the input.

Likewise, we identify three types of output of hypothesis generation tasks:
(1) *traditional hypothesis*, usually expressed as a single or multiple declarative sentences, e.g., [50, 58].
(2) *ideas*, enriched hypotheses as shown in Section 2.2. For example, in SciMon, generated ideas may contain claims, methods, and objectives extracted from abstracts.
(3) *hypothesis proposals*, comprehensive structured hypothesis description documents. For example, the output of SciAgents [33] is a document containing hypotheses, outcomes, mechanisms, design principles, unexpected properties, comparison, and novelty, each having its expanded version. AI co-scientist [32] also outputs a structured document but with different sections: introduction, recent findings, related research, rationale, specificity, experimental design, and validation. Si et al. [54] request LLM agents to generate an "idea", containing several components (e.g., problem, existing methods, motivation, proposed method, experiment plan), similar to a research proposal. Whereas, the Piflow framework [55] requires LLM agents to generate a "hypothesis structure" consisting of rationale, hypothesis, reiterate, and an experimental candidate.

## 4. Discussion and Conclusions

Our work here intends to be more descriptive than prescriptive. We outline the various definitions and instances of hypotheses in existing scientific literature (and beyond). In particular, we focus on definitions of hypotheses and related concepts in recent work in NLU.

We hope that this work may raise awareness within the hypothesis mining community about standardization of corpus-level tools, e.g., knowledge graphs representing connections amongst interdisciplinary hypotheses, or hypothesis generation models across multiple domains. In lieu of standardization, the inclusion of explicit, clear definitions of hypotheses (formal, where possible) could improve alignment and assembly.

For more than two decades, many in the research community have advocated for open data, open materials, preregistration, and other best practices as central to the vision of a *searchable and interpretable scholarly record*. With recent technological advances, this vision is on the horizon. The ultimate goals of an interpretable scholarly record are: robust and efficient scientific progress; thoughtful allocation of community resources toward important open problems; and honest dialogue with public and policymakers. How–precisely–a queryable scholarly corpus comes together is an open question. Here, we suggest that dissemination of clear, consistent, well-specified machine-readable hypotheses, claims, and evidence are critical to this mission.

## Acknowledgments

This project is partially supported by Open Philanthropy.

## Declaration on Generative AI

In accordance with the principles of responsible AI use, we disclose that generative AI was used solely for language editing and did not contribute to the scientific content or analysis presented in this work.

## References

[1] T. S. Kuhn, D. Hawkins, The structure of scientific revolutions, American Journal of Physics 31 (1963) 554–555. URL: https://doi.org/10.1119/1.1969660. doi:10.1119/1.1969660. arXiv:https://pubs.aip.org/aapt/ajp/article-pdf/31/7/554/12111921/554_1_online.pdf.

[2] K. R. Popper, Science as falsification, Conjectures and refutations 1 (1963) 33–39.

[3] I. Lakatos, History of science and its rational reconstructions, in: PSA: Proceedings of the biennial meeting of the philosophy of science association, volume 1970, Cambridge University Press, 1970, pp. 91–136.

[4] R. A. Fisher, Statistical methods for research workers, 5, Oliver and Boyd, 1928.

[5] R. Fisher, Statistical methods and scientific induction, Journal of the Royal Statistical Society Series B: Statistical Methodology 17 (1955) 69–78.

[6] D. J. Glass, N. Hall, A brief history of the hypothesis, Cell 134 (2008) 378–381.

[7] G. Fauconnier, Mappings in thought and language, Cambridge University Press, 1997.

[8] B. C. Malt, A. Majid, How thought is mapped into words, Wiley Interdisciplinary Reviews: Cognitive Science 4 (2013) 583–597.

[9] T. Yarkoni, The generalizability crisis, Behavioral and Brain Sciences 45 (2022) e1.

[10] V. Stodden, On emergent limits to knowledge—or, how to trust the robot researchers: A pocket guide, Harvard Data Science Review 6 (2024).

[11] V. Karasmanēs, V. Karasmanis, The hypothetical method in Plato's middle dialogues, Ph.D. thesis, University of Oxford, 1987.

[12] H. W. Ausland, Socrates' dialectical use of hypothesis, in: New Perspectives on Platonic Dialectic, Routledge, 2022, pp. 25–50.

[13] J. Barnes, Posterior analytics (1994).

[14] G. Galilei, Dialogue concerning the two chief world systems, Berkeley: University of California Press.·[1638] (1914).

[15] I. Newton, Philosophiae naturalis principia mathematica, volume 1, G. Brookman, 1833.

[16] R. Descartes, A discourse on method, JM Dent & Sons Limited, 1912.

[17] S. Sakellariadis, Descartes's use of empirical data to test hypotheses, Isis 73 (1982) 68–76.

[18] J. F. Quinn, A. E. Dunham, On hypothesis testing in ecology and evolution, The American Naturalist 122 (1983) 602–617.

[19] J. Leplin, A novel defense of scientific realism, Oxford University Press, 1997.

[20] D. Shapere, The structure of scientific revolutions, The Philosophical Review 73 (1964) 383–394.

[21] S. M. Rajtmajer, T. M. Errington, F. G. Hillary, How failure to falsify in high-volume science contributes to the replication crisis, Elife 11 (2022) e78830.

[22] B. A. Nosek, C. R. Ebersole, A. C. DeHaven, D. T. Mellor, The preregistration revolution, Proceedings of the National Academy of Sciences 115 (2018) 2600–2606.

[23] R. Mueller, S. Abdullaev, Deepcause: Hypothesis extraction from information systems papers with deep learning for theory ontology learning (2019).

[24] S. Kumar, T. Ghosal, V. Goyal, A. Ekbal, Can large language models unlock novel scientific

research ideas?, CoRR abs/2409.06185 (2024). URL: https://doi.org/10.48550/arXiv.2409.06185. doi:10.48550/ARXIV.2409.06185. arXiv:2409.06185.

[25] Q. Wang, D. Downey, H. Ji, T. Hope, SciMON: Scientific Inspiration Machines Optimized for Novelty, in: L. Ku, A. Martins, V. Srikumar (Eds.), Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024, Association for Computational Linguistics, 2024, pp. 279–299. URL: https://doi.org/10.18653/v1/2024.acl-long.18. doi:10.18653/V1/2024.ACL-LONG.18.

[26] S. E. Toulmin, The uses of argument, Cambridge university press, 2003.

[27] T. Heger, A. Algergawy, M. Brinner, J. M. Jeschke, B. König-Ries, D. Mietchen, S. Zarrieß, Natural language hypotheses in scientific papers and how to tame them: Suggested steps for formalizing complex scientific claims, in: Robust Argumentation Machines: First International Conference, RATIO 2024, Bielefeld, Germany, June 5–7, 2024, Proceedings, Springer-Verlag, Berlin, Heidelberg, 2024, p. 3–19. URL: https://doi.org/10.1007/978-3-031-63536-6_1. doi:10.1007/978-3-031-63536-6_1.

[28] N. Alipourfard, B. Arendt, D. M. Benjamin, N. Benkler, M. Bishop, M. Burstein, M. Bush, J. Caverlee, Y. Chen, C. Clark, et al., Systematizing confidence in open research and evidence (score), SocArXiv (2021).

[29] R. Vasu, C. Sarasua, A. Bernstein, Scihyp: A fine-grained dataset describing hypotheses and their components from scientific articles, in: International Semantic Web Conference, Springer, 2024, pp. 134–152.

[30] S. D. Koneru, J. Wu, S. Rajtmajer, Can large language models discern evidence for scientific hypotheses? case studies in the social sciences, in: N. Calzolari, M. Kan, V. Hoste, A. Lenci, S. Sakti, N. Xue (Eds.), Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy, ELRA and ICCL, 2024, pp. 2787–2797. URL: https://aclanthology.org/2024.lrec-main.248.

[31] D. Wadden, S. Lin, K. Lo, L. L. Wang, M. van Zuylen, A. Cohan, H. Hajishirzi, Fact or fiction: Verifying scientific claims, in: B. Webber, T. Cohn, Y. He, Y. Liu (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020, Association for Computational Linguistics, 2020, pp. 7534–7550. URL: https://doi.org/10.18653/v1/2020.emnlp-main.609. doi:10.18653/V1/2020.EMNLP-MAIN.609.

[32] J. Gottweis, W. Weng, A. N. Daryin, T. Tu, A. Palepu, P. Sirkovic, A. Myaskovsky, F. Weissenberger, K. Rong, R. Tanno, K. Saab, D. Popovici, J. Blum, F. Zhang, K. Chou, A. Hassidim, B. Gokturk, A. Vahdat, P. Kohli, Y. Matias, A. Carroll, K. Kulkarni, N. Tomasev, Y. Guan, V. Dhillon, E. D. Vaishnav, B. Lee, T. R. D. Costa, J. R. Penadés, G. Peltz, Y. Xu, A. Pawlosky, A. Karthikesalingam, V. Natarajan, Towards an AI co-scientist, CoRR abs/2502.18864 (2025). URL: https://doi.org/10.48550/arXiv.2502.18864. doi:10.48550/ARXIV.2502.18864. arXiv:2502.18864.

[33] A. Ghafarollahi, M. J. Buehler, Sciagents: Automating scientific discovery through bioinspired multi-agent intelligent graph reasoning, Advanced Materials 37 (2025) 2413523. URL: https://advanced.onlinelibrary.wiley.com/doi/abs/10.1002/adma.202413523. doi:https://doi.org/10.1002/adma.202413523. arXiv:https://advanced.onlinelibrary.wiley.com/doi/pdf/10.1002/adma.202413523.

[34] T. Khot, A. Sabharwal, P. Clark, SciTaiL: A Textual Entailment Dataset from Science Question Answering, Proceedings of the AAAI Conference on Artificial Intelligence 32 (2018). URL: https://ojs.aaai.org/index.php/AAAI/article/view/12022. doi:10.1609/aaai.v32i1.12022.

[35] B. Dalvi, P. Jansen, O. Tafjord, Z. Xie, H. Smith, L. Pipatanangkura, P. Clark, Explaining answers with entailment trees, in: M. Moens, X. Huang, L. Specia, S. W. Yih (Eds.), Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021, Association for Computational Linguistics, 2021, pp. 7358–7370. URL: https://doi.org/10.18653/v1/2021.emnlp-main.585. doi:10.18653/V1/2021.EMNLP-MAIN.585.

[36] I. Dagan, D. Roth, F. Zanzotto, M. Sammons, Recognizing textual entailment: Models and applications, Springer Nature, 2022.

[37] S. Storks, Q. Gao, J. Y. Chai, Recent advances in natural language inference: A survey of benchmarks, resources, and approaches, arXiv preprint arXiv:1904.01172 (2019).

[38] V. Do, O.-M. Camburu, Z. Akata, T. Lukasiewicz, e-snli-ve: Corrected visual-textual entailment with natural language explanations, arXiv preprint arXiv:2004.03744 (2020).

[39] L. Bentivogli, P. Clark, I. Dagan, D. Giampiccolo, The fifth pascal recognizing textual entailment challenge., TAC 7 (2009) 1.

[40] C. Shivade, Mednli—a natural language inference dataset for the clinical domain, (No Title) (2017).

[41] M. Bastan, M. Surdeanu, N. Balasubramanian, BioNLI: Generating a biomedical NLI dataset using lexico-semantic constraints for adversarial examples, in: Y. Goldberg, Z. Kozareva, Y. Zhang (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2022, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022, pp. 5093–5104. URL: https://aclanthology.org/2022.findings-emnlp.374/. doi:10.18653/v1/2022.findings-emnlp.374.

[42] T. Achakulvisut, C. Bhagavatula, D. Acuna, K. Kording, Claim extraction in biomedical publications using deep discourse model and transfer learning, arXiv preprint arXiv:1907.00962 (2019).

[43] X. Wei, M. R. U. Hoque, J. Wu, J. Li, Claimdistiller: Scientific claim extraction with supervised contrastive learning, in: Proceedings of Joint Workshop of the 4th Extraction and Evaluation of Knowledge Entities from Scientific Documents (EEKE2023) and the 3rd AI + Informetrics (AII2023) co-located with the JCDL 2023, Santa Fe, New Mexico, United States, June 26 – June 30, 2023, 2023, pp. 3487–3496. URL: https://ceur-ws.org/Vol-3451/paper11.pdf.

[44] E. White, K. B. Cohen, L. Hunter, The CISP annotation schema uncovers hypotheses and explanations in full-text scientific journal articles, in: Proceedings of BioNLP 2011 Workshop, BioNLP '11, Association for Computational Linguistics, USA, 2011, p. 134–135.

[45] Lancet, Information for Authors, https://www.thelancet.com/pb-assets/Lancet/authors/tl-info-for-authors-1740074875577.pdf, 2025.

[46] A. Saakyan, T. Chakrabarty, S. Muresan, COVID-fact: Fact extraction and verification of real-world claims on COVID-19 pandemic, in: C. Zong, F. Xia, W. Li, R. Navigli (Eds.), Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Association for Computational Linguistics, Online, 2021, pp. 2116–2129. URL: https://aclanthology.org/2021.acl-long.165/. doi:10.18653/v1/2021.acl-long.165.

[47] M. Sarrouti, A. Ben Abacha, Y. Mrabet, D. Demner-Fushman, Evidence-based fact-checking of health-related claims, in: M.-F. Moens, X. Huang, L. Specia, S. W.-t. Yih (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2021, Association for Computational Linguistics, Punta Cana, Dominican Republic, 2021, pp. 3499–3512. URL: https://aclanthology.org/2021.findings-emnlp.297/. doi:10.18653/v1/2021.findings-emnlp.297.

[48] B. P. Majumder, H. Surana, D. Agarwal, B. D. Mishra, A. Meena, A. Prakhar, T. Vora, T. Khot, A. Sabharwal, P. Clark, Discoverybench: Towards data-driven discovery with large language models, in: The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025, OpenReview.net, 2025. URL: https://openreview.net/forum?id=vyflgpwfJW.

[49] K. BahadarKhan, A. A Khaliq, M. Shahid, A morphological hessian based approach for retinal blood vessels segmentation and denoising using region based otsu thresholding, PLOS ONE 11 (2016) 1–19. URL: https://doi.org/10.1371/journal.pone.0158996. doi:10.1371/journal.pone.0158996.

[50] Y. Zhou, H. Liu, T. Srivastava, H. Mei, C. Tan, Hypothesis generation with large language models, in: L. Peled-Cohen, N. Calderon, S. Lissak, R. Reichart (Eds.), Proceedings of the 1st Workshop on NLP for Science (NLP4Science), Association for Computational Linguistics, Miami, FL, USA, 2024, pp. 117–139. URL: https://aclanthology.org/2024.nlp4science-1.10/. doi:10.18653/v1/2024.nlp4science-1.10.

[51] C. Tan, L. Lee, B. Pang, The effect of wording on message propagation: Topic- and author-controlled natural experiments on Twitter, in: K. Toutanova, H. Wu (Eds.), Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Baltimore, Maryland, 2014, pp. 175–185. URL:

https://aclanthology.org/P14-1017/. doi:10.3115/v1/P14-1017.

[52] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. Canton-Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P. S. Koura, M. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, T. Scialom, Llama 2: Open foundation and fine-tuned chat models, CoRR abs/2307.09288 (2023). URL: https://doi.org/10.48550/arXiv.2307.09288. doi:10.48550/ARXIV.2307.09288. arXiv:2307.09288.

[53] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual, 2020. URL: https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

[54] C. Si, D. Yang, T. Hashimoto, Can LLMs Generate Novel Research Ideas? A Large-Scale Human Study with 100+ NLP Researchers, in: The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025, OpenReview.net, 2025. URL: https://openreview.net/forum?id=M23dTGWCZy.

[55] Y. Pu, T. Lin, H. Chen, PiFlow: Principle-aware Scientific Discovery with Multi-Agent Collaboration, CoRR abs/2505.15047 (2025). URL: https://doi.org/10.48550/arXiv.2505.15047. doi:10.48550/ARXIV.2505.15047. arXiv:2505.15047.

[56] M. Chai, E. Herron, E. Cervantes, T. Ghosal, Exploring scientific hypothesis generation with mamba, in: L. Peled-Cohen, N. Calderon, S. Lissak, R. Reichart (Eds.), Proceedings of the 1st Workshop on NLP for Science (NLP4Science), Association for Computational Linguistics, Miami, FL, USA, 2024, pp. 197–207. URL: https://aclanthology.org/2024.nlp4science-1.17/. doi:10.18653/v1/2024.nlp4science-1.17.

[57] Z. Yang, X. Du, J. Li, J. Zheng, S. Poria, E. Cambria, Large language models for automated open-domain scientific hypotheses discovery, in: L.-W. Ku, A. Martins, V. Srikumar (Eds.), Findings of the Association for Computational Linguistics: ACL 2024, Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 13545–13565. URL: https://aclanthology.org/2024.findings-acl.804/. doi:10.18653/v1/2024.findings-acl.804.

[58] B. Qi, K. Zhang, H. Li, K. Tian, S. Zeng, Z.-R. Chen, B. Zhou, Large language models are zero shot hypothesis proposers, arXiv preprint arXiv:2311.05965 (2023).