SciHyp: A Fine-Grained Dataset Describing Hypotheses and Their Components from Scientific Articles

Rosni Vasu(✉), Cristina Sarasua, and Abraham Bernstein

Department of Informatics, University of Zurich, Zurich, Switzerland
{rosni,sarasua,bernstein}@ifi.uzh.ch

Abstract. Scientific discovery entails a detailed understanding and structuring of existing hypotheses—a challenging task due to the variety and complexity of the scientific texts. Despite efforts in domains like bio-medicine and invasion biology, there does not seem to be a curated fine-grained hypothesis dataset derived from scientific articles.

This paper presents SciHyp, a novel dataset containing the RDF description of 689 unique hypothesis sentences from 479 scientific articles in the computer science domain. The dataset describes hypotheses of two types: relation-finding hypotheses (526), which indicate a relationship between variables, and comparative (274) hypotheses, which specify comparisons between samples based on a variable and an operator. We created the dataset using a novel and multi-step annotation pipeline incorporating expert annotation, Large Language Models (LLMs) including BERT, Sci-BERT, and crowd-based refinement. Our pipeline effectively identified non-hypothesis sentences with a 96.1% consensus rate between the LLMs and crowd annotations, demonstrating its effectiveness in identifying relevant sentences that contain hypotheses. Furthermore, we extracted the individual components of hypotheses (i.e., their variables and the relation between them) using an in-context learning approach based on GPT-4.

We believe the SciHyp dataset will benefit the scientific community by offering a structured dataset for model training and evaluation, and adapting the procedure to curate and analyse large-scale hypothesis datasets.

Keywords: Scientific Hypotheses · Hypotheses Annotation · Large Language Models · Information Extraction · Knowledge Graph · Scientific Knowledge Management

1 Introduction

Scientific hypotheses are crucial for advancing data-driven decision-making and scientific inquiry. The automatic extraction of hypotheses from scientific articles and their structured representation can assist applications like hypothesis-driven literature search, trend analysis, and the development of knowledge graphs. However, current approaches [8,30] struggle due to the complexity of the language

G. Demartini et al. (Eds.): ISWC 2024, LNCS 15233, pp. 134–152, 2025.
https://doi.org/10.1007/978-3-031-77847-6_8

structures and the inadequate consideration of the context surrounding hypotheses [5,29], making the automatic extraction of both implicit and explicit hypotheses challenging. Moreover, manual hypothesis identification is time-consuming and often hindered by the lack of explicitly stated hypotheses with the necessary components (such as the variables measured and the relationship between them).

Various approaches to hypothesis annotation, ranging from rule-based systems to machine learning algorithms [5,30], have been developed to address the task of identifying and analysing the research hypotheses in scientific texts. However, the models face challenges from the complexity and diversity of disciplines, varying writing styles, and the implicit nature of some hypotheses used in the scientific texts, as posed in [5]. While there are datasets [5,16] in domains such as social science and invasion biology, they lack fine-grained, component-level annotations of hypotheses, which is important for hypothesis generation. To address the aforementioned challenges and to facilitate new algorithms for hypothesis identification in scientific literature, we present *SciHyp: a fine-grained and structured dataset containing the RDF description of the hypotheses specified in 479 scientific publications.* This dataset is specifically created for the classification of hypotheses and synthesis of their associated components from research articles. To create SciHyp, we use a novel, human-AI annotation pipeline that, given scientific publications, obtains fine-grained data about the hypotheses in the publications and the components of these hypotheses. We integrate the capabilities of LLMs, crowdsourcing, and expert knowledge, thereby creating a synergy that ensures quality and reliable annotations. This procedure forms the backbone of SciHyp, facilitating the development of more accurate models for hypothesis detection and their component tasks. Moreover, our pipeline considers the effects of contextual information on the performance of the models. In summary, SciHyp not only overcomes the limitations of existing datasets but also pushes the boundaries in large language models (LLMs) for hypothesis detection and synthesizing their components.

As a consequence, **our contributions** can be summarized as follows:

The SciHyp dataset: we create and release a novel dataset with fine-grained hypothesis annotations, along with contextual information (comprising the surrounding sentences), providing a valuable resource for researchers to develop and evaluate machine learning models for hypothesis detection and hypothesis extraction.

Extended data model for representing scientific hypotheses: we reuse, combine, and extend existing hypothesis ontologies to incorporate missing details regarding different types of hypotheses (i.e., relation-finding and comparative hypotheses; explicit and implicit hypotheses).

Pipeline for hypothesis detection and their component extraction: we present a pipeline that implements two major tasks, namely the detection of sentences in a scientific publication that refer to hypotheses, and the extraction of individual components in hypotheses (i.e., variables/groups and relations/operators), using expert, LLM, and crowd-annotations.

The remainder of the paper is structured as follows: we discuss the related works in Sect. 2. Section 3 provides the motivation and details of the data model, Sect. 4 describes the creation of the dataset, including the manual annotation (Sect. 4.1) and enhancement with LLMs (Sect. 4.2) followed by an evaluation in Sects. 5 and a discussion of use cases in Sect. 6. Finally, Sect. 7 concludes with a discussion of the limitations and future research directions.

2 Related Work

In recent years, there has been a growing interest in developing and applying Semantic Web technologies to improve the representation, accessibility, and usability of scientific knowledge. This section reviews existing efforts in modelling, extracting, and classifying hypotheses from the literature.

Formal Representations of Scientific Hypotheses. While LABORS [32] offers textual and logical representations for hypotheses within Systems Biology and Functional Genomics, and DISK [15] hypothesis ontology focuses on tracking hypothesis evolution and provenance, both lack the broader applicability required for cross-disciplinary scientific hypotheses. Our research reuse and extends DISK to create a general ontology applicable across scientific domains. The "super-pattern" (SP) semantic template representing scientific claims [7] enhances literature exploration by integrating findings into a knowledge graph across disciplines. We draw inspiration from the "super-pattern's" relation types property to represent the relation stated in the hypotheses. Moreover, LABORS and DISK do not capture the information about the different types of hypotheses we consider in this work. Our research builds upon this idea by adapting many constructs from the DISK and Scientific Question Ontology (SQO) [36], which offer semantic templates for customizable scientific questions, to ensure maximum interoperability.

Hypotheses Extraction and Classification. Previous methods have demonstrated advancement in this field, such as the fine-tuned BioBERT-large model [29] focus on causal relations in medical texts, DeepCause [24] utilizes sequence labeling and focus on cause-effect extraction, NLP models [8] for extracting explicit hypotheses (indicated by phrases such as 'H1:', 'Hypothesis 1:') in social science publications, and a supervised method [30] for extracting Meta-Knowledge (MK) dimensions, focusing on Research Hypotheses and New Knowledge from the scientific literature. However, the initial annotation efforts of the latter were primarily on abstracts with limited granularity within *Research Hypothesis*.

The INAS dataset [5] provides a network of domain-specific major hypotheses for invasion biology, labeling complete sentences to differentiate explicit mentions of the main hypothesis and textual references to it. Our SciHyp dataset goes further by labeling each of the components in the hypotheses (i.e., variables and relations), and the type of hypothesis (i.e., comparative and relation finding). Unlike INAS's use of BERT-based models for abstract classification,

Table 1. Comparison of ontologies across features. ✓ indicates the feature is supported by the ontology.

Features Supported	LABORS [32]	DISK [15]	SP [7]	SQO [36]	SciHyp
Scientific paper source	✓	–	✓	–	✓
Types of hypothesis	–	–	–	–	✓
Variables/Groups of hypothesis	–	✓	–	✓	✓
Relation/Operator	✓	–	✓	–	✓
Cross-disciplinary applicability	–	✓	✓	✓	✓

our multi-step pipeline leverages LLMs and crowdsourced human computation for detailed sentence-level annotation. SciHyp extends beyond DeepCause [24], which focuses solely on cause-effect extraction, by covering relation-finding and comparative hypotheses, incorporating surrounding information for better context, and addressing implicitly stated hypotheses with an efficient annotation framework.

While the studies above substantially contribute to the representation and detection of hypotheses within specific fields, our research advances them further. As Table 1 shows, SciHyp uniquely enhances hypothesis modeling by supporting structured analysis of scientific texts, detailed extraction of hypothesis components, and accommodating multiple hypothesis types. Although our dataset is currently focused on the computer science domain, our cross-disciplinary methodology is designed to analyze scientific texts regardless of the sections to take into account. We uniquely combine LLMs and human knowledge for the accurate detection and fine-grained annotation of hypotheses, resulting in a more detailed dataset. This exhaustive procedure expands the scope and applicability of hypothesis representation and detection, facilitating scientific discovery.

3 The SciHyp Dataset

The SciHyp Dataset is a novel, fine-grained dataset describing scientific hypotheses and their components. This section provides the details of the data model and corpus collection.

3.1 SciHyp Data Model

In scientific research, *a hypothesis is a prediction or testable explanation that researchers will verify with experiments* [34]. The purpose of the SciHyp data model is to capture detailed information about the components present in hypothesis statements written as natural language text in scientific articles. Following the standard classification in hypothesis testing methods [3], we distinguish between *relation finding* and *comparative* hypotheses.

Relation-Finding Hypotheses identify the relationship between two variables, differentiating between the dependent variable (DV) and the independent variable (IV). For instance, "the complexity of the search query influences

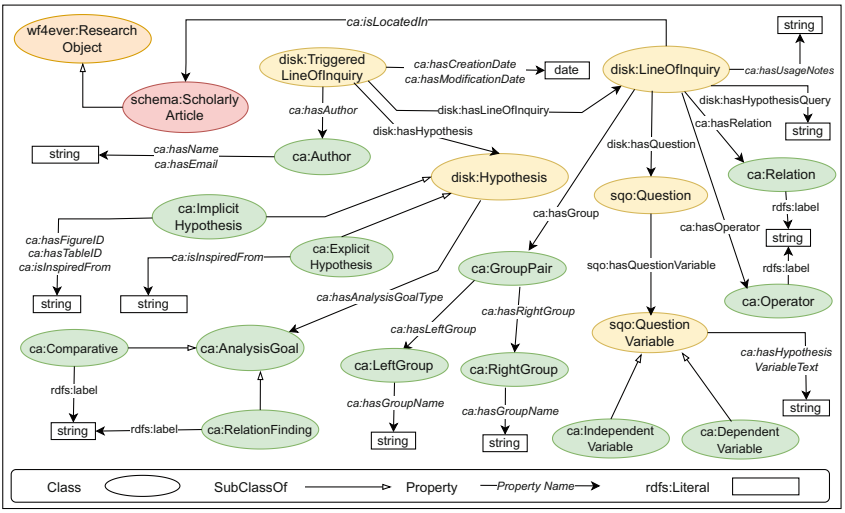


Fig. 1. Description of the data model. We implemented the green elements and the properties with prefix ca as an extension the DISK disk and Scientific Questions sqo ontologies. (Color figure online)

the time spent browsing results" is a relation-finding hypothesis, where *the complexity of the search query* is the independent variable, *influences* is the relation, and *time spent browsing results* is the dependent variable.

Comparative Hypotheses compare two samples in terms of a variable, using an operator (e.g., "greater than," "less than," "equal to"). For example, "the accuracy of a recommender system using temporal data is higher compared to the accuracy of a recommender system using temporal and spatial data," where "the accuracy" is the variable, "a recommender system using temporal data" is one sample, "a recommender system using temporal and spatial data" is the second sample, and "higher" is the operator.

As indicated in Fig. 1, to implement these hypothesis components, we reuse parts of the DISK hypothesis [15] (using the prefix "disk") and the Scientific Questions [36] (using the prefix "sqo:") ontologies. The DISK ontology models hypotheses as lines of inquiry containing question specified into SQO's question variable. We used the question variable to represent our hypothesis variables. Our data model extends DISK to incorporate some new properties. First, we needed to capture the information of the relations between variables (e.g., "increases", "decreases", "influences"). To model that, we reuse the relation types defined in the Super-Pattern ontology [7]. However, to better suit the specific requirements of our data model, we introduced the property has relation and class ca:Relation for *relation finding* hypothesis, inspired by the Super-Pattern ontology.

To model the two aforementioned types of hypotheses, we added new classes and properties that we connect to DISK elements. We define two subclasses of

the class disk:Hypothesis, namely ca:ExplicitHypothesis, which refers to a hypothesis that is explicitly formulated in the text, and ca:ImplicitHypothesis, which refers to a hypothesis that can be inferred from evaluation results, such as tables or figures. We define a class ca:AnalysisGoal and the analysis goal type property to represent the type of the hypotheses. We added two new subclasses ca:RelationFinding and ca:Comparative of the class ca:AnalysisGoal to capture the type of analysis goal. We included the subclasses ca:IndependentVariable and ca:DependentVariable of the class sqo:QuestionVariable for *relation-finding* hypotheses, where sqo:QuestionVariable is used to represent the variable measured in the *comparative* hypotheses. The has hypothesis variable text property captures the variable text.

Further, we included the ca:GroupPair class to represent the two samples in the *comparative* hypothesis with the properties has left group and has right group. We added the classes ca:LeftGroup and ca:RightGroup to further specify these samples and the ca:hasGroupName property captures the textual information. Similar to has relation, we incorporated the has operator property and the ca:Operator class to denote the comparison between two samples in a comparative hypothesis.

Furthermore, to indicate the provenance of the hypotheses, we use *Schema.org*[1] and the is located in property to represent the scholarly article the hypotheses belong to. We validated our resulting ontology using the OOPS! tool [28], which confirmed that the ontology does not contain common pitfalls.

3.2 The SciHyp Corpus

We used the S2ORC dataset [23] comprising 8.1 million open-access scientific publications with metadata, paper abstracts, resolved bibliographic references, and structured full text. We selected open access publications[2] from a set of computer science conferences that tend to publish publications containing explicit hypotheses: CSCW, HCOMP, IUI, and CHI. This pre-selection resulted in 3354 articles. The detailed statistics of this corpus are summarized in Table 2. The complexity of the dataset is reflected in its broad vocabulary and technical text as indicated by metrics including the Flesch Reading Ease score [14] and Type-Token Ratio (TTR) [18].

4 The SciHyp Pipeline: Integrating Expert, Crowd, and LLM Annotations

This section outlines the overview of the dataset generation pipeline (see Fig. 2), starting with the annotation of a sample by expert annotators. This data is used both as ground truth and as input to automated annotation processes. The remaining pipeline shows the multi-step annotation process combining human- and AI-based annotation.

[1] https://schema.org/ScholarlyArticle.
[2] Publications that contained a link to an accessible PDF.

Table 2. SciHyp Scientific Article Corpus Statistics

Metric	Value
#article	3354
#sentences	748213
Avg. words per article	5177.414 ± 2810.29
Avg. sentences per article	223.081 ± 122.066
#unique words	390773
Avg. readability score	34.37 ± 28.4
Type-Token Ratio (TTR)	0.023

In this pipeline, we split the larger task of annotating hypothesis statements in scientific papers into two different sub-tasks:

T1. Hypothesis Sentence Detection: The objective of this task is to identify whether individual sentences within scientific texts contain hypotheses. For each sentence x_i and its surrounding context (defined as the adjacent $\pm n$ sentences), the goal is to obtain a binary label y_i that indicates the presence or absence of a hypothesis in x_i. The context c_i is important as it provides additional information that may influence the detection of hypotheses in x_i.

T2. Hypotheses Component Extraction: This task aims to dissect a hypothesis into its fundamental structure. ***2.1. Variable Extraction:*** Given a sentence x_i previously classified as a hypothesis and its corresponding context c_i, the objective is to extract specific variables and groups (in the case of comparative hypotheses) that adhere to the predefined structure we have outlined earlier. These components, marked as $z_1, \ldots z_m$, are extracted using prompt augmentation or decomposition methods [9]. The model trained for this task uses features from x_i and c_i to accurately synthesize these components. ***2.2. Relation Classification:*** This task focuses on classifying the relationships (in relation-finding hypotheses) or operators (in comparative hypotheses) present within sentences labeled as hypotheses. Given a sentence, x_i, which contains components expressed as $z_1, \ldots z_m$, the objective is to determine the specific relation or operator between the variables or groups respectively, denoted as y_i. The set Y encompasses all possible relationships/operators that can be inferred. The classification model aims to assign an appropriate relation within Y to x_i by leveraging both the sentence and its variables as input.

We implement task T1 (sentence detection) using LLMs and crowd-computation, while task T2 is implemented with LLMs. Due to the complexity of scientific texts (see Table 2) and the need for an efficient annotation process, we use a set of 229 expert labels to automatically generate the weak labels for the hypothesis detection task (T1). Furthermore, a subset of weakly labeled sentences, along with their context, is sent to be processed by the crowd, who is asked to identify if a sentence contains a hypothesis or not. The resulting crowd-refined labels are used in a few-shot in-context learning method that extracts

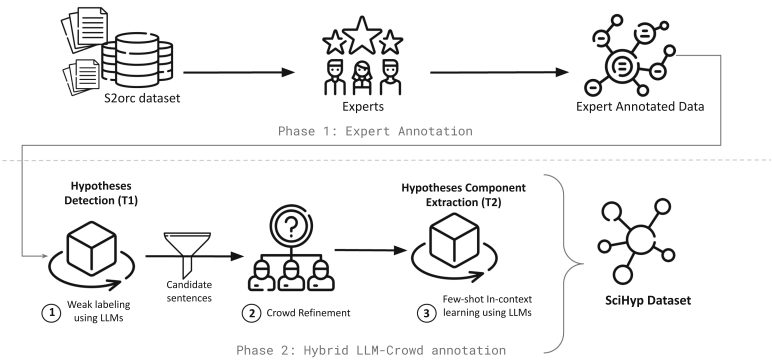


Fig. 2. SciHyp Pipeline: Dataset Annotation Process.

the hypothesis variables or groups (T2.1) and identifies the relation or operator stated in the hypothesis (T2.2), as illustrated in Fig. 2. [3,4]

4.1 Ground Truth Generation

As a first step (Phase 1 in Fig. 2), we proceeded with an expert annotation process to be able to curate the ground truth. From the SciHyp corpus (Table 2), we randomly sampled 50 publications from the set of 1013 publications that contained the word "hypothesis" or derived words from it for further expert annotation process. We asked two expert annotators to label the publications, mark the sentences containing hypotheses, and identify the hypothesis components. The two experts are two of the authors of this work, who are very familiar with scientific publications.

The annotators used an in-house developed tool (the *CrowdAlytics Annotation Tool*) that provides a GUI to annotate the hypothesis statements and the hypothesis components in a publication's text based on the ontology described in Sect. 3.1. Each annotator labeled publications individually. Subsequently, we measured the disagreement between the annotators. An exhaustive analysis of the disagreement cases and the ensuing discussion to resolve discrepancies led to increased agreement and a final ground truth annotation.

Given the imbalanced nature of this data, widely-used inter-rater reliability measures, such as Cohen's Kappa [10], are not suitable in our case [37]. Hence, we measured the percentage of cases in which both annotators agreed.

Following best practices in data annotation procedures, we implemented the annotation in two batches of 25 publications for early consensus discussion. Moreover, due to the complexity of the task, we implemented the discussion and disagreement resolution procedure for T1 (sentence classification) separately from T2.1 (variable/groups extraction) and T2.2 (relation/operator extraction).

[3] Icons made by Freepik from www.flaticon.com.
[4] Icons made by pictogramer from www.flaticon.com.

Table 3. Results of Hypothesis Classification on Test Dataset (f1-score)

Model	Without Context	With Context
BERT *fine-tuned*	0.956	0.956
Sci-BERT *fine-tuned*	0.924	**0.967**
Sci-BERT *not fine-tuned*	0.392	**0.661**

Sentence Classification. In the phase with the first 25 publications, the annotators initially reached agreement on 39.05% of the sentences, which increased after the discussion up to 97.14%—a set of 102 sentences were classified as hypotheses by both annotators and 3 sentences were classified as hypotheses by only one annotator. This process resulted in a revision of the annotation guidelines, as there were criteria jointly determined from concrete cases. For instance, one such case was *questions* (e.g.,"*How will the sensemaking translucence interface affect participants cognitive workload?*") mentions a relationship between variables, it implies an underlying hypothesis. In the second phase, with the remaining 25 publications, an agreement was reached in 35.66% of the cases, increasing to 94.07% after the discussions—a set of 127 sentences were classified as hypotheses by both annotators, and 16 sentences only had the vote of one annotator. The cases in which the two annotators did not reach an agreement were discarded.

The disagreement in the annotation was attributed to several factors. Firstly, there were statements in the papers that the authors did not explicitly present as hypotheses. In some other cases, the statements were poorly structured, which often led to different interpretations. Secondly, some of the statements were particularly difficult to classify, as they contained a similar structure to hypotheses. However, they were actually claims related to exploratory procedures that did not adhere to hypothesis testing. Additionally, the high effort required to annotate full articles in a batch led to either of the annotators overlooking some statements, resulting in higher disagreement because when one annotator labeled a sentence that the other annotator did not label, we computed this case as an instance of disagreement. As a result, we obtained 229 (sentence-level) annotations from 41 unique articles (out of the 50 randomly sampled articles).

Variables/Groups and Relation/Operator Extraction. The annotators revised the data pertaining to the components of the hypotheses. In some cases, even when a hypothesis is specified explicitly as such by the authors of the paper, the expert annotators rephrased the names of variables. In other cases, the names of the variables were captured verbatim. The annotators included the original sentence in the publications' text for each hypothesis. During this phase, the annotators agreed to remove one sentence that failed to adequately represent the variables/groups and the corresponding relation/operator, leading to 228 (sentence-level) annotations and their associated components from 41 unique articles (out of the 50 randomly sampled articles).

4.2 Hybrid LLM-Crowd Annotation

In this subsection, we provide a detailed description of each step of the hybrid LLM-Crowd annotation process (phase 2 in Fig. 2), including AI-generated weak labeling, crowd refinement, and a fine-grained annotation technique.

Step 1: Weak labeling using LLMs. We use BERT [20] and Sci-BERT [4] as foundational models for sentence-level annotation (Task 1). We evaluated a total of 6 variants (i.e., the two models including or excluding context, and fine-tuning[5]) using the expert annotations for fine-tuning (Sect. 4.1). Additionally, we sampled negative sentences that we identified as dissimilar using cosine similarity and added them to the annotated texts. Data were divided into training, validation, and test sets (274, 92, 92 sentences, respectively). The validation set was used for hyperparameter tuning and model validation in our standard dataset split (i.e., training/validation/test split). We see in Table 3, which lists the F1-score of the models on the test sets, that the models were able to learn the task by fine-tuning.

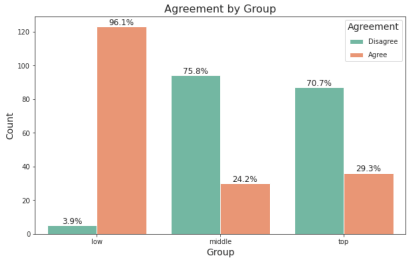
Model Ensemble: Instead of relying on the output of a single model, we leverage the collective intelligence of all the models. A logistic regression ensemble weights each model's contribution by its F1-score ($f1\_score_i$), creating a unified scoring mechanism. We compute the aggregated weak labelling score by summing each models' predicted probability times its F1-score weight (w_i), i.e., $\sum_{i=1}^{6} p_i * w_i$, where $w_i = f1\_score_i / \sum_{i=1}^{6} f1\_score_i$. We used an empirically[6] determined threshold value of 0.60 to ensure predictions with a high confidence.

Re-sampling Process: We observed from the predicted probability distribution that only 9.42% of sentences likely represent hypotheses. Based on this, we categorize the sentences of Scihyp corpus into confidence bins—*low* $[0, 0.5)$, *middle* $[0.5, 0.8)$, or *high* $[0.8, 1.0)$—totaling about 49K sentences with the *high* label. These *weak* labels, however, introduce some inherent noise as evident from the mean number of filtered sentences per article for certain models (e.g., 24.49 for *fine-tuned bert-with-no-context*). Moreover, sentence specificity is shown to be a key property to obtain good candidates for argument extraction [33]. Thus, we selected sentences with a specificity score [21][7] above 0.41—which is the mean score—for the crowd annotation process, to have a more efficient use of human annotation. Although the LLM-generated weak labels may be less precise and accurate compared to human annotations, they offer a valuable starting point for humans, as the LLMs can identify potentially interesting sentences, reducing the effort that finding them from scratch requires. This significantly reduces the workload for human annotators, potentially enhancing the effectiveness of the workflow.

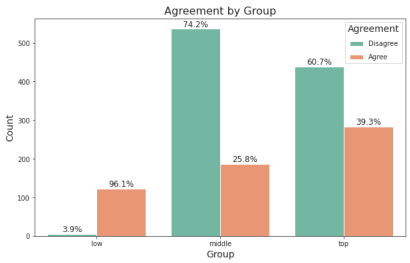
[5] Given that Sci-BERT is trained explicitly on scientific literature, its non-fine-tuned variant is considered relevant for our evaluation, potentially providing valuable insights even without additional fine-tuning specific to our task.

[6] by running cross-validation in a threshold range $0.5 - 1.0$ with 0.05 step.

[7] https://github.com/wjko2/Domain-Agnostic-Sentence-Specificity-Prediction.



Crowd-sourced annotations (before removing further the *low* category sentences)



Complete crowd-sourced annotations

Fig. 3. Group-wise Agreement Rates between human and the ensemble model, detailing matches and mismatches by group percentage.

Step 2: Refinement by the Crowd. In this step, we recruit Prolific crowd workers and ask them to classify a set of sentences into hypotheses or non-hypotheses to enhance the accuracy of the labels produced by AI models. The participants were rewarded at a rate of 9 GBP/hr. We implemented crowd tasks deploying the Potato annotation tool [26] and redirecting Prolific users to our Potato instance.

For the crowd tasks, we sampled 2400 sentences, evenly distributing across "low," "middle," or "high" categories based on the LLM's weak labels (Sect. 4.2). We employed a batch-by-batch approach with 15 sentences to be annotated per batch. To ensure high-level expertise, we introduced two types of pre-screening mechanisms in our tasks that help us filter crowd workers adequately: firstly, we deployed a survey to select English-speaking participants with a *Doctorate* as their highest education level. Secondly, we used an accuracy-based pre-screening, using one batch of 15 sentences to test the performance of crowd workers on our task. The crowd workers who passed this test (i.e., workers with an accuracy higher than 0.8 in this test) could complete their initial batches. Further, the annotations are processed using MACE (Multi-Annotator Competence Estimation) [25] to identify the reliability of each worker. Workers with a reliability score higher than a set threshold ($\geq$0.8) are re-invited to participate in additional annotation tasks.

Human vs. LLM Annotations: The hybrid process including model predictions (step 1) and human refinement (step 2), was essential in our attempt to gather a comprehensively annotated dataset. After reviewing initial batches, it became evident that the ensemble model was proficient at filtering out non-relevant sentences, specifically those that did not constitute hypotheses (Fig. 3, left). This allowed us to focus on potentially subtle or complex hypothesis statements, ensuring our annotated dataset was comprehensively developed. In the subsequent annotation task, we considered the sentences from *middle* and *high* categories. In the right Fig. 3, the agreement between the final predictions (using MACE) and the AI annotations across the sentence categories ('low,' 'middle,' and 'top') is displayed. The high alignment in the 'low' group indicates the model's capability to correctly filter out non-relevant sentences. We noticed the imprecise ensemble annotations by comparing them with expert annotations.

This emphasized the importance of human expertise in the refinement phase. Despite their imprecision, the initial weak labels serve as a starting point for human refinement. The crowd-refinement phase resulted in a total of 1575 annotations, and 460 of those were positively labeled as *hypothesis*.

Step 3: Fine-grained Annotation using few-shot in-context learning. In this step, we implement fine-grained annotations of hypotheses with the help of experts and LLMs. To this end, we consider the manually annotated data from Sect. 4.1 for component extraction and use the in-context learning paradigm [17] to annotate the sampled hypotheses statements from Step 2. In our methodology, we adopt few-shot-in-context learning with GPT-4 [1] for fine-grained hypotheses annotation, a cost-effective approach ideal for tasks with limited data. This technique, leveraging few-shot [31,38] and in-context learning [6], enables the model to learn from minimal examples within prompts [17,22], which is essential for understanding the detailed aspects of the sentences without extensive training samples.

Dynamic Prompting Approach: We implemented a dynamic prompting approach based on [22] to instruct the model more effectively for the component annotation regarding new hypotheses. We chose GPT-4 for the in-context learning approach due to its demonstrated effectiveness and promising results in similar tasks [27]. This technique involves crafting the prompts that accommodate the data from the prior manual annotations. For each new hypothesis, denoted as S_{new}, we first identify the sentences from our manual annotations, that exhibit the highest semantic similarity to S_{new}, with a similarity threshold set at 0.1 and considering up to 5 instances. Using these selected samples, we then craft the prompt, represented as P. This process then integrates the new sentence as a query along with its contextual information. By doing so, P becomes tailored specifically to guide the GPT-4 model in generating the structured components of the hypotheses that align with the format of expert annotations. This method maintains a consistent, high-quality structure and enhances the overall quality of the synthesized components.

The procedure conducted by experts and LLMs in this task facilitated the extraction and synthesis of rich information associated with forming hypotheses in scientific literature. These hybrid LLM-Crowd annotations resulted in 460 hypotheses refined by the crowd. The annotations were initially captured in CSV format, which allowed us to efficiently manage and process data through various steps and by different contributors.

The comprehensive annotations obtained from this phase were automatically converted to RDF and have been integrated into the SciHyp dataset[8], concluding the hybrid LLM-crowd annotation process.

4.3 SciHyp Dataset Description

The SciHyp dataset from the pipeline includes annotations from 479 papers. Experts annotated 228 unique sentences, whereas the Human-AI pipeline cre-

[8] the final data can be explored via the SPARQL endpoint, with example queries available in our documentation.

Table 4. Summary of Dataset Statistics. Hypotheses Types (Relation Finding, Comparative) are represented as respective counts.

Detail	Expert Annotation	Human-AI Annotation
# Unique Sentences	228	460
# Hypotheses Types	178, 161	348, 113
# Unique Independent Variables	136	346
# Unique Dependent Variables	141	348
# Groups in LeftGroup	94	113
# Groups in RightGroup	102	113
# Unique Relation Types	17	34
# Unique Variables	103	112
Most Frequent Hypothesis Type	RelationFinding	RelationFinding

ated 460 sentence-level annotations. Each hypothesis involves a different set of variables and relations or different comparisons and groups. This complexity is captured during the manual annotation process. As a result, the number of individual hypotheses (178 and 161 for relation-finding and comparative, respectively) identified is higher than the count of unique hypotheses sentences (228). *SciHyp* showcases a total of 800 components: 526 *relation finding* (178 from expert annotation + 348 from Human-AI annotation) and 274 *comparative* hypotheses (161 from expert annotation + 113 from Human-AI annotation), all extracted from the scientific articles. The respective component count for each type of hypotheses is detailed in Table 4. Finally, the statistics indicate that the most frequent type of hypothesis in our dataset is *relation finding*.

5 Pipeline Evaluation

We conducted a human assessment of crowd-refined data (for Task 1) as well as the component synthesized by the LLMs (for Tasks 2) with the two expert annotators. We sampled 50 instances from 460 crowd-refined hypotheses and their generated components with two GPT-4 back-boned (Sect. 4.2) model versions—one with and one without using sentence context (as *Model 1* and *Model 2* respectively). We employed the two versions to evaluate the impact of sentence context on the component extraction quality. This comparison helps to understand how contextual information affects the accuracy and relevance of the generated components.

The evaluators rated the samples in five different dimensions: (i) *Pipeline Accuracy* denotes if the given sentence states an explicit or implicit scientific hypothesis, (ii) *Component Relevance* denotes the generated variables are closely described by the hypothesis and the context, (iii) *Component Extraction Accuracy* denotes the generated variables accurately represent the true variables in the given hypothesis, based on the context, (iv) *Relation Accuracy* denotes the classified relation reflects the connections described within the hypothesis variables accurately, and (v) *Understandability*—denotes the generated variables are clear and understandable in relation to the provided context. For each dimensions, each evaluator provided a value from a 5-point Likert scale where 1 means *Strongly Disagree* and 5 means *Strongly Agree*. **Results:** The human evaluation

results in Table 5 showed that the models' outputs were generally well-received by the experts. *Model 1* outperforms *Model 2* in variable relevance and relation accuracy. The agreement of 76% of the data from the crowd-refined sentences during the final evaluation highlights the difficulty and subjective nature of this classification task. Our study primarily aimed to evaluate the accuracy of LLM-crowd annotations, acknowledging their variance from expert standards. This assessment is important to understand the practicality and reliability of such annotations. The Cohen's Kappa score [2], as detailed in Table 5, reflects generally strong agreement among evaluators across various dimensions for both models. 72% of the components from Model 1 were rated above score 3 in terms of relevance, whereas Model 2 components were rated similarly by 62%. We believe that the contextual information provided would help in the synthesis and extraction of components that are more closely described by the hypotheses and the context. Both the models have the same ratio of components generated accurately, however, the mean rating of Model 1 is slightly above compared to Model 2. Model 1 is more accurate in depicting the relation/operator between the variables/groups in the hypotheses. 70% of the content rated above score 3 indicates that most of the components generated by the models are perceived as equally understandable by the evaluators. Overall, there are similarities in certain aspects like variable accuracy and understandability, but Model 1 shows a slight advantage in variable relevance and relation accuracy. We believe that future studies could conduct a detailed analysis of the generated hypotheses components to understand their applicability.

Table 5. Human Evaluation Scores (X%)(mean ± std. dev) and Inter-Rater Agreement (Cohen's Kappa) between evaluators. X is the percent of samples with a rating score higher than 3. In Model 1, contextual information along with examples is provided, whereas in Model 2, the contextual information was not provided.

Model	Pipeline Accuracy	Component Relevance	Component Accuracy	Relation Accuracy	Understandability
Model 1	(76%) 3.86 ± 1.65	(72%) 3.46 ± 2.06	(64%) 3.20 ± 1.99	(66%) 3.28 ± 2.12	(70%) 3.53 ± 2.11
Cohen's Kappa	0.72	0.72	0.72	0.82	0.85
Model 2	(76%) 3.86 ± 1.65	(62%) 3.45 ± 1.87	(64%) 3.04 ± 1.70	(62%) 3.14 ± 1.83	(70%) 3.57 ± 1.93
Cohen's Kappa	0.72	0.64	0.44	0.60	0.73

Discussion: Here, we discuss briefly the efficiency of LLMs for hypotheses detection, the difficulties in accurate component extraction, and the crucial role of context in understanding the scientific hypotheses. Despite the superior performance of LLMs in the task of hypothesis detection, as Table 3 shows, LLMs struggled to classify sentences from new scientific texts. Our ensemble model (Sect. 4.2) better filter out non-hypothesis sentences, as Fig. 3 illustrates, yet the evaluation scores suggest room for improvement. One promising solution would be increasing the volume of training data to refine the ensemble model's precision. Given the task complexity and the absence of large-scale expertly annotated data, our human evaluation revealed that the GPT-4 back-boned in-context learning approach moderately succeeds in synthesizing the hypothesis

components, with 72% showing effectiveness in component relevance. However, 64% component accuracy highlights the need for improvement; hence, human intervention is crucial to verify the model outcomes and ensure accuracy. While in-context learning shows promise in synthesizing hypothesis components without large-scale manually labeled data, we need to thoroughly investigate in this direction to improve precision and train open-source models like Llamma-2 [35] for our tasks. Contextual information played a crucial role in the performance of models for all three tasks. Even the manual annotation heavily depended on the full article context to mark each statement as a hypothesis or non-hypothesis. Statements often appear non-hypothesis-like in isolation, but considering the context would help to infer the underlying hypothesis.

6 Use Cases of SciHyp Dataset, Models and Pipeline

Our work's primary contribution is the generation of detailed structured hypotheses data extracted from the corpus of scientific articles. The SciHyp dataset and pipeline may support a variety of use cases. Below, we outline four potential SciHyp use cases.

Exploration of Scientific Hypotheses: The SciHyp dataset allows users to explore and navigate structured hypotheses across scientific domains. This can facilitate an intuitive understanding of current research ideas, their components, and the relationships between them, allowing the researchers to familiarise themselves with the existing scientific hypotheses of interest. By linking hypotheses using disk:lineOfInquiry and disk:hypothesis and other variable textual properties, researchers can identify common variables and relationships across papers, supporting literature analysis.

Knowledge-Guided Scholarly Search: The structured SciHyp data can act as an enabler for enhancing scholarly search engines by identifying key sentences and integrating the hypothesis-driven search capabilities. Users can obtain highlights of the variables and relations between variables of hypotheses from the articles and explore related scientific questions, making the search process more targeted and efficient.

Trend-Analysis: SciHyp will facilitate researchers to conduct trend analysis of hypotheses per topic or domain. Integrating SciHyp with a system like ScholarSight [13] for trend analysis enhances the granularity of research insights. This can highlight emerging areas of interest in a detailed view of developing hypotheses and low-level concepts from scientific texts.

Facilitating Knowledge Synthesis: SciHyp models pre-trained on the expert dataset and the pipeline can support the aggregation of hypotheses and their components described in the scientific articles across disciplines, encouraging advances in interdisciplinary research. The kind of data described in SciHyp complements the data in ORKG [19]—which focuses on annotations about the background, the contributions, the methods, the problem statement, and results

in a paper—and the data in SemOpenAlex [12]—which contains information about scientific publications and corresponding authors, institutions, journals, and concepts as RDF knowledge graph. Applying SciHyp models to ORKG and SemOpenAlex publications can enable the detailed mapping of hypotheses and corresponding components to a vast amount of scientific publications. This enables efficient cross-domain knowledge synthesis by leveraging the relationships between hypotheses, facilitating new research hypotheses development. Furthermore, linking hypotheses to external resources, such as the Data Set Knowledge Graph [11] opens avenues for automated hypotheses verification.

Additionally, adding identity links between variables (*similar_to*), linking related hypotheses (*related_to*) or sub-hypotheses to their parent hypotheses can enhance connections across domains. For example, linking a hypothesis about impact of remote work (ca:hasHypothesisVariableText) on productivity in computer science shares a variable (*remote-work*) with a hypothesis about its effects on *employee well-being* (ca:hasHypothesisVariableText) in psychology. This facilitates interdisciplinary research.

7 Conclusion

Our study introduces *SciHyp*, a scientific hypotheses dataset featuring 689 unique hypotheses sentences from 479 articles and 800 detailed annotations of hypotheses components. We reused and extended existing ontologies to represent hypotheses and defined two key annotation tasks. SciHyp offers a valuable resource for automated hypothesis and component extraction, and relation classification. Our novel hybrid human-AI pipeline showcases the synergy between human and LLMs capabilities in data curation, setting an example process to curate similar scientific resources. Our dataset can be a useful resource for automated literature review and automated knowledge discovery systems. Future work will focus on enhancing the model's applicability to other scientific domains, such as climate science or bio-medicine. Furthermore, we aim to introduce an additional crowd refinement step to process the model-synthesized components. As such, *SciHyp* is both a valuable resource and a first step to a web of hypotheses.

Resource Availability Statement: SciHyp dataset is available from Zenodo at https://doi.org/10.5281/zenodo.10949488 (Resource DOI) under the CC BY 4.0 international license, and on GitLab at https://gitlab.ifi.uzh.ch/DDIS-Public/scihyp for long-term access and version control, respectively. The dataset can be accessed through the SPARQL endpoint at https://crowdalytics.ifi.uzh.ch/sparql/dataset.html. For the source code and additional resources, visit the repository at https://gitlab.ifi.uzh.ch/DDIS-Public/scihyp. We plan to apply the SciHyp pipeline in various disciplines (e.g., bio-medicine) and make necessary adaptations to fit the specific needs of those domains. The dataset will be promoted at scientific events, and we welcome collaborations to integrate the pipeline into larger projects.

Acknowledgments. This work is partly funded by the Swiss National Science Foundation through project "CrowdAlytics" (contract no 184994) and project "Digital Deliberative Democracy" (contract no 205975).

References

1. Achiam, J., et al.: GPT-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Artstein, R., Poesio, M.: Inter-coder agreement for computational linguistics. Comput. Linguist. **34**(4), 555–596 (2008)
3. Barroga, E., Matanguihan, G.J.: A practical guide to writing quantitative and qualitative research questions and hypotheses in scholarly articles. J. Korean Med. Sci. **37**(16), e121 (2022)
4. Beltagy, I., Lo, K., Cohan, A.: SciBERT: a pretrained language model for scientific text. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 3615–3620 (2019)
5. Brinner, M., Heger, T., Zarriess, S.: Linking a hypothesis network from the domain of invasion biology to a corpus of scientific abstracts: the INAS dataset. In: Proceedings of the first Workshop on Information Extraction from Scientific Publication, pp. 32–42. Association for Computational Linguistics, Online (2022). https://aclanthology.org/2022.wiesp-1.5
6. Brown, T., et al.: Language models are few-shot learners. Adv. Neural. Inf. Process. Syst. **33**, 1877–1901 (2020)
7. Bucur, C.I., Kuhn, T., Ceolin, D., van Ossenbruggen, J.: Expressing high-level scientific claims with formal semantics. In: Proceedings of the 11th on Knowledge Capture Conference, pp. 233–240 (2021)
8. Chen, V.Z., Montano-Campos, F., Zadrozny, W., Canfield, E.: Machine reading of hypotheses for organizational research reviews and pre-trained models via R shiny app for non-programmers. arXiv preprint arXiv:2106.16102 (2021)
9. Cho, H., et al.: Prompt-augmented linear probing: scaling beyond the limit of few-shot in-context learners. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, pp. 12709–12718 (2023)
10. Cohen, J.: A coefficient of agreement for nominal scales. Educ. Psychol. Measur. **20**(1), 37–46 (1960)
11. Färber, M., Lamprecht, D.: The data set knowledge graph: creating a linked open data source for data sets. Quant. Sci. Stud. **2**(4), 1324–1355 (2021)
12. Färber, M., Lamprecht, D., Krause, J., Aung, L., Haase, P.: SemOpenAlex: the scientific landscape in 26 billion RDF triples. In: Payne, T.R., et al (eds.) International Semantic Web Conference, vol. 14266, pp. 94–112. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-47243-5_6
13. Färber, M., Nishioka, C., Jatowt, A.: ScholarSight: visualizing temporal trends of scientific concepts. In: 2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL), pp. 438–439. IEEE (2019)
14. Farr, J.N., Jenkins, J.J., Paterson, D.G.: Simplification of flesch reading ease formula. J. Appl. Psychol. **35**(5), 333 (1951)
15. Garijo, D., Gil, Y., Ratnakar, V.: The disk hypothesis ontology: capturing hypothesis evolution for automated discovery. In: K-CAP Workshops, pp. 40–46 (2017)

16. de Haan, R., Tiddi, I., Beek, W.: Discovering research hypotheses in social science using knowledge graph embeddings. In: Verborgh, R., et al. (eds.) ESWC 2021. LNCS, vol. 12731, pp. 477–494. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-77385-4_28
17. Han, X., Simig, D., Mihaylov, T., Tsvetkov, Y., Celikyilmaz, A., Wang, T.: Understanding in-context learning via supportive pretraining data. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 12660–12673 (2023)
18. Hess, C.W., Ritchie, K.P., Landry, R.G.: The type-token ratio and vocabulary performance. Psychol. Rep. **55**(1), 51–57 (1984)
19. Jaradeh, M.Y., Oelen, A., Prinz, M., Stocker, M., Auer, S.: Open research knowledge graph: a system walkthrough. In: Doucet, A., Isaac, A., Golub, K., Aalberg, T., Jatowt A., (eds.) Digital Libraries for Open Knowledge: 23rd International Conference on Theory and Practice of Digital Libraries, TPDL 2019, Oslo, Norway, 9-12 September 2019, Proceedings 23, pp. 348–351. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-30760-8
20. Kenton, J.D.M.W.C., Toutanova, L.K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of NAACL-HLT, vol. 1, p. 2 (2019)
21. Ko, W.J., Durrett, G., Li, J.J.: Domain agnostic real-valued specificity prediction. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 6610–6617 (2019)
22. Liu, J., Shen, D., Zhang, Y., Dolan, W.B., Carin, L., Chen, W.: What makes good in-context examples for GPT-3? In: Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures, pp. 100–114 (2022)
23. Lo, K., Wang, L.L., Neumann, M., Kinney, R., Weld, D.S.: S2ORC: the semantic scholar open research corpus. arXiv preprint arXiv:1911.02782 (2019)
24. Mueller, R., Abdullaev, S.: DeepCause: hypothesis extraction from information systems papers with deep learning for theory ontology learning. In: Proceedings of the 52nd Hawaii International Conference on System Sciences (2019)
25. Paun, S., Carpenter, B., Chamberlain, J., Hovy, D., Kruschwitz, U., Poesio, M.: Comparing Bayesian models of annotation. Trans. Assoc. Comput. Linguist. **6**, 571–585 (2018)
26. Pei, J., et al.: Potato: the portable text annotation tool. In: Proceedings of the The 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pp. 327–337 (2022)
27. Polak, M.P., Morgan, D.: Extracting accurate materials data from research papers with conversational language models and prompt engineering. Nat. Commun. **15**(1), 1569 (2024)
28. Poveda-Villalón, M., Gómez-Pérez, A., Suárez-Figueroa, M.C.: Oops!(ontology pitfall scanner!): an on-line tool for ontology evaluation. Int. J. Semant. Web Inf. Syst. (IJSWIS) **10**(2), 7–34 (2014)
29. Reklos, I., Meroño-Peñuela, A.: Medicause: causal relation modelling and extraction from medical publications. In: CEUR Workshop Proceedings, vol. 3184, pp. 1–18. CEUR-WS (2022)
30. Shardlow, M., Batista-Navarro, R., Thompson, P., Nawaz, R., McNaught, J., Ananiadou, S.: Identification of research hypotheses and new knowledge from scientific literature. BMC Med. Inform. Decis. Mak. **18**(1), 1–13 (2018)
31. Snell, J., Swersky, K., Zemel, R.: Prototypical networks for few-shot learning. In: Advances in Neural Information Processing Systems, vol. 30 (2017)

32. Soldatova, L.N., Rzhetsky, A.: Representation of research hypotheses. In: J. Biomed. Semant. **2**, 1–15 (2011). https://doi.org/10.1186/2041-1480-2-S2-S9
33. Swanson, R., Ecker, B., Walker, M.: Argument mining: extracting arguments from online dialogue. In: Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pp. 217–226 (2015)
34. Thompson, W.H., Skau, S.: On the scope of scientific hypotheses. Royal Soc. Open Sci. **10**(8), 230607 (2023)
35. Touvron, H., et al.: Llama 2: open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023)
36. Vargas, H., Garijo, D., Gil, Y.: The scientific questions ontology (2017). https://w3id.org/sqo/1.3.1/, revision: v1.3.1
37. Viera, A.J., Garrett, J.M., et al.: Understanding interobserver agreement: the kappa statistic. Fam. Med. **37**(5), 360–363 (2005)
38. Wang, Y., Yao, Q., Kwok, J.T., Ni, L.M.: Generalizing from a few examples: a survey on few-shot learning. ACM Comput. Surv. (csur) **53**(3), 1–34 (2020)