LLMs Tackle Meta-Analysis: Automating Scientific Hypothesis Generation with Statistical Rigor

Tung-Wei Lin[1[0000−0003−2352−2867]], Runing Yang[2], Zain ul Abdeen[2], Alberto Sangiovanni-Vincentelli[1[0000−0003−1298−8389]], Haibo Huang[2], and Ming Jin[2]

[1] UC Berkeley, USA
{twlin, alberto}@berkeley.edu,
[2] Virgina Tech, USA
{yruning, zabdeen, huang151, jinming}@vt.edu

Abstract. We propose the use of Large Language Models (LLMs) for generating statistically supported hypotheses from scientific literature. We present a two-stage framework that effectively leverages LLMs' capacity to analyze vast literature and extract pertinent information to formulate evidence-based hypotheses. Our method comprises two phases: 1) data extraction via decomposed zero-shot prompting, and 2) hypothesis generation by auto-formulating and solving an optimization problem. We demonstrate this framework in agricultural science, where field data is particularly limited.

1 Introduction

Large Language Models (LLMs) demonstrate useful applications in scientific research, including assisting in chemical synthesis [22] and formulating hypotheses [18]. This paper introduces a novel application of LLMs as scientific research assistants, capable of generating hypotheses with effect size measures (e.g., standardized mean differences (SMD)) and statistical significance based on a corpus of scientific literature. Our methodology leverages a two-stage framework inspired by meta-analysis [8], a method essential for deriving robust statistical conclusions from multiple studies. The proposed approach is particularly relevant given that manual meta-analyses are time-consuming, on average exceeding 1000 hours to complete [3], thereby presenting a substantial bottleneck in scientific progress.

The first stage of our framework utilizes decomposed zero-shot prompting [12] to extract information from full-text articles and the tables within, where valuable numerical data resides. Yet existing LLM-based data extractions for scientific articles focus only on abstracts [11] and simple sentence-level information [17,7,22].

The second stage generates hypotheses, formulated as an optimization problem [20]. The LLM proposes, validates, and refines hypotheses iteratively analogous to a scientist honing theories per experimental observations. This process is

guided by counter-examples [4,15], aiming to improve the reliability of generated hypotheses, which can be further validated by human experts.

This paper addresses the gap in existing literature on hypothesis generation, where they fall short in producing hypotheses with quantifiable measures of effect size and statistical significance [18,20,6,13,5,23]. Our framework enhances LLMs' role in scientific research by improving statistical analysis rigor in hypothesis generation. It diverges from prior methods by fully leveraging LLMs for detailed statistical analysis, effectively tackling complex scientific inquiries.

To demonstrate the effectiveness of our framework, we conduct a case study in agricultural science, utilizing a high-impact Nature publication [16] as benchmark to show LLMs' potential of producing premier research. Our contribution is poised to significantly reduce the time and labor associated with traditional meta-analyses while opening new avenues for LLMs in scientific discovery.

2 Problem Formulation

Given an open-ended research question **Q** with an effect size measure and a scientific corpus **C**, which is manually curated or compiled through automatic criteria screening [14,2], the goal is for the LLM to propose a predicate **A** as the answer to **Q** with statistically significant effect **E** derived from **C**. A predicate is defined as a binary-valued function that outputs True or False given an expression.

Open-ended research questions include inquiries like "*When does caffeine improve muscle strength?*" quantified by SMD [9]. A potential answer **A** to this opend-ended question is "*Is male and is upper body.*" with **E** being SMD= 0.21.

This task stands in direct contrast to close-ended question answering [21] and zero-shot hypothesis generation [18] by demanding not only complex reasoning to generate hypotheses but also statistical validation **E** supported by **C**. We contend that hypotheses generated by LLMs should always be supported by statistical evidence for better communication and validation.

3 Case Study in Agricultural Science

3.1 Preliminaries

We consider **Q** as "**When does zero tillage (ZT) out-yield conventional tillage (CT)?**" and **C** as the references in [16]. Tillage refers to the soil disturbance process before planting. In ZT, this step is omitted, while in CT, tillage is conducted. GPT-4 [1] serves as our LLM of choice to derive **A** and **E**. The effect size is defined as $ln(\frac{\text{ZT yield}}{\text{CT yield}})$, representing the logarithm of the yield ratio between ZT and CT for a *comparable* pair of treatments. A comparable pair of treatments differ in as few variables as possible to reduce confounding (see AppendixA Subtask 7 for an example).

Subgroup analysis aids in examining effect size heterogeneity. A predicate P can be used to characterize a subgroup G_P. Specifically, $P(e)$ =True if and only

Table 1. Rows of reference 103 in extracted and expert-curated database [16]. (* - uncontrolled.) TP, FP, FN are respectively 2,1,0 since the second and third GPT-4 extracted rows are matched. Precision and Recall are 67% and 100%. Accuracies(%) of Crop, Irrigation, Rotation, Residue, and Climate (all mapped to "Dry") are respectively, 50, 100, 100, 0, 100.

ID	Country	Location	Crop Type	Irrigation	Rotation	Residue	CT yield	ZT yield	Effect Size
GPT-4 Extracted Database									
103	Argentina	Anguil Experimental Station	oat + hairy vetch	no	yes	*	970kg ha^{-1}	1500kg ha^{-1}	0.44
103	Argentina	Anguil Experimental Station	corn	no	yes	*	6140kg ha^{-1}	8160kg ha^{-1}	0.28
103	Argentina	Anguil Experimental Station	wheat	no	yes	*	2600kg ha^{-1}	3340kg ha^{-1}	0.25
Expert-Curated Database									
103	Argentina	Semiarid Pampean Region, Anguil Experimental Station	maize	no	yes	retained	6140	8160	0.28
103	Argentina	Semiarid Pampean Region, Anguil Experimental Station	maize	no	yes	retained	2600	3340	0.25

if an element e is in G_P. Therefore, **Q** is answered by finding a predicate such that the corresponding subgroup has a positive average effect size, indicating ZT yield > CT yield when all other variables are controlled for. According to [16], the subgroup with the predicate **A** – **dry climate, no irrigation, retained residue, crop rotation** demonstrates higher ZT yield. The objective of this study is thus using GPT-4 to answer **Q** by extracting information and generating **A**.

3.2 Automated Information Extraction

Given a reference paper and all its tables in HTML, GPT-4 identifies the tables containing numerical yields for effect size calculation, excluding tables with only irrelevant information such as rainfall statistics, short text descriptions arranged as a table, etc. From the n relevant tables $\{T_1, \ldots, T_n\}$, GPT-4 enumerates the tillage treatments verbatim in each table $T_i, i \in [n]$[3], resulting in m treatment names, $\{N_{i,1}, \ldots, N_{i,m}\}$. However, the extracted names are notations with no semantic meaning when out of context. For instance, the extracted name "vT-$\frac{1}{2}$F" in reference 642 stands for hairy vetch + conventional tillage $+\frac{1}{2}$ the recommended fertilizer, where hairy vetch is a type of plant.

Therefore, GPT-4 is prompted to featurize each raw name $N_{i,j}, j \in [m]$ into $f_{i,j}$, extracting all relevant features, including a short description of the treatment, the crop type, irrigated or not, crop rotated or not, residue retained or not, the numerical yields, and finally, the country and locations of the experiment sites, which determine the climate by table lookup of the WorldClim database [24]. These features are automatically compiled by GPT-4 given T_i and the methods section of the reference.

With the extracted and featurized data, GPT-4 generates a list of ℓ comparable pairs $\{T_{i,1}, \ldots, T_{i,\ell}\}$ to minimize confounding. Each $T_{i,k}, k \in [\ell]$ consists of

[3] $[n]$ denotes the set $\{1, 2, \cdots, n\}$

Table 2. Extraction results by GPT-4 averaged across 224 references. (* - uncontrolled.)

	Crop Type	Irrigation	Rotation	Residue	Climate	Avg Precision	Avg Recall
Type	Multi-Class/*	Yes/No/*	Yes/No/*	Retained/Removed/*	Dry/Humid/*	62.18%	72.15%
Avg Accuracy	82.44%	76.66%	69.19%	38.63%	87.72%		

$(f^-_{i,k}, f^+_{i,k})$, representing the ZT and CT treatments. Finally, each featurized name selected for pairing is standardized by GPT-4 according to a predefined JSON template so that each pair $T_{i,k}$ is written as a row for analysis (see Appendix for an example).

Table 1 shows the rows of reference 103 of the extracted and expert-curated database [16]. Note that the number of rows for each reference is variable and unknown, in contrast to simpler data extractions [11]. The effect size of each row is calculated by parsing the numerical values from ZT yield and CT yield.

For each row u produced by GPT-4 of a reference x, we calculate the absolute difference of the effect size with each row v of reference x in the expert-curated database, denoted by $\Delta_u(v)$. If $\min_v \Delta_u(v)$ is smaller than a threshold of 0.01, u is defined to be matched with $\arg\min_v \Delta_u(v)$. One row in the expert-curated database is matched at most once. Furthermore, define true positives (TP) as extracted rows that are matched, false positives (FP) as extracted rows that are not matched, and false negatives (FN) as rows in the expert-curated database not matched by any extracted rows. We also define precision $= \frac{\text{TP}}{\text{TP+FP}}$ and recall $= \frac{\text{TP}}{\text{TP+FN}}$.

For extracted rows that are true positives, we calculate the accuracy of each feature by dividing the number of exact string matches by the number of true positives. To do this, we manually create a set of rules to map semantically equivalent keywords. For instance, ("maize", "corn") for crop type. Except for country and location, the climate is first determined, followed by exact string matching of the climate.

Out of the 610 references, 224 are extracted, with the rest inaccessible, not digitized, or the numerical yields not reported in tables but likely in figures, as indicated by GPT-4's inability to find relevant tables (i.e., $n = 0$). The GPT-4 extracted database has 2604 rows and the expert-curated database has 2737 rows. As shown in Table 2, residue management has the lowest extraction accuracy, often because it is not explicitly stated, requiring inference from the text. Moreover, GPT-4 exhibits a bias towards positively correlating CT treatments with residue removal and ZT treatments with residue retention. It interprets the soil disturbance in CT as a process that incorporates the residue, thus removing it from the surface. This positive correlation is spurious because even after tillage, a significant amount of residue can remain on the surface and farmers may even deliberately retain residue on the surface in certain CT practices.

When $f^+_{i,k}$ and $f^-_{i,k}$ differ on a feature, it indicates an uncontrolled variable between the compared treatments and is indicated by * as shown in Table 1. This can potentially be mitigated by an LLM with more agriculture domain knowl-

edge. The requirement of document-level understanding and domain knowledge to correctly extract residue management illustrates the difficulty of information extraction in academic papers.

3.3 Hypothesis Generation

For hypothesis generation, the goal is to find a predicate P such that the corresponding subgroup G_P has a positive average effect size given a significance level α. This is formulated as an optimization problem where GPT-4 iteratively proposes predicates, maximizing the average effect size of G_P. To account for increased false discoveries in multiple hypothesis testing, we apply Bonferroni correction [10] to guarantee Type I error below α.

LLM as an Optimizer The database S is first partitioned into S^- and S^+, which are rows with effect sizes smaller and greater than 0 respectively. By contrasting S^- with S^+ in the prompt, GPT-4 is asked to propose a predicate P_0 under λ words so that rows satisfying P_0 have a positive average effect size. In particular, λ is set to 15 as a regularization to avoid trivial predicates. GPT-4 is then asked to generate a Python function that takes a row R_i as an argument and returns $P_0(R_i)$, which forms a subgroup $G_0 \coloneqq \{R_i | P_0(R_i) = \text{True}, \forall R_i \in S\}$. P_0 is assessed by calculating the average effect size ES_0 and the p-value p_0. Specifically, ES_0 is estimated through bootstrapping and p_0 is computed by the randomization test.

Average Effect Size. The average effect size ES_G of a subgroup G is estimated through bootstrapping. The procedure goes as follows. Randomly sample a set of rows S_i from G with replacement so that $|S_i| = |G|$. Then calculate the average effect size of the rows in S_i. We get ES_G by repeating this process for q times and averaging the results.

p-Value. After the average effect size of a subgroup G is calculated, its p-value can be calculated using the randomization test. The null hypothesis in the randomization test assumes that there is no difference between ZT yield and CT yield of a row in G, suggesting that the observed ES_G is merely by chance. The alternative hypothesis is that ZT yield > CT yield. To test this, the method randomly inverts the sign of each row's effect size in G. Following this, the average effect size of these sign-flipped values is computed and then compared with ES_G. This process is repeated q times and the proportion of instances where ES_G exceeds the average of these randomized effect sizes serves as the p-value, indicating the statistical significance of ES_G. In our implementation, q is set to 4999.

For $i \geq 0$, GPT-4 is asked to propose a new predicate P_{i+1} aimed at maximizing ES_{i+1} and minimizing p_{i+1}. This is achieved by refining all predicates proposed in the past $P_j, \forall j \leq i$. Toward this goal, we collect negative (CN_j^-) and positive (CN_j^+) counterexamples of P_j, which are respectively rows that satisfy P_j but with an effect size smaller than ES_j, and rows that do not satisfy P_j but with an effect size larger than ES_j. The exclusion of some negative

Algorithm 1 Hypothesis Generation

Inputs: S^-, S^+, t, α, Regularization λ
Output: Final predicate P^*

```
P_0 ← Prompt LLM for predicate within λ words by contrasting S^- with S^+
Calculate G_0, ES_0, p_0
for i = 0, 1, ..., t − 1 do
    Collect CN_i^-, CN_i^+, CR_i^-, CR_i^+
    P_{i+1} ← Prompt LLM for predicate within λ words by looking at P_j, ES_j, p_j
        and contrasting CN_j^-, CN_j^+, CR_j^-, CR_j^+, ∀j ≤ i
    Calculate G_{i+1}, ES_{i+1}, p_{i+1}
end for
α' ← α/t // Bonferroni Correction
P* ← P_k, where k = arg max_{i:p_i ≤ α'} {ES_i}
return P*
```

counterexamples and the inclusion of some positive counterexamples would have increased ES_j. Also, we collect negative (CR_j^-) and positive correct examples (CR_j^+), which are rows with effect sizes smaller than ES_j that do not satisfy P_j and rows with effect sizes larger than ES_j that satisfy P_j. In the prompt, we list the history of proposed predicates P_j, average effect sizes ES_j, and p-values p_j, $\forall j \leq i$, and ask GPT-4 to contrast samples of CN_j^-, CN_j^+ and CR_j^-, CR_j^+ to come up with a refined predicate P_{i+1}. Examples for each j are uniformly sampled to fit within the context window (see Appendix for prompting templates). Once the predefined hypothesis budget t is reached, we apply Bonferroni correction and divide α by t. The predicate with the largest effect size and a p-value below the Bonferroni-adjusted significance level is selected as the final output, as detailed in Algorithm 1.

Theorem 1. *Given P^*, the result of Algorithm 1, and significance level α, the probability that the proposed answer $\boldsymbol{A} = P^*$ is false is bounded by α.*

Proof. Let $G_0, \cdots, G_{t-1}$ be the set of proposed subgroups, $(H0_i, H1_i)$ be the pair of null and alternative hypotheses associated with each subgroup G_i, p_i be the p-value, and ES_i be the average effect size. Given significance level α, the Bonferroni-corrected $\alpha' = \frac{\alpha}{t}$ and predicate P_k, where $k = \arg\max_{i:p_i \leq \alpha'}\{ES_i\}$. Assume that $H0_k$ is true, then the probability that our answer $\mathbf{A} = P_k$ is false is $Pr(p_k \leq \alpha')$.

$$\begin{aligned} Pr(p_k \leq \alpha') &\leq Pr\left(\bigcup_{i=0}^{t-1}(p_i \leq \alpha')\right) \\ &\leq \sum_{i=0}^{t-1} Pr\left(p_i \leq \alpha'\right) \\ &= t\alpha' \\ &= \alpha \end{aligned} \tag{1}$$

Given that Bonferroni correction divides α by t, careful selection of t is crucial to balance the risk of Type II errors.

Table 3. Predicates P_i, average effect size ES_i and p-value p_i of subgroup G_i on GPT-4 extracted and expert-curated databases. P^* for each database is in bold.

	i	P_i	$ES_i(\%)$	p_i
GPT-4 Extracted Database	0_G	**crop rotation is yes and irrigated is no and climate is dry**	**12.98**	**0.0**
	1_G	crop rotation is yes and irrigated is no and climate is dry and residue management is retained	10.14	0.0
	2_G	crop rotation is yes and irrigated is no and climate is dry	12.93	0.0
	3_G	crop rotation is yes and irrigated is no and climate is dry	12.91	0.0
	4_G	crop rotation is yes and irrigated is no and climate is dry	12.93	0.0
Expert Curated Database	0_E	residue management is retained and climate is dry	3.11	0.0168
	1_E	residue management is retained, climate is dry, crop rotation is yes	7.12	0.0
	2_E	residue management is retained, climate is dry, irrigated is no	7.09	0.0004
	3_E	residue management is retained, climate is dry, irrigated is no, crop rotation is yes	13.39	0.0
	4_E	**residue management is retained, climate is dry, irrigated is no, crop rotation is yes**	**13.51**	**0.0**

Results and Discussion In Table 3, we show the predicates proposed during optimization and the selected P^* using our GPT-4 extracted and expert-curated databases. We set the significance level $\alpha = 0.05$ and hypothesis budget $t = 5$. Henceforth, we report back-transformed effect sizes in percentage yield change: $\frac{\text{ZT yield}}{\text{CT yield}} - 1$. Our framework reproduces the answer to the research question mentioned in the preliminaries section when using the expert-curated database (P_{4_E}). Although P_{0_G} does not ultimately reproduce P_{4_E}, it correctly characterizes 3 out of 4 features, except for the residue management. This difference is attributed to the low extraction accuracy of "Residue" as indicated in Table 2. Note that during the optimization on the GPT-4 extracted database, P_{1_G} reproduces P_{4_E} but was discarded due to a smaller average effect size. We thus also tested P_{0_G} on the expert-curated dataset, yielding $ES = 13.03\%$ and $p = 0$. This indicates that there is little heterogeneity in residue retention and removal, the marginalization of which makes little difference.

As an additional experiment, we queried GPT-4 without providing access to GPT-4 extracted or expert-curated database but only the feature names. It proposes *"irrigated, humid climate, residue retained, crop rotation yes"*, which results in $ES = -2.9$ and $p = 0.94$ on the expert-curated database. Since the average effect size of this predicate is negative, this shows the importance of literature access for generating scientifically valid hypotheses (see Appendix for details).

4 Conclusion

We introduced the application of LLMs for addressing open-ended research questions and developed a two-stage framework that effectively leverages statistical evidence, ensuring the Type I error of the generated hypothesis remains under the prescribed threshold. The presented case study illustrated this framework's ability to parse scientific literature and formulate evidence-based hypotheses. This paper underscores the promising role of LLMs in scientific inquiry, encour-

aging further exploration and adoption of automated approaches across various disciplines.

5 Limitations

While our study demonstrates GPT-4's capability of correctly answering the question, "When does ZT outyield CT?" using an expert-curated database, it's important to note certain limitations. First, we cannot verify whether GPT-4's pre-training dataset includes our benchmark [16], as it does not provide token probabilities for membership inference [19]. Additionally, biases within GPT-4, as discussed in the Automated Information Extraction section and the Results and Discussion section, could lead to biased or unfounded hypotheses. Finally, the effectiveness of our approach in domains other than agricultural science remains uncertain, calling for further validation to confirm GPT-4's domain-specific knowledge and hypothesis generation capability across different fields.

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Alchokr, R., Borkar, M., Thotadarya, S., Saake, G., Leich, T.: Supporting systematic literature reviews using deep-learning-based language models. In: Proceedings of the 1st International Workshop on Natural Language-based Software Engineering. pp. 67–74 (2022)
3. Allen, I.E., Olkin, I.: Estimating time to conduct a meta-analysis from number of citations retrieved. Jama **282**(7), 634–635 (1999)
4. Alur, R., Bodik, R., Juniwal, G., Martin, M.M., Raghothaman, M., Seshia, S.A., Singh, R., Solar-Lezama, A., Torlak, E., Udupa, A.: Syntax-guided synthesis. IEEE (2013)
5. Banker, S., Chatterjee, P., Mishra, H., Mishra, A.: Machine-assisted social psychology hypothesis generation (2023)
6. Ciucă, I., Ting, Y.S., Kruk, S., Iyer, K.: Harnessing the power of adversarial prompting and large language models for robust hypothesis generation in astronomy. arXiv preprint arXiv:2306.11648 (2023)
7. Gilligan, L.P., Cobelli, M., Taufour, V., Sanvito, S.: A rule-free workflow for the automated generation of databases from scientific literature. arXiv preprint arXiv:2301.11689 (2023)
8. Glass, G.V.: Primary, secondary, and meta-analysis of research. Educational researcher **5**(10), 3–8 (1976)
9. Grgic, J., Trexler, E.T., Lazinica, B., Pedisic, Z.: Effects of caffeine intake on muscle strength and power: a systematic review and meta-analysis. Journal of the International Society of Sports Nutrition **15**(1), 11 (2018)
10. Haynes, W.: Bonferroni correction. Encyclopedia of systems biology pp. 154–154 (2013)

11. Kartchner, D., Ramalingam, S., Al-Hussaini, I., Kronick, O., Mitchell, C.: Zero-shot information extraction for clinical meta-analysis using large language models. In: The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks. pp. 396–405 (2023)
12. Khot, T., Trivedi, H., Finlayson, M., Fu, Y., Richardson, K., Clark, P., Sabharwal, A.: Decomposed prompting: A modular approach for solving complex tasks. arXiv preprint arXiv:2210.02406 (2022)
13. Koneru, S., Wu, J., Rajtmajer, S.: Can large language models discern evidence for scientific hypotheses? case studies in the social sciences. arXiv preprint arXiv:2309.06578 (2023)
14. Kusa, W., Mendoza, O.E., Samwald, M., Knoth, P., Hanbury, A.: Csmed: Bridging the dataset gap in automated citation screening for systematic literature reviews. arXiv preprint arXiv:2311.12474 (2023)
15. Madaan, A., Tandon, N., Gupta, P., Hallinan, S., Gao, L., Wiegreffe, S., Alon, U., Dziri, N., Prabhumoye, S., Yang, Y., et al.: Self-refine: Iterative refinement with self-feedback. arXiv preprint arXiv:2303.17651 (2023)
16. Pittelkow, C.M., Liang, X., Linquist, B.A., Van Groenigen, K.J., Lee, J., Lundy, M.E., Van Gestel, N., Six, J., Venterea, R.T., Van Kessel, C.: Productivity limits and potentials of the principles of conservation agriculture. Nature **517**(7534), 365–368 (2015)
17. Polak, M.P., Morgan, D.: Extracting accurate materials data from research papers with conversational language models and prompt engineering–example of chatgpt. arXiv preprint arXiv:2303.05352 (2023)
18. Qi, B., Zhang, K., Li, H., Tian, K., Zeng, S., Chen, Z.R., Zhou, B.: Large language models are zero shot hypothesis proposers. arXiv preprint arXiv:2311.05965 (2023)
19. Shi, W., Ajith, A., Xia, M., Huang, Y., Liu, D., Blevins, T., Chen, D., Zettlemoyer, L.: Detecting pretraining data from large language models. arXiv preprint arXiv:2310.16789 (2023)
20. Yang, Z., Du, X., Li, J., Zheng, J., Poria, S., Cambria, E.: Large language models for automated open-domain scientific hypotheses discovery. arXiv preprint arXiv:2309.02726 (2023)
21. Zheng, C., Zhou, H., Meng, F., Zhou, J., Huang, M.: Large language models are not robust multiple choice selectors. arXiv e-prints pp. arXiv–2309 (2023)
22. Zheng, Z., Zhang, O., Borgs, C., Chayes, J.T., Yaghi, O.M.: Chatgpt chemistry assistant for text mining and prediction of mof synthesis. arXiv preprint arXiv:2306.11296 (2023)
23. Zhong, R., Zhang, P., Li, S., Ahn, J., Klein, D., Steinhardt, J.: Goal driven discovery of distributional differences via language descriptions. arXiv preprint arXiv:2302.14233 (2023)
24. Zomer, R.J., Xu, J., Trabucco, A.: Version 3 of the global aridity index and potential evapotranspiration database. Scientific Data **9**(1), 409 (2022)

A Information Extraction Example

In the following, we show the prompts and responses in extracting reference 642 in [16].

Subtask 1. This template is applied to all 4 tables in reference 642 to prune those with no numerical yields. We show the prompt and response of table 1. The words in brown are replaced with the corresponding tables and abstract of the reference.

Prompt: You read the abstract of a paper and one of its tables at a time to check if the table satisfies the following condition. Does the table contain the actual numerical yields (unit: weight/area) of at least one zero/no tillage treatment and one with-tillage treatment? Note that sometimes zero tillage is equivalently called no-till and with-tillage treatment is equivalently called conventional tillage or simply tillage, in contrast with the conservational agriculture practice. Zero tillage is sometimes abbreviated as ZT or NT, and with-tillage treatment is sometimes abbreviated as CT, standing for conventional tillage. Note that some tables only contain the statistical analyses (e.g. root mean square error) of the yields but not the actual yields. Pay attention to the caption for the nature of the table. If the actual yields of zero tillage and with-tillage are in the table, list the treatment names verbatim in a JSON object. e.g.,
`{‘methods’:[NT, MM, MR, MC, CT]}.`
If not, output an empty list, i.e.,
`{‘methods’: []}`
Table 1 Effect of the various treatments on weed cover rate and weed dry biomass <...**omitted**...>
Abstract Using mulch to tackle the problems of soil <...**omitted**...>

Response:
`{‘methods’: []}`

Subtask 2. This template is applied to the result of subtask 1, which resulted in a non-empty list only on table 2 after running on all 4 tables. That is, $n = 1$ and the list of relevant tables is $[T_2]$. Therefore, we show the prompt and response of only table 2 in this subtask. The words in brown are replaced with the corresponding table when used for other references.

Prompt: List the name of the tillage treatments in the table verbatim if their numerical yield data are available. Output a JSON object
`{‘methods’:[treatment1, treatment2, ...]}`
Note that some tables contain only numerical analyses of the yield data instead of the actual yield data. In this case, output a JSON object with an empty list
`{‘methods’:[]}`
Table 2 Effect of treatments on N input, N uptake and N use efficiency <...**omitted**...>

Response:

```
{'methods':['vT-1/2F', 'vZ-1/2F', 'vZ-0F', 'fT-1F', 'fZ-1F',
     'fT-0F']}
```

Subtask 3. This template is applied to the result of subtask 2. We show the prompt and response of featurizing the first treatment, 'vT-1/2F'. The words in brown are replaced accordingly.

Prompt: Given the provided text and table, answer the following for each treatment of 'vT-1/2F' in the table. Ignore the following methods 'vZ-1/2F', 'vZ-0F', 'fT-1F', 'fZ-1F', 'fT-0F'.
(i) brief description of the treatment as described in the text
(ii) weights of yields as shown in the table. Ignore other metrics.
(iii) crop type whose yield is measured.
(iv) crop part whose yield is measured.
Use a dictionary if the table contains multiple treatments of 'vT-1/2F'. e.g.,

```
{'tillage description': 'answer',
 'treatment names':{'X', 'Y'},
 'yields': {'X': {'2002': '5.3 Mg ha -1', 'crop type':'ans',
                  'crop part':'ans'},
            'Y': {'2002': '4.5 Mg ha -1', 'crop type':'ans',
                  'crop part':'ans'}}}
```

Sometimes, yields of individual years are unavailable. Only average yields are available, in that case, use avg (year start-year end) to denote. e.g.,

```
{'tillage description': 'answer',
 'treatment names':{'X'},
 'yields': {'X': {'avg(1995-2000)': '1.9 Mg ha -1',
                  'crop type':'ans', 'crop part':'ans'}}}
```

Replace 'X', 'Y' with actual treatment names and 'ans' with answers. Format your answer in JSON.
Table 2 effect of treatments on N input, N uptake and N use efficiency <...**omitted**...>
MATERIAL AND METHODS Site description and experimental design The present <...**omitted**...>

Response:

```
{'tillage description': 'Hairy vetch with conventional
    tillage and half the recommended fertilization',
 'treatment names': {'vT-1/2F': {'yields': {'vT-1/2F':
     {'2005': '5.038 Mg ha -1', 'crop type': 'maize', 'crop
     part': 'kernel'}}}}}
```

Subtask 4. This template is applied to the result of subtask 3. We show the prompt and response of continuing featurization of 'vT-1/2F'. Residue management requires more instructions due to its complexity. The words in brown are replaced accordingly.

Prompt: Given a JSON object, which is extracted from the given text and table, answer whether residue is retained or not for each treatment of 'vT-1/2F' in the JSON object. Ignore the following methods 'vZ-1/2F', 'vZ-0F', 'fT-1F', 'fZ-1F', 'fT-0F'. Use a dictionary if the table contains multiple treatments of 'vT-1/2F'. e.g.,

```
{'residue': {'treatment1': 'retained',
             'treatment2': 'removed'},
 'reasoning':'answer'}.
```

Reason by considering the following questions. Is there cover crop or crop rotation? What is the weed control strategy? What is the treatment of the residue of crop and weed? If burned, incorporated into the soil, grazed, or removed, then residue is removed. If left on the surface, then residue is retained. Is there artificial or natural mulching? Mulching indicates retention, while explicit statement of no-mulch indicates removal. Important note: conventional tillage does NOT always incorporate residue into the soil. It is possible to retain residue for conventional tillage. Answer 'not specified' if not inferrable'

{'tillage description': 'Hairy vetch with conventional tillage <...**omitted**...>

Table 2 effect of treatments on N input, N uptake and N use efficiency <...**omitted**...>

MATERIAL AND METHODS Site description and experimental design The present <...**omitted**...>

Response:

```
{'residue': {'vT-1/2F': 'removed'},
  'reasoning': 'The treatment 'vT-1/2F' involved hairy vetch
       with conventional tillage and half the recommended
      fertilization. According to the text, for the 'vT-1/2F
      ' treatment, the hairy vetch biomass was incorporated
      through tillage before maize sowing. Incorporation of
      the biomass into the soil through tillage indicates
      that the residue is not left on the surface but is
      instead mixed into the soil, which means the residue
      is removed.'}
```

Subtask 5. This template is applied to the result of subtask 3. We show the prompt and response of continuing featurization of 'vT-1/2F'. The words in brown are replaced accordingly.

Prompt: Given a JSON object, which is extracted from the given text and table, answer the following for each treatment of 'vT-1/2F' mentioned in the JSON object. Ignore the following methods 'vZ-1/2F', 'vZ-0F', 'fT-1F', 'fZ-1F', 'fT-0F'. Use a dictionary if there are multiple treatments, e.g.
`'irrigated': {'site A': 'yes', 'site B': 'no'}.`
(i) Irrigated or not (identify first whether irrigation is described. If not, but rainfall is stated, answer no since it is probably rainfed; otherwise, 'not specified' if not inferrable).
(ii) Crop rotation or not (identify first whether crop rotation is described. If there is no mention of cover crops or crop rotation, there is probably no rotation).
(iii) Location.
(iv) Country.
{'tillage description': 'Hairy vetch with conventional tillage <...**omitted**...>
Table 2 Effect of treatments on N input, N uptake and N use efficiency <...**omitted**...>
MATERIAL AND METHODS Site description and experimental design The present <...**omitted**...>

Response:
```
{'vT-1/2F': {'irrigated': 'no', 'crop rotation': 'yes',
    'location': 'Ishigaki Island', 'country': 'Japan'}}
```

Subtask 6. This template is applied to the result of subtask 3, 4, and 5. We show the prompt and response of merging the features of 'vT-1/2F'. The structures and keys of the resultant JSON objects from subtasks 3, 4, and 5 are variable, making it impossible to merge with a hard-coded program. The words in brown are replaced according to the treatment being featurized.

Prompt: Merge these JSON objects into one JSON object ignoring the reasoning fields:
```
{'tillage description': 'Hairy vetch with conventional
    tillage and half the recommended fertilization',
 'treatment names': {'vT-1/2F': {'yields': {'vT-1/2F':
     {'2005': '5.038 Mg ha -1', 'crop type': 'maize', 'crop
     part': 'kernel'}}}}}
```

```
{‘residue’: {‘vT-1/2F’: ‘removed’}, ‘reasoning’: ‘The
    treatment ‘vT-1/2F’ involved hairy vetch with
    conventional tillage and half the recommended
    fertilization. According to the text, for the ‘vT-1/2F’
    treatment, the hairy vetch biomass was incorporated
    through tillage before maize sowing. Incorporation of
    the biomass into the soil through tillage indicates that
     the residue is not left on the surface but is instead
    mixed into the soil, which means the residue is removed
    .’}
{‘vT-1/2F’: {‘irrigated’: ‘no’, ‘crop rotation’: ‘yes’,
    ‘location’: ‘Ishigaki Island’, ‘country’: ‘Japan’}}
```

Response:

```
{‘tillage description’: ‘Hairy vetch with conventional
    tillage and half the recommended fertilization’,
  ‘treatment names’: {‘vT-1/2F’: {‘yields’: {‘vT-1/2F’:
      {‘2005’: ‘5.038 Mg ha -1’, ‘crop type’: ‘maize’, ’crop
       part’: ‘kernel’}}}},
  ‘residue’: {‘vT-1/2F’: ‘removed’},
  ‘vT-1/2F’: {‘irrigated’: ‘no’,
    ‘crop rotation’: ‘yes’,
    ‘location’: ‘Ishigaki Island’,
    ‘country’: ‘Japan’}}
```

Subtask 7. This template is applied to the result of subtask 6. Once all treatments [‘vT-1/2F’, ‘vZ-1/2F’, ‘vZ-0F’, ‘fT-1F’, ‘fZ-1F’, ‘fT-0F’] are fully featurized, the following prompt pairs comparable treatments to minimize confounding. The words in brown are replaced accordingly.

Prompt: You are conducting a meta-analysis on the effects of tillage on yields. You read the extracted experiment settings from a paper and:
Pick exactly one tillage method representing the least soil disturbance (typically no-till or zero till), named T1 in the following, and one tillage method representing the greatest soil disturbance, (typically conventional tillage), named T2 in the following. Ignore other reduced tillage methods, such as minimum tillage, moderate tillage, etc. Do not count them.
T1 and T2 are paired so that they have as many similar variables as possible excluding the yields. There may be multiple pairs. Output in JSON your reasoning and the NUMERICAL indices of pairings.
Importantly, an index is only allowed to be paired once hence some variations might be left unpaired. For example, given

```
[index 0: no-till without mulch, index 1: reduced tillage
    with mulch,
 index 2: CT with mulch, index 3: CT without mulch]
```
i.e.,
```
{'reasoning':'No-till and CT are picked as T1 and T2. Then,
    variations are identified, i.e., 0, 2, and 3. However,
    since indices are only allowed to be paired once, only
    (0, 3) is produced', 'pairs':[{'T1':0, 'T2':3}]}.
```
A counterexample, given
```
[index 0: ZT, zero tillage, more info: {ZT_mulch,
    ZT_no_mulch},
 index 1: CT, conventional tillage], more info: {CT_mulch,
     CT_no_mulch}.
```
Output
```
{'reasoning': 'answer',
 'pairs':[{'T1':'ZT_mulch', 'T2':'CT_mulch'},
          {'T1':'ZT_no_mulch', 'T2':'CT_no_mulch'}]}.
```
This is disallowed because the values of 'T1', 'T2' must be numerical indices. Instead, output
```
{'reasoning': 'answer', 'pairs':[{'T1':0, 'T2':1}]}
```
Index 0: {'tillage description': 'Hairy vetch with conventional tillage and half the <...**omitted**...>
Index 1: {'tillage description': 'Hairy vetch with zero tillage and half the <...**omitted**...>
<...**omitted**...>
Index 5: {'tillage description': 'Natural fallow with conventional tillage and <...**omitted**...>

Response:
```
{'reasoning': 'T1 is chosen as the tillage method with the
    least soil disturbance, which is zero tillage (ZT), and
    T2 is chosen as the tillage method with the greatest
    soil disturbance, which is conventional tillage (CT).
    The pairings are made by matching the treatments with
    similar variables such as fertilization levels, residue
    management, irrigation, crop rotation, location, and
    country, while excluding the yields. The pairs are (1,
    0) and (4, 3) because they have matching variables
    except for the tillage method and yields. Index 2 and 5
    are not paired because there are no corresponding ZT
    treatments with no fertilizer supply to match with CT
    treatments.',
  'pairs': [{'T1': 1, 'T2': 0}, {'T1': 4, 'T2': 3}]}
```

Subtask 8. This template is applied to the result of subtask 7. All treatments selected for pairing are standardized by this subtask. We show the standardization of 'vT-1/2F'. The words in brown are replaced accordingly with the treatment to be standardized.

Prompt: Standardize the following experiment setting into a list of JSON objects with fixed and predefined fields. The value of the 'yields' field is a single-level dictionary, in which the numerical yields are values indexed by keys. Ignore other metrics than yields. All fields correspond to one value except for 'yields', which corresponds to a single-level dictionary. Therefore, sometimes it is required to split the setting into multiple JSON object(s) to follow the format when location, crop type, residue management, etc., have multiple values. For example, a setting with locations Iowa and Wisconsin is split into two JSON objects as follows,

```
{'objects':[{'tillage method': 'ZT',
 'residue management': 'removed', 'irrigated':'no',
 'crop rotation':'yes', 'crop type':'wheat',
 'crop part':'grain', 'location':'Iowa', 'country':'USA',
 'yields':{'1999':'3000 kg ha -1', '2000':'3100 kg ha -1'}},
{'tillage method': 'ZT',
 'residue management': 'removed', 'irrigated':'no',
 'crop rotation':'yes', 'crop type':'wheat',
 'crop part':'grain','location':'Wisconsin','country':'USA',
 'yields':{'1999':'3300 kg ha -1', '2000':'3000 kg ha -1'}}]
 }
```

On the other hand, the following is disallowed,

```
{'objects':[{'tillage method': 'ZT', 'residue management': '
    removed',
   'irrigated':'no', 'crop rotation':'yes',
   'crop type':'wheat', 'crop part':'grain',
   'location':'Iowa and Wisconsin', 'country':'USA',
   'yields':{'1999':{'Iowa':'3000 kg ha -1',
                     'Wisconsin':'3300 kg ha -1'},
             '2000':{'Iowa':'3100 kg ha -1',
                     'Wisconsin':'3000 kg ha -1'}}]}}
```

Because 'location' corresponds to two values and 'yields' corresponds to a two-level dictionary.

Experiment setting to standardize:

{'tillage description': 'Hairy vetch with conventional tillage and half the

<...**omitted**...>

Response:

```
{'objects': [{'tillage method': 'conventional',
       'residue management': 'removed',
       'irrigated': 'no', 'crop rotation': 'yes',
```

```
        'crop type': 'maize', 'crop part': 'kernel',
        'location': 'Ishigaki Island', 'country': 'Japan',
        'yields': {'2005': '5.038 Mg ha -1'}}]}
```

Subtask 9. This subtask is achieved by a hard-coded python program that writes a pair of standardized JSON objects into a row. We show the example of the paired treatments ('vZ-1/2F', 'vT-1/2F').

Input:

```
{'objects': [{'tillage method': 'zero tillage',
        'residue management': 'retained',
        'irrigated': 'no', 'crop rotation': 'yes',
        'crop type': 'maize', 'crop part': 'kernel',
        'location': 'Ishigaki Island', 'country': 'Japan',
        'yields': {'2005': '4.215 Mg ha -1'}}]}
{'objects': [{'tillage method': 'conventional',
        'residue management': 'removed',
        'irrigated': 'no', 'crop rotation': 'yes',
        'crop type': 'maize', 'crop part': 'kernel',
        'location': 'Ishigaki Island', 'country': 'Japan',
        'yields': {'2005': '5.038 Mg ha -1'}}]}
```

Output:

```
{'2005' :
    ['country': 'Japan', 'location': 'Ishigaki Island',
      'residue management': '*', 'irrigated': 'no',
      'crop rotation': 'yes', 'crop type': 'maize',
      'ZT yield': '4.215 Mg ha -1', 'CT yield': '5.038 Mg ha
          -1']}
```

B Hypothesis Generation Template

In the following, we show the templates of the prompts for hypothesis generation.

Template 0. This template prompts GPT-4 for an initial predicate when no predicate has been proposed. Experiments are sampled until the context window fills up as shown in brown.

Prompt: You are investigating why in some scenarios; zero tillage yields more than conventional tillage. You look across different experiments published in the literature to generate one predicate 'P' that defines the subgroup with maximum mean effect size and minimum p-value. The effect size is defined as the natural log of the ratio of zero tillage yield to conven-

tional tillage yield. The p-value is calculated using randomization test. In the following, the experiment settings and their effect sizes are provided in two distinct groups. In group 1, each experiment has an effect size greater than 0. In group 2, each experiment has an effect size smaller than 0. Formulate your predicate P so that the average effect size of the set of experiments satisfying P is maximized and its p-value minimized. You should output one JSON object without explanations. The following are some examples.
Example 1. {'P':'fertilization is applied'} (This means that the experiments satisfying 'fertilization is applied' has the maximum effect size compared with other predicates and with minimum p value.)
Example 2. {'P':'crop type is yam'} (This means that the experiments satisfying 'crop type is yam' has the maximum effect size compared with other predicates and with minimum p value.)
Importantly, P must be within 15 words. Note that all experiments compare yields of zero-tillage with conventional tillage. Therefore, whether zero-tillage is applied should not be a part of your proposed predicate.
Group 1: experiments with positive effect size

```
{'country': 'Spain', 'location': 'Cordoba', 'crop type': '
    bean', ...
{'country': 'Nigeria', 'location': 'Owo', 'crop type': '
    cocoyam', ...
```

<...**omitted**...>
Group 2: experiments with negative effect size

```
{'country': 'Brazil', 'location': 'Parana', 'Londrina', '
    Embrapa Soybean' ...
{'country': 'Chile', 'location': 'Antumapu', 'crop type': '
    wheat', ...
```

<...**omitted**...>

Template 1. This template prompts GPT-4 for predicate P_2 given predicates P_0 and P_1. The brown words are replaced accordingly. The prompt for $P_i, i > 2$ is constructed similarly by appending descriptions of $P_j, j < i$ and the corresponding samples of correct examples and counterexamples.

Prompt: You are investigating why in some scenarios; zero tillage yields more than conventional tillage. You look across different experiments published in the literature to generate one predicate 'P' that defines the subgroup with maximum mean effect size and minimum p-value. The effect size is defined as the natural log of the ratio of zero tillage yield to conventional tillage yield. The p-value is calculated using randomization test. Your initial predicate was P=P_0. The mean effect size of the experiments satisfying P is ES_0, and the p-value is p_0.

Correct examples and counterexamples of P_0

The following are experiments with effect sizes greater than ES_0 and do satisfy P=P_0, which means the inclusion of these experiments increases the mean.

```
{'country': 'China', 'location': 'Gansu, Dingxi
    Experiment ...
```

<...**omitted**...>

The following are experiments with effect sizes smaller than ES_0 and do not satisfy P=P_0, which means the exclusion of these experiments increases the mean.

```
{'country': 'Argentina', 'location': 'Buenos Aires
    province ...
```

<...**omitted**...>

The following are counterexamples with effect sizes greater than ES_0 but do not satisfy P=P_0.

```
{'country': 'USA', 'location': 'Texas, Etter ...
```

<...**omitted**...>

The following are counterexamples with effect sizes smaller than ES_0 but do satisfy P=P_0.

```
{'country': 'USA', 'location': 'Montana, Culbertson
    '...
```

<...**omitted**...>

Then, you refined P to become P_1. The mean effect size of the experiments satisfying P is ES_1, and the p-value is p_1.

Correct examples and counterexamples of P_1

<...**omitted**...>

To maximize the mean and minimize the p-value further, can you propose a refined predicate considering the counterexamples mentioned? Importantly, the refined predicate should not be longer than 15 words. Note that all experiments compare yields of zero-tillage with conventional tillage. Therefore, whether zero-tillage is applied or not should not be a part of your proposed predicate. Output in JSON format as follows. {'P':'<your refined predicate>'}

Template 3. This template prompts GPT-4 to translate a natural-language predicate P into a python function. The words in brown are replaced according to the predicate being translated.

Prompt: You translate a natural language predicate P into a Python lambda function whose input is a dictionary and output is the evaluation of P.
The possible keys in the input dictionary include the following. However, a key may be absent, so please check whether the key is in the dictionary: 'country', 'location', 'crop type', 'climate', 'crop rotation', 'residue management', and 'irrigated'
For example, P='crop type is maize' and the corresponding Python lambda function is

```
lambda obj: True
       if 'crop type' in obj and obj['crop type'] == 'maize'
            else False
```

Translate the following predicate and output in JSON: {'lambda function':'answer'}
crop rotation is yes and irrigated is no and climate is dry

C Hypothesis Generation Without Data

The following shows the prompt of asking GPT-4 to generate a hypothesis without access to any databases.

Prompt: You are analyzing in what situation, zero tillage yields more than conventional tillage.
The following are the variables considered.
Country: multi-class, uncontrolled
Location: multi-class, uncontrolled
Crop type: multi-class, uncontrolled
Irrigated: yes, no, uncontrolled
Climate: humid, dry, uncontrolled
Residue: retained, removed, uncontrolled
Crop rotation: yes, no, uncontrolled
Provide a predicate P such that zero tillage yields more than conventional tillage if and only if the predicate is true. P is less than 15 words.
Format your answer in JSON: {'P':'answer'}

Response:

```
{'P': 'Irrigated, humid climate, residue retained, crop
    rotation yes'}
```