Large Language Models for Automated Open-domain Scientific Hypotheses Discovery

Anonymous ACL submission

Abstract

Hypothetical induction is recognized as the main reasoning type when scientists make observations about the world and try to propose hypotheses to explain those observations. Past research on hypothetical induction is under a constrained setting: (1) the observation annotations in the dataset are carefully manually handpicked sentences (resulting in a close-domain setting); and (2) the ground truth hypotheses are mostly commonsense knowledge, making the task less challenging. In this work, we tackle these problems by proposing the first NLP dataset for social science academic hypotheses discovery, consisting of 50 recent top social science publications; and a raw web corpus that contains enough information to make it possible to develop all the research hypotheses in the 50 papers. The final goal is to create systems that automatically generate valid, novel, and helpful scientific hypotheses, given only a pile of raw web corpus. Different from the previous settings, the new dataset requires (1) using open-domain data (raw web corpus) as observations; and (2) proposing hypotheses even new to humanity. A multi-module framework is developed for the task, as well as three different feedback mechanisms that empirically show performance gain over the base framework. Finally, our framework exhibits superior performance in terms of both GPT-4 based evaluation and expert-based evaluation.

1 Introduction

Logical reasoning is central to human cognition (Goel et al., 2017). It is widely recognized as consisting of three components, which are deductive, inductive, and abductive reasoning (Yang et al., 2023b). Hypothetical induction is considered to be an important sub-type of inductive reasoning (Norton, 2003). It is recognized as the main reasoning type when scientists make observations about the world and try to propose hypotheses to explain the observations. For example, the proposal of Geocentrism, Heliocentrism, and Newton's law of universal gravitation based on the observations of the motion of (celestial) objects can be seen as a result of hypothetical induction. Hypothetical induction is a process of knowledge exploration from observations to hypotheses: it is challenging because it involves the exploration of knowledge that is even new to humanity.

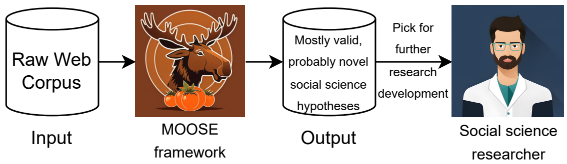

Figure 1: Overview of the new task setting of hypothetical induction and the role of the MOOSE framework.

The latest research on hypothetical induction (Yang et al., 2022b) has two main limitations. Firstly, the observations in their dataset have already been manually selected from the raw web corpus, resulting in a close-domain setting. As a result, a developed system for this dataset relies on already manually selected observations, and cannot utilize the vast raw web corpus to propose hypotheses. Secondly, the ground truth hypotheses are mostly commonsense knowledge (e.g., Newton's law), making the task less challenging since LLMs might have already seen them during pretraining.

To this end, we propose a new task setting of hypothetical induction, which is to generate novel and valid research hypotheses targeting being helpful to researchers while only given (vast) raw web corpus (Figure 1)[1]. This hypothesis formation process is seen as the first step for scientific discovery (Wang et al., 2023a). We call this task as "au**TO**mated open-do**MA**in hypo**T**hetical inducti**O**n (TOMATO)". It is "automated" since a method for this task should automatically propose

[1]Dataset and Code available at https://anonymous.4open.science/r/TOMATO/.

hypotheses with few human efforts; It is open-domain since it is not restricted by any manually collected data. For the TOMATO task, we constructed a dataset consisting of 50 recent social science papers published after January 2023 in top social science journals. For each paper, social science experts collect its main hypothesis, identify its background and inspirations, find semantically similar contents for its background and inspirations from the web corpus, collect the full passage for each matched content, and use all collected web passages as raw web corpus. Although the new dataset involves many manual selection processes, the manually selected contents are used more as benchmarking human performance for comparison. In the TOMATO task, a method is required to only utilize the raw web corpus in the dataset to propose hypotheses. In addition, the raw web corpus is mostly from common news, Wikipedia, and business reviews, which means it can easily expand in scale without much human involvement.

To tackle the TOMATO task, we develop a multi-module framework called MOOSE based on large language model (LLM) prompting (Figure 4). To further improve the quality of the generated hypotheses, we also propose three different feedback mechanisms (present-feedback, past-feedback, and future-feedback) to use LLMs to retrospect and improve the LLM-generated hypotheses for better quality. For present-feedback, the intuition is that, for some modules, their generation can be evaluated by other LLMs and be provided with feedback, which can be utilized by the modules to refine their generation by taking the feedback and previous generation as input and generating again. Some modules can have feedback instantly after their generation to improve themselves. But just like the reward mechanism in reinforcement learning, some rewards (feedback) might be hard to obtain instantly, but need to wait for feedback for a future module. Similarly, we develop past-feedback where a module can benefit from the feedback for a future module. The last one is future-feedback, where a current module can provide justifications for the current module's generation to help a future module's generation, or can provide some initial suggestions which a future module can build upon to further provide more in-depth generation.

For both GPT-4 (OpenAI, 2023) evaluation and social science expert evaluation, our experiment indicates that our framework performs better than an LLM (Ouyang et al., 2022) based baseline, and each of the three feedback mechanisms can progressively improve the base framework. During human analysis, many hypotheses generated by our framework are recognized by social science researchers to be valid, novel, and helpful in the same time.

2 Related Work

2.1 NLP Methods for Scientific Discovery

Zhong et al. (2023) propose a dataset where each data consists of a research goal, a corpus pair, and a language-described discovery. However, (1) their task needs a human-provided research goal and a pair of corpus for discovery, which is not an automated setting and has a limited application scope; (2) the ground truth discovery is not from recent publications. Wang et al. (2023b) is a concurrent work of ours, proposing an automatic method to collect NLP publications to construct a dataset, and a method to propose hypotheses in the NLP domain. However, (1) their task needs humans to input seed terms and background context, which is not an automated setting; (2) their dataset is not manually collected, and their background text and seed terms are collected in the same paper which proposes the ground truth hypothesis, which might cause data contamination problem; (3) their dataset is composed of ACL anthology papers before 2021, so the papers in the dataset are likely to appear in the training corpus of ChatGPT as well as LLaMA-based models (Touvron et al., 2023); (4) their method does not leverage feedback mechanism and is not specifically designed to propose novel hypotheses. Bran et al. (2023) focuses on integrating computational tools in the chemistry domain, but not on providing novel chemistry findings or hypotheses. Boiko et al. (2023) focuses on using LLMs to design, plan, and execution of scientific experiments, but not on finding novel hypotheses.

2.2 LLM-based Self Feedback

Self-refine (Madaan et al., 2023) is a concurrent work, but it only focuses on present-feedback (our framework also proposes past-feedback and future-feedback), and it is not specially designed for inductive reasoning tasks. Other similar works to self-refine (Press et al., 2022; Peng et al., 2023; Yang et al., 2022a; Shinn et al., 2023) also only focus on present-feedback, and their feedback is not multi-aspect nor iterative compared to ours. Our present-feedback is developed upon a multi-aspect

Hypothesis 2. *Customers whose preceding customers use FR payment technology are more likely to use FR payment technology than those whose preceding customers do not use FR payment technology.*

Figure 2: A selected hypothesis in a social science publication collected in our dataset.

2. Hypothesis Development **2.2. Herding Effect**

Figure 3: Hypothetical development section and a particular theory subsection for developing hypotheses.

over-generate-then-filter mechanism (Yang et al., 2022b). However, they only utilize LLMs to "filter" but not to provide feedback.

3 Dataset Collection

In this section, we take one publication (Gao et al., 2023) in our dataset as an example to illustrate the dataset collection process. In total, there are 50 papers published after January 2023. Table 1 shows the statistics of the subject distribution.

Most social science publications highlight their hypotheses. Figure 2 shows our selected main hypothesis in the example publication. The research backgrounds are given in the introduction section. In this example paper, the background is about facial recognition payment technology's usage in society. Most social science publications also have a "Hypothesis Development" section (some may call it by other names, e.g., "Theoretical Development"). For example, the left part ("Hypothesis Development") in Figure 3 shows the title of this section in the example paper. In this section, several theories used to develop the main hypothesis are separately introduced. Usually, each theory takes one subsection. For example, the right part ("Herding Effect") in Figure 3 shows the title of a subsection, which is a particular theory being used as an inspiration, which with the background can develop the hypothesis in Figure 2.

For each publication in our dataset, we identify its main hypothesis, research background, and inspirations, where the background and inspirations together provide enough information to be possible to develop the hypothesis. We also abstract the reasoning process from background and inspirations to hypothesis and note it down for each publication in our dataset. In this selected example, the reasoning process is easy, but it has medium difficulty for researchers to associate the inspiration (herding effect) to the background. For each publication, we include an expert-evaluated complexity for both the reasoning process and the association of the inspiration to the background (details in §A.3).

Social Science	Communication	5
	Psychology	7
Business	Human Resource Management	8
	Information System	8
	International Business	5
	Management	6
	Marketing	11

Table 1: Statistics of subject distribution of the dataset.

Instead of directly copying the background and inspirations from the paper to construct the dataset, we try to find semantically similar text contents from the web corpus as a substitution to avoid data contamination and fit the requirement of TOMATO task that a system should propose novel and valid research hypotheses only given raw web corpus. In the example paper, we find news sentences reporting the usage of facial recognition payment as ground truth background and a Wikipedia description of the herding effect as ground truth inspiration. We also collect the web link and the full text of the manually selected web passages for backgrounds and inspirations to be used as raw web corpus.

In addition, we collect the link and the publication date for all publications in the dataset. We also collected fourteen survey papers in related fields that might help check the novelty of the hypotheses. The dataset is fully constructed by domain experts. We illustrate why the dataset shouldn't be collected by automatic methods in §A.4.

4 Methodology

In general, our method consists of a base multi-module framework and three feedback mechanisms (past-feedback, present-feedback, and future-feedback). We call the full framework as **M**ulti-m**O**dule framew**O**rk with pa**S**t present future fe**E**dback (MOOSE). The base framework without any feedback is called MOOSE-base. MOOSE is described in Figure 4 and Algorithm 1.

4.1 Base Framework

The base framework is developed based on the intuitive understanding of how social science researchers propose an initial research hypothesis.

Firstly, a researcher needs to find a proper research background, e.g., facial recognition payment system's social impact. A proper background

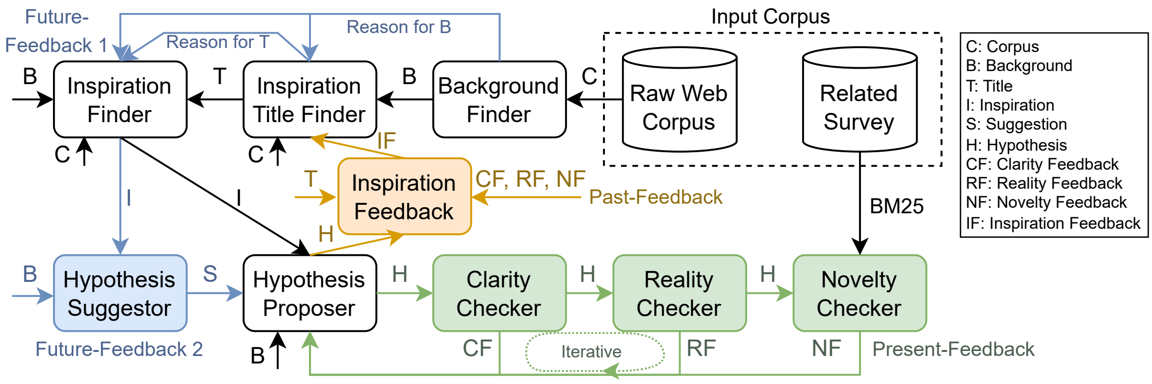

Figure 4: MOOSE: Our multi-module framework for TOMATO task. The black part is the base framework; orange part represents past-feedback.; green part represents present-feedback; blue part represents future-feedback. Each capitalized letter represents the generation of one of the modules. If a module has an input arrow pointing in with a capitalized letter, it represents that this module utilizes one of its previous modules' generation (which has the same letter pointing out) as input.

should be proposed with a proper understanding of the world. Accordingly, we develop a background finder module, which reads through raw web corpus to find a reasonable research background.

Secondly, since the proposed hypothesis should be novel, directly copying from raw web corpus usually is not enough. A good social science hypothesis should contain an independent variable and a dependent variable, and describe how the independent variable can influence the dependent variable. Therefore, building connections between two variables that have not been known for established connections contributes to a novel hypothesis. We hypothesize that proper inspiration can help this connection-building process, since a proper inspiration might serve as one of the variables itself, or might help to find such variables. However, it could consume lots of computing resources and even be practically impossible if the framework searches over the full web for every found background. Nevertheless, it could be much more viable if only searching over the titles of the full web corpus, and then only finding inspirational sentences in the passages which match the selected titles. Accordingly, we develop an inspiration title finder module and an inspiration finder module, together to find proper inspirations given a background.

Lastly, a hypothesis proposer module can utilize backgrounds and inspirations for hypotheses.

4.2 Present-Feedback

Now we have a hypothesis proposer module to propose hypotheses, but the base framework might overly rely on it. In other words, we cannot rely on one module to perform inference once to generate perfect enough research hypotheses (many might have flaws). Previous work on hypothetical induction (Yang et al., 2022b) tackles this problem by building an overly-generate-then-filter mechanism, which leverages LLMs to identify flaws in the generation and filters those with huge flaws. However, current LLMs are so powerful that they can not only identify whether there are any flaws but also provide feedback on possible modifications to avoid the flaws. Therefore we take a step further that the LLMs for filtering also provide feedback, so that the hypothesis proposer module can generate the hypothesis again, leveraging the feedback.

In terms of what aspects should the feedback focus on, Yang et al. (2022b) propose four aspects according to the philosophical definition and requirement for hypothetical induction (Norton, 2003). The aspects are (1) whether the hypothesis is consistent with observations; (2) whether the hypothesis reflects reality; (3) whether the hypothesis generalizes over the observations; (4) whether the hypothesis is clear, complete, and meaningful.

In our framework, we basically adopt the four aspects but reframe them to better fit the current task, and make them more concise. Specifically, aspect (2) contains aspect (1) most of the time (unless the observations are wrongly described). To save computing power, we adopt aspect (2) but not aspect (1). In addition, we reframe aspect (3) as whether the hypothesis is novel, and reframe aspect (4) as whether the hypothesis is clear and provides enough details. Accordingly, we develop a reality checker module, novelty checker module, and clarity checker module in Figure 4.

We call this feedback "present-feedback" since after generating the hypothesis, feedback can be instantly provided towards the hypothesis.

4.3 Past-Feedback

Just like the reward mechanism in reinforcement learning, some modules' generation can only be given feedback at a future time point. For instance, it is hard to evaluate selected inspirations unless we know what hypotheses these inspirations (combined with a given background) could lead to.

Sometimes low-quality generated hypotheses are caused by improper inspirations. Accordingly, we developed an inspiration feedback module, which utilizes generated hypotheses and previously selected titles to provide feedback to the inspiration title finder to find better titles. We call this feedback "past-feedback" since it is based on the future module's generation and is for a past module.

4.4 Future-Feedback

We also develop future-feedback, which is that the current module provides justifications for its generation to future modules (future-feedback 1, or abbreviated as FF1), or an additional module being placed previous to a key module to provide suggestions to reduce the reasoning burden of the key module (future-feedback 2, or abbreviated as FF2).

For future-feedback 1, the justifications are the reasons or analyses of the selected background or inspiration titles. No additional modules are needed to provide this information, instead, we modify the prompt to require a module to not only give an answer but also provide the reason or analysis of the answer. The intuition is that it could be helpful if the inspiration title finder module knows not only the background but also what possible research topics could be conducted for this background so as to select suitable titles; it could be also helpful for the inspiration finder module to know why this background was selected and what potentially helpful inspirations could be found from the passage with the corresponding selected titles.

For future-feedback 2, the intuition is that it can be still challenging for the hypothesis proposer module to propose high-quality hypotheses. Therefore we may have an additional module to undertake some reasoning burdens of the hypothesis proposer module. Accordingly, we develop a hypothesis suggestor module to provide some initial suggestions on how to utilize the inspirations and background first, and then the hypothesis proposer can build upon the suggestions to propose more novel or more complicated hypotheses.

5 Experiments

5.1 Evaluation Metrics & Details

We conduct both automatic evaluation and human evaluation for the experiments.

For automatic evaluation, we adopt validness, novelty, and helpfulness as three aspects for GPT-4 to evaluate. We choose validness and novelty because they are the two basic requirements for hypothetical induction illustrated in philosophical literature (Norton, 2003; Yang et al., 2022b). In addition, these two scores also highly resemble the current ACL review form, which requires reviewers to score submitted papers on soundness and excitement aspects. We choose helpfulness because the final goal of the TOMATO task is to provide help and assistance for human scientists.

In §A.6 we illustrate why we don't adopt evaluation metrics such as (1) relevance and significance, and (2) BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), or METEOR (Banerjee and Lavie, 2005).

For human (expert) evaluation, evaluation metrics are the same. Three experts (social science PhD students) take charge of the expert evaluation. They evaluate on 400 randomly selected hypotheses from the baseline and variants of the MOOSE framework. To avoid any bias, they are not told which methods we are comparing; the order of generated hypotheses to compare is also randomized. We introduce how the 400 hypotheses are selected in §A.7, and the high expert agreement in §A.8.

Each metric is on a 5-point scale. Both experts and GPT-4 are given the same description of the scale and evaluation standard of the three aspects (listed in §A.10).

Out of the metrics, we consider the novelty metric to be relatively more important than the validness metric. Because the goal of the TOMATO task is to assist human researchers, but not to directly add the machine-proposed hypotheses to the literature. If the hypotheses are fully valid but not novel, then they are not helpful at all; but if the hypotheses are novel but not valid, then they can still be possible to inspire human researchers to develop novel and valid hypotheses. Helpfulness is also an important metric since it could be seen as an overall evaluation of a hypothesis.

In §A.9, we introduce the surprisingly high consistency between expert evaluation and GPT4 eval-

	Validness	Novelty	Helpfulness
Baseline	3.954	2.483	3.489
MOOSE-base	3.907	3.081	3.859
w/ future-feedback	**3.955**	3.226	**3.953**
w/ future- and past-feedback	3.916	**3.390**†	3.931†

Table 2: Effect of MOOSE-base, future-feedback and past-feedback (evaluated by *GPT-4*). MOOSE-related results are averaged over iterations of present-feedback. Results with † mean the difference compared to the baseline is statistically significant ($p < 0.01$) using Bootstrap method (Berg-Kirkpatrick et al., 2012).

	Validness	Novelty	Helpfulness
MOOSE (w/o present-feedback)	3.823	3.114	3.809
w/ 1 iteration of present-feedback	3.918	3.199	3.900
w/ 2 iterations of present-feedback	3.951	3.293	3.956
w/ 3 iterations of present-feedback	3.969	3.270	**3.962**
w/ 4 iterations of present-feedback	**3.970**†	**3.329**†	3.951†

Table 3: Effect of present-feedback (evaluated by *GPT-4*). Results with † mean the difference compared to MOOSE w/o present-feedback is significant ($p < 0.01$).

uation, indicating that GPT-4 might be able to provide a relatively reliable evaluation for machine-generated (social science and business) hypotheses.

5.2 Baselines & Base Model Selection

Since the TOMATO task is to propose hypotheses given only corpus, a natural baseline is to use a corpus chunk as input, and directly output hypotheses.

We use `gpt-3.5-turbo` for each module in MOOSE. To be fair, the baseline is also instantiated with `gpt-3.5-turbo`. The training data of the model checkpoint is up to September 2021, while all papers in our dataset are published after January 2023, so the model has not seen any of the collected papers in the dataset.

5.3 Main Results

In this subsection, we compare MOOSE-base with the baseline and examine the effect of each of the three feedback mechanisms to MOOSE-base.

We first introduce the number of generated hypotheses being evaluated in §5.3 and §6. For experiments evaluated with GPT-4, fifty backgrounds are selected for each method. For MOOSE-related methods, for each background, on average around 6 inspirations are extracted, resulting in 4 different hypotheses. Each hypothesis leads to another 4 more refined ones with present-feedback. Therefore on average for each MOOSE-related method in GPT-4 evaluation tables, around 50*4*5=1000 hypotheses are evaluated. For experiments evaluated

	Validness	Novelty	Helpfulness
Baseline	3.579	2.276	2.632
MOOSE-base	3.500	2.855	3.026
w/ future-feedback	3.645	3.105	3.303
w/ future- and past-feedback	**3.750**	**3.197**	**3.368**

Table 4: Effect of MOOSE-base, future-feedback and past-feedback (evaluated by *experts*). MOOSE results are selected from the 5^{th} iteration of present-feedback.

	Validness	Novelty	Helpfulness
MOOSE-base (w/o present-feedback)	3.342	2.382	2.500
w/ 2 iterations of present-feedback	**3.539**	2.803	2.934
w/ 4 iterations of present-feedback	3.500	**2.855**	**3.026**
MOOSE (w/o present-feedback)	3.224	2.737	2.855
w/ 2 iterations of present-feedback	3.579	**3.250**	3.342
w/ 4 iterations of present-feedback	**3.750**	3.197	**3.368**

Table 5: Effect of present-feedback (eval. by *experts*).

with expert evaluation, in general, we randomly select one hypothesis for each background, resulting in 50 hypotheses evaluated for each line of the method in expert evaluation tables.

Table 2 shows GPT-4's evaluation targeting at comparing MOOSE-base and the baseline and shows the effect of future-feedback and past-feedback. In this table, MOOSE-related results are averaged over iterations of present-feedback to not be influenced by present-feedback. MOOSE-base largely outperforms the baseline in terms of both novelty and helpfulness, but slightly lower in terms of validness. As illustrated in §5.1, since the purpose of the TOMATO task is to inspire and help human researchers, novelty and helpfulness metrics should be more important. In practice, we find many hypotheses from baseline almost only rephrasing some sentences in the input corpus, adding little novelty content. MOOSE-base with future-feedback comprehensively outperforms MOOSE-base in terms of all three metrics. MOOSE-base with both future and past-feedback largely outperforms MOOSE-base with future-feedback in novelty and performs slightly lower in validness and helpfulness metrics. One of the reasons is that the past-feedback may focus more on the novelty aspect because the novelty checker module provides more negative present-feedback than the reality checker module.

Table 3 shows the effect of present-feedback with GPT-4 evaluation. In this table, the results are averaged over three experiments: MOOSE-base, MOOSE-base with future-feedback, and MOOSE-base with both future and past-feedback to focus on

	Validness	Novelty	Helpfulness
Rand background	3.954	2.483	3.489
Rand background and rand inspirations	3.773	2.957	3.643
Rand background and BM25 inspirations	3.585	**3.364**	3.670
Gpt-3.5 picked background and inspirations	**3.812**	2.818	**3.733**
Groundtruth background and inspirations	**3.876**	3.000	3.806
Groundtruth hypotheses	3.700	**3.380**	**3.880**

Table 6: Analysis of retrieval's effect on generated hypotheses (evaluated by *GPT-4*). No methods here utilize any feedback mechanisms. Every method here uses the same ChatGPT-based hypothesis proposer module.

	Validness	Novelty	Helpfulness
MOOSE	3.916	3.390	3.931
w/o future-feedback 2	3.895	3.281	3.918
w/o future-feedback 1	3.882	3.355	3.935
w/o access to related survey	3.889	**3.431**	3.886
w/ randomized corpus	**3.941**	3.227	**3.955**

Table 7: More ablation study (evaluated by *GPT-4*). Results are averaged over iterations of present-feedback.

present-feedback. It shows that as more iterations of present-feedback are conducted, validness and novelty steadily go up; helpfulness also steadily goes up but reaches the best performance with 3 iterations of present-feedback.

Table 4 shows expert evaluation results on the comparison between MOOSE-base and the baseline, and the effect of future-feedback and past-feedback. MOOSE-related results are selected from the 5^{th} iteration of present-feedback. Similar to GPT-4 evaluation, MOOSE-base largely outperforms the baseline in terms of Novelty and Helpfulness; MOOSE-base with future-feedback comprehensively outperforms MOOSE-base. Different from GPT-4 evaluation, MOOSE-base with future and past-feedback also comprehensively outperforms MOOSE-base with future-feedback. We think one of the reasons could be that GPT-4 might grade validness based on how frequently it has seen relevant texts, but not true understanding of the world. Therefore a more novel hypothesis might tend to have a relatively lower score in validness and helpfulness under GPT-4 evaluation.

Table 5 shows the expert evaluation of present-feedback. MOOSE-base and MOOSE are both evaluated. Overall performance generally goes up with more iterations of present-feedback, but there might be an optimal number of iterations.

6 Analysis

6.1 Background and Inspirations

Here we try to answer "Is ChatGPT necessary for background and inspiration selection?".

Table 6 shows various methods for background and inspiration selection. In general, there might be a validness-novelty trade-off that if a method reaches a high novelty score, then it is usually hard for it to reach a high validness score. It is surprising that a randomly selected background and randomly selected inspirations can lead to hypotheses with relatively comparable validness and novelty to ChatGPT-picked background and inspirations. Empirically we hypothesize the reason is that randomly picked inspirations are mostly not related to the background, resulting in a high novelty (but less validness and helpfulness). In addition, BM25 (Robertson et al., 2009) picked background and inspirations reach a much higher novelty score compared to ChatGPT-picked ones. Empirically we do not find BM25 retrieved inspirations to be similar to the background, but they are usually with more concrete contents compared with random inspirations. Not surprisingly, ChatGPT picked background and inspirations reach the highest helpfulness score among those without any ground-truth annotations. Lastly, ground-truth hypotheses reach the highest novelty and helpfulness.

6.2 More Ablation Studies

Table 7 shows ablation studies on future-feedback, access to surveys, and the selection of corpus.

Firstly, for future-feedback, we separately test the effect of future-feedback 1 (FF1) and future-feedback 2 (FF2). Without FF2, performance comprehensively drops; without FF1, performance drops on validness and novelty, with helpfulness remaining comparable. It seems that FF2 is more significant than FF1. However, the fact that FF1 works on inspiration title finder and inspiration finer modules does not mean that it works on all modules. Empirically we find that adding the reasons (or prospects) for background and inspirations to the hypothesis proposer module will cause a more valid but much less novel generation of hypotheses. The reason is that the hypothesis proposer module tends to simply follow the prospects, which do not have a global view of both background and all inspirations, but only focus on one background or one inspiration. Instead, FF2 (the hypothesis suggestor module) has the global view and only provides soft initial suggestions on how to combine the background and inspirations together. With the hypotheses suggestor module, the hypotheses proposer module is prompted to further combine the initial suggestions and other inspirations to propose hypotheses. To be fair, MOOSE-base, which is not equipped with the hypothesis suggestor module, has the same prompt to combine the inspirations together (just without suggestions) to propose hypotheses.

Secondly, we cut the access of novelty detector to related surveys to check the effect of related surveys. As a result, novelty largely goes up (0.04), and validness goes down to around 0.26. Empirically one of the main reasons is that BM25 hardly retrieves enough similar survey chunks, so that access to the survey leads novelty detector to tend to reply the hypotheses are novel since it is not mentioned in the related survey. Without present-feedback, MOOSE and MOOSE w/o access to survey perform quite comparably.

Lastly, the raw corpus in the dataset is from two sources: passages that contain the ground truth backgrounds and passages that contain the ground truth inspirations. In all of the previous experiments, backgrounds are extracted from the background passages, and inspirations are extracted from the inspirations passages. To see whether the passages are only restricted to their designed role, in MOOSE w/ randomized corpus experiment, we use inspiration corpus for background extraction and use both inspiration and background corpus for inspiration extraction. As a result, validness goes up by about 0.025, while novelty goes down by about 0.16. We think one of the reasons is that, in this setting, after selecting a background from an inspiration passage, MOOSE tends to retrieve the same inspiration passage to find inspirations, which leads to less novel results.

6.3 Qualitative Analysis

The following box shows one generated counter-intuitive hypothesis (expert evaluation appended).

> *In collectivist cultures, individuals engage in more conspicuous consumption behaviors compared to individualistic cultures.* (Validness: 3.3; Novelty: 4.0; Helpfulness: 4.0)

Here is the assessment from one of the experts:

The main reason I give a high mark for both three dimensions of this hypothesis is because:

(1) For validness, this hypothesis is based on existing cultural theories and empirical evidence that suggests cultural values significantly impact consumer behavior. It aligns with established concepts like collectivism and individualism that have been widely studied in cross-cultural psychology.

(2) For novelty, this hypothesis is counter-intuitive to some extent. Prior research has shown that collectivist cultures often prioritize group harmony, cooperation, and social cohesion over individual desires or displays of wealth. This emphasis on collective well-being might suggest a reduced inclination toward overt displays of personal wealth or status through conspicuous consumption. However, this hypothesis suggests the opposite and says individuals in collectivist cultures could engage in more conspicuous consumption, which is more commonly linked to individualistic societies in popular perceptions. This challenges the notion that members of collectivist cultures avoid conspicuous consumption behaviors.

(3) For helpfulness, if this hypothesis is confirmed, it could have significant practical implications. Understanding the impact of cultural values on conspicuous consumption can assist businesses and marketers in crafting more effective cross-cultural marketing strategies. It could also aid policymakers in addressing societal issues related to consumerism.

In addition to the analysis of this counter-intuitive example, we also provide qualitative analysis on the difference between hypotheses generated from the baseline, MOOSE-base, MOOSE-base w/ future-feedback, and MOOSE-base w/ future and past-feedback in §A.12. More about qualitative analysis are in §A.13 (high expert evaluation) and §A.14 (factors for good hypotheses).

7 Conclusion

In this paper, we propose a novel task, automated open-domain hypothetical induction (TOMATO), which is the first task in NLP to focus on social science and business research hypotheses discovery. Along with the task, we construct a dataset consisting of 50 recent social science and business papers published in top academic journals. We also developed a multi-module framework MOOSE for the TOMATO task, which contains a base framework and three novel feedback mechanisms. Our experiments indicate that MOOSE-base outperforms an LLM-based baseline, and the three feedback mechanisms can progressively further improve over MOOSE-base. Surprisingly, evaluated by PhD students, MOOSE produces many valid, novel (means "not existing in the literature"), and helpful (for scientists' research development) hypotheses.

Limitations

From the first look, it might seem that the proposed dataset consists of only 50 recent papers. However, they are all manually collected by experts (PhD students), and are annotated with lots of details (e.g., identifying background and inspirations, finding relevant raw web passages for background and inspirations, reasoning process, complexity level). In addition, each paper has been published in a top social science journal, representing the pinnacle of human intelligence. This means it would be incredibly exciting if LLMs could propose a hypothesis from even a single one of these recent papers.

It might also seem that it is not clear whether the design of the framework can apply to other disciplines. However, to the best of our knowledge, this is the first paper using LLMs that can propose novel scientific hypotheses that are new to humanity. We choose social science as the breakthrough point since the main data format of social science is language. Table 1 shows that the dataset covers 7 different disciplines (e.g., Psychology, Management, Marketing). It would be nearly impossible for the first few works to develop a general method to propose novel hypotheses for all disciplines.

Ethics Statement

This article follows the ACL Code of Ethics. To our knowledge, there are no foreseeable potential risks to use the datasets and methods in this paper.

References

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Taylor Berg-Kirkpatrick, David Burkett, and Dan Klein. 2012. An empirical investigation of statistical significance in nlp. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 995–1005.

Daniil A. Boiko, Robert MacKnight, and Gabe Gomes. 2023. Emergent autonomous scientific research capabilities of large language models. *CoRR*, abs/2304.05332.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. COMET: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4762–4779, Florence, Italy. Association for Computational Linguistics.

Andres M Bran, Sam Cox, Andrew D White, and Philippe Schwaller. 2023. Chemcrow: Augmenting large-language models with chemistry tools. *arXiv preprint arXiv:2304.05376*.

Rajarshi Das, Manzil Zaheer, Dung Thai, Ameya Godbole, Ethan Perez, Jay Yoon Lee, Lizhen Tan, Lazaros Polymenakos, and Andrew McCallum. 2021. Case-based reasoning for natural language queries over knowledge bases. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9594–9611, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jia Gao, Ying Rong, Xin Tian, and Yuliang Yao. 2023. Improving convenience or saving face? an empirical analysis of the use of facial recognition payment technology in retail. *Information Systems Research*.

Vinod Goel, Gorka Navarrete, Ira A Noveck, and Jérôme Prado. 2017. The reasoning brain: The interplay between cognitive neuroscience and theories of reasoning.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132, Austin, Texas. Association for Computational Linguistics.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Sean Welleck, Bodhisattwa Prasad Majumder, Shashank Gupta, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. *CoRR*, abs/2303.17651.

John D Norton. 2003. A little survey of induction.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *NeurIPS*.

Feng Pan, Rutu Mulkar-Mehta, and Jerry R. Hobbs. 2011. Annotating and learning event durations in text. *Comput. Linguistics*, 37(4):727–752.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Liden, Zhou Yu, Weizhu Chen, and Jianfeng Gao. 2023. Check your facts and try again: Improving large language models with external knowledge and automated feedback. *CoRR*, abs/2302.12813.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A. Smith, and Mike Lewis. 2022. Measuring and narrowing the compositionality gap in language models. *CoRR*, abs/2210.03350.

Biqing Qi, Kaiyan Zhang, Haoxiang Li, Kai Tian, Sihang Zeng, Zhang-Ren Chen, and Bowen Zhou. 2023. Large language models are zero shot hypothesis proposers. *CoRR*, abs/2311.05965.

Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Noah Shinn, Beck Labash, and Ashwin Gopinath. 2023. Reflexion: an autonomous agent with dynamic memory and self-reflection. *CoRR*, abs/2303.11366.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

Hanchen Wang, Tianfan Fu, Yuanqi Du, Wenhao Gao, Kexin Huang, Ziming Liu, Payal Chandak, Shengchao Liu, Peter Van Katwyk, Andreea Deac, et al. 2023a. Scientific discovery in the age of artificial intelligence. *Nature*, 620(7972):47–60.

Qingyun Wang, Doug Downey, Heng Ji, and Tom Hope. 2023b. Learning to generate novel scientific directions with contextualized literature-based discovery. *CoRR*, abs/2305.14259.

Kevin Yang, Yuandong Tian, Nanyun Peng, and Dan Klein. 2022a. Re3: Generating longer stories with recursive reprompting and revision. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 4393–4479. Association for Computational Linguistics.

Zonglin Yang, Li Dong, Xinya Du, Hao Cheng, Erik Cambria, Xiaodong Liu, Jianfeng Gao, and Furu Wei. 2022b. Language models as inductive reasoners. *CoRR*, abs/2212.10923.

Zonglin Yang, Xinya Du, Erik Cambria, and Claire Cardie. 2023a. End-to-end case-based reasoning for commonsense knowledge base completion. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3509–3522, Dubrovnik, Croatia. Association for Computational Linguistics.

Zonglin Yang, Xinya Du, Rui Mao, Jinjie Ni, and Erik Cambria. 2023b. Logical reasoning over natural language as knowledge representation: A survey. *CoRR*, abs/2303.12023.

Zonglin Yang, Xinya Du, Alexander Rush, and Claire Cardie. 2020. Improving event duration prediction via time-aware pre-training. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3370–3378, Online. Association for Computational Linguistics.

Ruiqi Zhong, Peter Zhang, Steve Li, Jinwoo Ahn, Dan Klein, and Jacob Steinhardt. 2023. Goal driven discovery of distributional differences via language descriptions. *CoRR*, abs/2302.14233.

A Appendix

A.1 Hyper-parameters, Anonymous Github Link, and Full Prompts

All experiments are conducted with `gpt-3.5-turbo`, with 0.9 temperature and 0.9 top_p.

The hyperparameters for GPT-4 evaluation are 0.0 temperature to ensure the evaluation scores are stable, and 0.9 top_p.

The dataset and code of this submission are already public on GitHub. An anonymous version can be found at: `https://anonymous.4open.science/r/TOMATO/`. Particularly, the dataset can be found at `https://anonymous.4open.science/r/TOMATO/Data/business_research.xlsx`.

The full prompts for MOOSE framework is shown in prompts_for_tomato_modules() function in utils.py.

A.2 More Related Works on Reasoning and Scientific Discovery

This paper is a successive work in inductive reasoning and is different from commonsense reasoning (Bosselut et al., 2019; Yang et al., 2020) in that the novel social science hypotheses do not belong to commonsense.

	Reasoning Complexity	Association Complexity
Easy	24	12
Medium	17	25
Hard	9	13

Table 8: Statistics of the complexity of the dataset.

Case-based reasoning (Das et al., 2021; Yang et al., 2023a) also falls in the domain of inductive reasoning, but case-based reasoning is more about high-level guidance on methodology design (case retrieve, reuse, revise, and retain), which is not involved in this paper.

Qi et al. (2023) work on zero-shot hypothesis proposing, which is a concurrent work to our paper. They don't focus on social science and business disciplines, and mostly adopt a single LLM as method (prompting, finetuning).

A.3 Dataset Complexity Distribution

Table 8 illustrates the complexity distribution of the proposed dataset from both reasoning and association perspectives. "Easy" in the table means it is relatively easy compared to other publications in the dataset, but does not mean it is actually easy to induce the hypotheses.

A.4 Why the Tomato Dataset Shouldn't Be Collected by Automatic Methods

Firstly, there are many hypotheses in a social science publication, which might need an expert to identify which hypothesis is suitable for this task (e.g., whether it is a main hypothesis, whether the background and inspirations are properly introduced).

Secondly, the background and inspirations scatter in a publication. It needs a deep domain understanding of the hypothesis, related background, and inspirations to select the background and inspirations out to form a complete reasoning chain to conclude the hypothesis.

Thirdly, it needs enough domain knowledge to find semantically similar texts (similar to the groundtruth selected background and inspirations) from the web, where the texts should contain enough details to help elicit the hypothesis.

A.5 Full Algorithm for the Proposed Multi-Module Framework

Algorithm 1 shows the full algorithm of the proposed framework.

A.6 Why Not Using Other Evaluation Metrics

Other relevant aspects from related literature include relevance (Wang et al., 2023b) and significance (Zhong et al., 2023).

We do not adopt relevance because our task setting is the automated and open domain, without a manually given background; neither for significance because social science is different from engineering subjects — (1) every hypothesis is to reflect the reality of the world, and as long as it reflects the world, it is significant. Therefore it is hard to tell which one is more significant even by experts; (2) the evaluation standard of significance varies from time to time. For example, in the 60s, conducting research on how to improve the assembly line's efficiency as much as possible was seen as very significant. However, in recent decades, how to alleviate the psychological depression of assembly line workers is seen as more significant.

We do not adopt BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), or METEOR (Banerjee and Lavie, 2005) as evaluation metric to compare the proposed hypothesis and the ground truth hypothesis since (1) proposing novel research hypotheses is an open problem, and (2) TOMATO has an automated open domain setting, which means the automatically selected background and inspirations are hardly the same as a few given ground truth ones (if background and inspirations are not the same, then it is meaningless to compare the hypothesis). Liu et al. (2016) have conducted a comprehensive analysis that they also reached a similar conclusion that BLEU, METEOR, or ROUGE is not suitable for an open-ended task (such as a dialogue system).

A.7 Hypotheses Selection for Expert Evaluation

In total, we randomly selected 400 hypotheses to be evaluated by experts. Specifically, for each background passage in the dataset (out of 50), we use 4 methods (which are to be compared) to collect in total 8 hypotheses.

The 8 hypotheses are from (1) the baseline; (2) the MOOSE-base framework; (3) MOOSE-base + future-feedback; (4) MOOSE-base + future-feedback + past-feedback. For (2) and (4), we collect three hypotheses, which are (a) without present-feedback; (b) after 2 iterations of present-feedback; and (c) after 4 iterations of present-feedback. For (1) and (3), we only collect one hypothesis, which is without present-feedback.

	Validness	Novelty	Helpfulness
Hard Consistency	0.298	0.337	0.361
Soft Consistency	0.755	0.793	0.791

Table 9: Hard and soft consistency scores between evaluation from different experts in terms of Validness, Novelty, and Helpfulness metrics.

With these collections, we can evaluate the effect of both the MOOSE-base framework and the three feedback methods, leading to results in Table 4 and Table 5.

Out of the three experts, one expert evaluates the full 400 hypotheses, and the other two each evaluate 104 hypotheses (the first and second 104 hypotheses out of 400). The reason we choose the number "104" is that (1) social science PhD students are quite busy and two of them can only have time to evaluate around 100 hypotheses; (2) the number should be dividable by 8 (since every 8 hypotheses form a group for comparison).

The results of the expert evaluation are averaged over the three experts. Specifically, expert evaluation essentially compares the 8 hypotheses within a group. The 400, 104, and 104 hypotheses evaluation scores can be written as arrays of [50, 8], [13, 8], and [13, 8]. We concatenate them to [76, 8], and average them across the first dimension.

The payment for expert evaluation is $1 per hypothesis.

A.8 Expert Qualification and Expert Agreement

The constructed dataset covers many subjects, but every collected publication is somewhat related to Marketing, which is a big topic in Business research. It is common in social science to conduct research that connects with other social science domains. The experts for expert evaluation are three PhD students majoring in Marketing. Therefore the experts are qualified enough to provide assessment for machine-generated hypotheses in the domain.

The consistency scores between experts are shown in Table 9. The soft consistency and hard consistency are defined in §A.9. All soft consistency scores are above 0.75 means, and the average difference between experts in terms of each metric is less than 1 (out of a 5-point scale), exhibiting high expert evaluation agreement.

	Validness	Novelty	Helpfulness
Hard Consistency	0.485	0.392	0.321
Soft Consistency	0.850	0.823	0.773

Table 10: Hard and soft consistency scores between expert evaluation and GPT-4 evaluation in terms of Validness, Novelty, and Helpfulness metrics.

A.9 Consistency Between Expert Evaluation and GPT-4 Evaluation

To check the consistency between expert evaluation and GPT-4 evaluation, we use the expert evaluation results and find the corresponding GPT-4 evaluation results. In total, there are 400 hypotheses evaluated by experts, so the sample we use to calculate the consistency score is 400.

Specifically, similar to Pan et al. (2011), for soft consistency, if the absolute difference between expert evaluation and GPT-4 evaluation (both are on a 5-point scale) is 0/1/2/3/4, then we assign a consistency score of 1.00/0.75/0.50/0.25/0.00; for hard consistency, if only the difference is 0, can the consistency score be 1.00, otherwise consistency score is 0.00. The hard and soft consistency scores shown in Table 10 are averaged for each metric.

The consistency scores are surprisingly high. All soft consistency scores are above 0.75 means, and the average difference between expert and GPT-4 evaluation in terms of each metric is less than 1 (out of a 5-point scale). The results indicate that GPT-4 might be able to provide a relatively reliable evaluation for machine-generated hypotheses.

A.10 Evaluation Aspects Description

Aspect 1: Validness.

5 points: the hypothesis completely reflects the reality;
4 points: the hypothesis almost completely reflects the reality, but has only one or two minor conflictions that can be easily modified;
3 points: the hypothesis has at least one moderate conflict or several minor conflicts;
2 points: the hypothesis has at least one major confliction with the reality or only establishes in very rare circumstances that are not mentioned in this hypothesis;
1 point: the hypothesis completely violates the reality.

Aspect 2: Novelty.

5 points: the hypothesis is completely novel and has not been proposed by any existing literature;
4 points: the main argument or several sub-

arguments of the hypothesis are novel;
3 points: the main argument is not novel, only one or two sub-arguments appear to be novel;
2 points: the full hypothesis is not novel, but the way it combines the topics can be inspiring for human researchers;
1 point: the hypothesis is not novel at all and not inspiring for human researchers.

Aspect 3: Helpfulness.

5 points: the hypothesis is novel, valid, clear, and specific enough that it is itself a mature research hypothesis, and human researchers can directly adopt it for publication with no modifications needed;
4 points: the hypothesis is novel enough and can be directly adopted by human researchers for publication after minor modifications;
3 points: the hypothesis should be largely modified or reconstructed by human researchers to adopt it;
2 points: modifying this hypothesis might not deserve the efforts, but a small part of this hypothesis is inspiring for human researchers to develop a new hypothesis;
1 point: the hypothesis is not helpful and not inspiring at all.

A.11 More Details About Past-Feedback Design

In practice, we find that ChatGPT is not capable enough to generate past-feedback with enough good quality for the Inspiration Feedback module. Instead, it tends to provide feedback as "the previous inspiration titles are not very relevant to the hypotheses or the background". As a result, the ChatGPT Inspiration Title Finder module tends to select inspiration titles that are very related to the background, resulting in a less novel hypotheses generation.

Therefore instead of instantiating with ChatGPT for the Inspiration Feedback module, we experiment with leveraging human heuristics. The heuristics are "if the inspiration titles are less related to the background, then more novel hypotheses are likely to be proposed.". With this heuristics-based past-feedback, MOOSE does perform better (as shown in the tables in §5 and §6).

This heuristics-based feedback is possible to be obtained by a language model since it has access to the novelty feedback of each hypothesis as well as the inspiration titles the hypothesis leveraged. Here our contribution is to propose a useful framework for the TOMATO task, which is not limited by any LLMs for any module in the framework. In the future, it is possible for more powerful LLMs to find better inspiration feedback than human heuristics.

A.12 Qualitative Analysis on Hypotheses Generated From Different Methods

We analyze four hypotheses from the baseline, MOOSE-base, MOOSE-base w/ future-feedback, and MOOSE-base w/ future and past-feedback (MOOSE), where the four methods use the same passage to extract background.

- Hypothesis from the baseline: *Companies that prioritize customer understanding will have higher profitability than companies that do not prioritize customer understanding.* (Evaluated by the expert, Validness: 4; Novelty: 1.5; Helpfulness: 2)

- Hypothesis from MOOSE-base: *The level of empathy displayed by leaders in a startup environment influences employees' job satisfaction and organizational success through the mediation of employees' perceived likelihood of negative outcomes and expectation of enjoyment, as well as their propensity toward risky choices.* (Evaluated by the expert, Validness: 3.5; Novelty: 3; Helpfulness: 3)

- Hypothesis from MOOSE-base w/ future-feedback: *Female CMOs in startups, leveraging their higher levels of empathy, are more likely to prioritize customer satisfaction by actively listening to customer feedback, incorporating customer insights into decision-making processes, and providing personalized customer experiences. This employee prioritization of customer satisfaction is positively associated with higher levels of customer engagement, increased customer loyalty, and improved brand recall, as measured by objective metrics such as sales figures, customer retention rates, and brand recognition in the market. The influence of female CMOs on employee behavior is mediated by their ability to foster a caring relationship with customers, as supported by empirical data and statistical analysis.* (Evaluated by the expert, Validness: 3.5; Novelty: 3.5; Helpfulness: 3.5)

- Hypothesis from MOOSE-base w/ future and past-feedback (MOOSE): *Female CMOs' empathy advantage influences their considera-*

tion of negative consequences of CSR initiatives, specifically in terms of employee well-being and job security. This relationship is moderated by individual differences in emotional intelligence. Additionally, the organizational culture and industry context will further influence the relationship between empathy advantage and consideration of negative consequences. The hypothesis will investigate whether female CMOs with higher levels of empathy are more likely to prioritize employee well-being and job security in the implementation of CSR initiatives, and whether this relationship is stronger in industries with a stronger emphasis on employee well-being and job security. It will also explore the mediating role of organizational culture and the moderating role of emotional intelligence in shaping the relationship between empathy advantage and consideration of negative consequences. (Evaluated by the expert, Validness: 4.5; Novelty: 4; Helpfulness: 4)

Analysis from the expert:

- *H1 falls short of challenging established assumptions or introducing a novel perspective beyond the widely accepted link between customer understanding and profitability.*

- *Both H2 & H3 center around a specific scenario involving female CMOs in startups and delve into their influence on customer satisfaction, employee behavior, and overall business results. From a research standpoint, this more focused approach points to a potential gap in the existing body of knowledge. Moreover, these two hypotheses surpass conventional understanding by considering how the empathy of female CMOs impacts employee behavior and business outcomes. They put forth a fresh viewpoint, suggesting that cultivating a compassionate rapport with customers, fostered by female CMOs, could positively affect customer engagement, loyalty, and brand recognition. These two hypotheses zoom in on a more specific context, introduce an innovative perspective, and probe a potential void in current research. They are anchored in the dynamic world of innovative business settings and propose a more nuanced and all-encompassing connection between variables.*

- *H4 retains its relevance within a modern business landscape by scrutinizing the intersection of empathy, CSR initiatives, and the dynamics of organizations. This syncs seamlessly with the criterion of being rooted in an innovative business environment. Moreover, it shakes up established assumptions by considering the potential adverse outcomes of CSR initiatives and the role empathy plays in shaping decision-making within this context. This hypothesis delves into a more intricate and thorough exploration, examining a broader spectrum of factors and interactions within a specific context. Additionally, it imparts a deeper comprehension of the interplay between empathy, business choices, and organizational results. It grapples with a more complex and distinctive scenario, unearths possible gaps in the existing literature, and introduces a new angle on the role of empathy in the realm of business decisions.*

A.13 Qualitative Analysis on Two MOOSE-Generated Hypotheses With High Expert Evaluation Scores

In the following two grey boxes are two generated hypotheses from MOOSE with high expert evaluation scores (appended to each hypothesis). The expert's assessment of the two hypotheses is:

> Hypothesis 1: *The level of personalization in crowdfunding campaign storytelling, the influence of social media influencers who align with the campaign, the presence of trust indicators, and the emotional appeal of the campaign will positively impact potential donors' likelihood of making a donation. Additionally, the timing of donation requests and the type of social media influencers (e.g., celebrities vs. micro-influencers) will moderate this relationship. The perceived risk associated with the crowdfunding campaign will negatively moderate the relationship between the emotional appeal and donation likelihood.* (Validness: 4.5; Novelty: 4.5; Helpfulness: 4.5)

These two hypotheses both present a comprehensive view of the research narrative. It encompasses multiple hypotheses, including the primary one, as well as the mediation effect, which serves to elucidate the causal connection between the independent and dependent variables. Concurrently, both

Hypothesis 2: *Limited financial resources and limited access to networks and markets of women entrepreneurs in the manufacturing sector in developing countries may negatively impact their investment in corporate social responsibility (CSR) initiatives that promote gender equality in host countries. This relationship is further influenced by the intersectionality of gender and race, with women of color facing additional challenges. Additionally, the hypothesis considers the role of institutional factors, such as legal frameworks and policies, and the influence of patriarchal structures on women entrepreneurs' ability to invest in CSR initiatives.* (Validness: 3.5; Novelty: 4; Helpfulness: 4)

hypotheses outline the range of the effect — namely, the circumstances in which this effect is applicable, under which scenarios where it might be weakened, and under which situation it could potentially be inverted.

In terms of novelty: 1. Limited prior research or a gap in the existing literature. This means that there is a dearth of studies or information available on the subject, making it an unexplored area. 2. Based on a new business setting. It is grounded in an innovative business environment, characterized by novel technologies, contemporary themes, and evolving business requirements. 3. The topic offers a fresh and unique perspective that goes beyond conventional understanding. It might challenge existing assumptions, propose new theories, or present an unconventional approach.

A.14 Essential Factors for Good Social Science (and Business) Hypotheses

According to social science PhD students, counter-intuitive and novel hypotheses are the mostly favoured (by top social science and business journals). Intuitive and novel hypotheses are also good but not as good as the counter-intuitive ones. Here "novel" refers to "not pointed out by existing literatures".

Empirically they think of all the hypotheses on top social science journals, around 20% are counter-intuitive, leaving the remaining 80% intuitive.

Counter-intuitive hypotheses tend to receive a lower validness evaluation compared to intuitive ones. For this reason, we highlight the counter-intuitive hypothesis in §6.3, even if it receives a lower score in validness than hypotheses in §A.13.

A.15 Future Directions

This work discovered the possibility of LLMs to propose novel research hypotheses. But it mainly focuses on the social science and business disciplines. It would be very interesting to investigate how LLMs can induce novel hypotheses for other disciplines (especially engineering domains).

In addition, the MOOSE framework could be further improved to induce more valid and novel hypotheses for social science and business domains.

From the aspect of human-AI interaction, it would be also interesting to see how MOOSE can act as an AI Copilot to assist scientists in hypothesis discovery.

A.16 License of the New Dataset (TOMATO)

The license is CC-BY 4.0. It should be used for research purposes.

A.17 Method for Prevention of Personal Information

During the dataset collection process, we make sure that the dataset is constructed only with public information (published papers, Wikipedia, business review, and news).

A.18 Dataset Split of TOMATO

The full dataset is used only as test set.

Algorithm 1 Algorithm for MOOSE

Input: Raw web corpus C, related survey S (, previous selected titles $prev\_t$ which is selected without past-feedback, previous generated hypotheses $prev\_h$ which is generated without past-feedback, previous present-feedback for previous generated hypotheses $prev\_prf$)

Parameter: Number of iterations for present-feedback N

Output: List of hypotheses H

for c in C **do**
  b, b_reason = Background_Finder(c)
  if $b ==$ None **then**
    continue
  end if
  f = Inspiration_Feedback($prev\_t$, $prev\_h$, $prev\_prf$)
  t, t_reason = Inspiration_Title_Finder(C, b, b_reason, f)
  p = Find_Passage_with_Title(t, C)
  i = Inspiration_Finder(b, b_reason, p, t_reason)
  s = Hypothesis_Suggestor(b, i)
  h = Hypothesis_Proposer(b, i, s)
  for iteration t $\in$ 0...N **do**
    $cfdbk$, $rfdbk$, $nfdbk$ = Clarity_Checker(h), Reality_Checker(h), Novelty_Checker(h, S)
    cur_prf = [$cfdbk$, $rfdbk$, $nfdbk$]
    h = Hypothesis_Proposer(b, i, s, h, cur_prf)
  end for
  H.append(h)
end for
return H