Concept Paper

A Framework for Automated Hypothesis Testing

Hardik Tiwari *

Posted Date: 12 September 2025

doi: 10.20944/preprints202509.1105.v1

Keywords: automated hypothesis testing; natural language processing; machine learning; statistics; large language models

 doi:10.20944/preprints202509.1105.v1

Concept Paper

A Framework for Automated Hypothesis Testing

Hardik Tiwari

India; i@hardiktiwari.com

Abstract

Hypothesis testing is a foundational process in scientific discovery and data driven decision making, yet it traditionally demands expert intervention and manual interpretation of both hypotheses and data. Tools like PSPP or IBM SPSS offer interfaces for analysis, but their integration requires analysts to translate natural language questions into formal statistical tests. On the other hand, recent advances in NLP and ML offer tools to automate elements of scientific analysis, but their integration for full-cycle hypothesis testing remains unsolved. This indicates that a significant gap exists in creating an integrated system that can automate this translation from human intent to statistical execution i.e. ability to interpret natural language hypotheses, align them with appropriate datasets, and execute relevant statistical or ML models without human input. Here I propose the development of a cognitive framework that synthesizes LLMs with a statistical decision engine to fully automate the hypothesis testing workflow. The system parses hypotheses into structured analytical intents using NLP Techniques, then maps them to structured data and accurately selects, executes the appropriate statistical test. The framework concludes by translating the technical results into a clear, human readable summary, replicating the outcome of a manual analysis. Using transformer-based models for semantic parsing and rule based statistical selection, we will demonstrate that our system can accurately validate causal and correlational hypotheses across benchmark datasets. This system's performance will be validated against benchmark datasets to ensure validity with expert-led analysis. This framework significantly reduces the cognitive load required for early-stage hypothesis evaluation, making exploratory research more scalable. The immediate implication is a significant acceleration of the research and discovery cycle across numerous fields.

Keywords: automated hypothesis testing; natural language processing; machine learning; statistics; large language models

1. Introduction

Hypothesis testing has long served as the foundation of scientific inquiry and data-informed decision-making. It offers a systematic means of formulating questions, analysing data, and drawing conclusions. Yet, the traditional hypothesis testing workflow has seen minimal innovation, remaining dependent on manual effort and specialized statistical expertise. Tools like IBM SPSS and PSPP provide robust statistical capabilities, but they act merely as computational aids requiring users to manually translate real-world questions into precise statistical queries (IBM Corp., 2021; Free Software Foundation, 2023). This human-dependent translation stage creates a bottleneck in speed, accessibility, and scalability.

Meanwhile, transformative advancements in artificial intelligence particularly in Natural Language Processing (NLP) and Machine Learning (ML) have enabled machines to understand and generate human language with unprecedented accuracy. The emergence of models such as BERT and GPT has opened the door for machine comprehension of scientific language (Devlin et al., 2019; Brown et al., 2020). These Large Language Models (LLMs) present an opportunity to automate the end-to-end hypothesis testing pipeline. This proposal introduces a cognitive framework that leverages such models to interpret natural language hypotheses, align them with structured data, execute statistical tests, and generate human-readable interpretations.

 doi:10.20944/preprints202509.1105.v1

2. Problem Statement and Research Gap

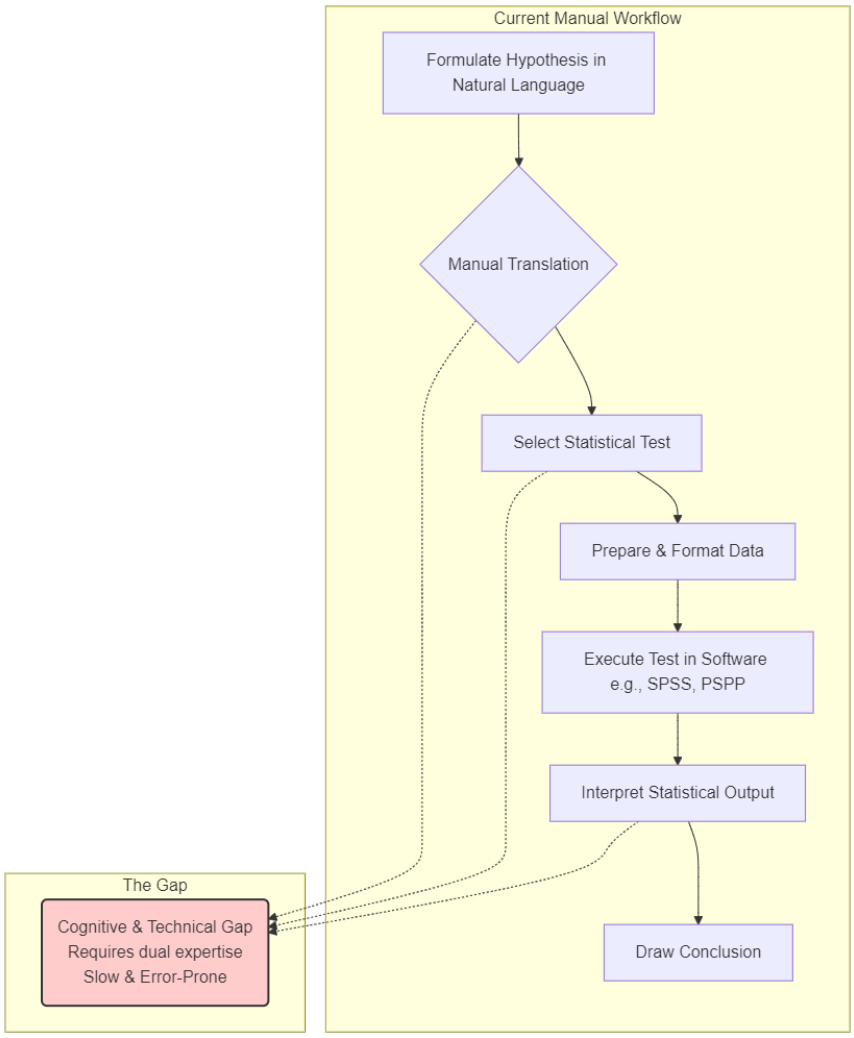

Figure 1. Problem Statement and Research Gap.

The central problem is the persistent and inefficient gap between the expression of a scientific hypothesis in natural language and its formal execution through statistical software. While powerful tools for both language understanding and statistical analysis exist, their integration into a seamless, end-to-end workflow for scientific inquiry remains an unsolved challenge (Shen et al., 2025; Huang et al., 2025). This disconnect manifests in several ways:

2.1. **Cognitive Load and Expertise Bottleneck:** Researchers must possess dual expertise in their own domain and in statistical methodology to correctly select and apply tests.

2.2. **Scalability Issues:** The manual process is slow and laborious, making large-scale exploratory research, where hundreds of hypotheses might be evaluated, impractical.

2.3. **Reproducibility Challenges:** The manual selection of tests can introduce subjective biases and errors, hindering the reproducibility of findings.

While some platforms have attempted to automate parts of the scientific process, they often fall short of providing a holistic solution. They may focus on a narrow set of statistical models or fail to cover the full cycle from raw linguistic input to fully interpreted output. A significant research gap therefore exists for a unified system that can intelligently and autonomously manage the entire hypothesis testing pipeline.

3. Proposed Framework

 doi:10.20944/preprints202509.1105.v1

We propose the development of a multi-stage cognitive framework that automates hypothesis testing by integrating advanced NLP with a robust statistical decision engine. This system is designed to accept a hypothesis in plain English, process it through a series of analytical stages, and return a clear, concise, and statistically sound conclusion. The framework is composed of three core stages:

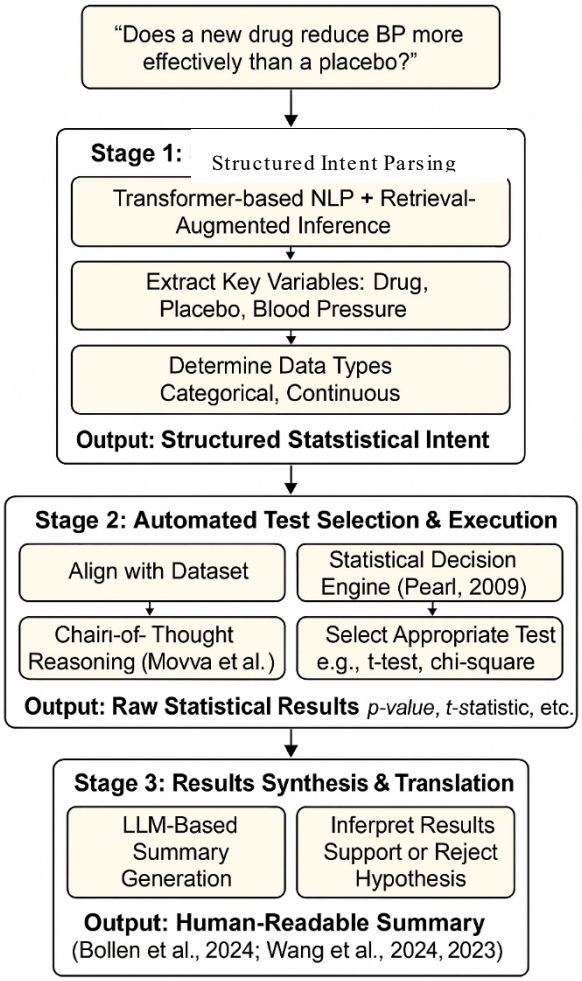

Figure 2. Proposed framework - Automated Hypothesis Testing Pipeline.

3.1. **Stage 1: Structured Intent Parsing:** The system ingests a user's hypothesis (e.g., "Does a new drug reduce blood pressure more effectively than a placebo?"). Using retrieval-augmented and transformer-based NLP techniques, it parses this sentence into a structured statistical intent. This involves identifying key components: the variables of interest (drug, placebo, blood pressure), the relationship being tested (causal, correlational), and the nature of the data (An et al., 2024; Devlin et al., 2019).

3.2. **Stage 2: Automated Test Selection and Execution:** The structured intent is then used to programmatically align with a provided dataset. A rule-based statistical decision engine, informed by principles of causal inference and statistical theory (Pearl, 2009), automatically selects the most appropriate statistical test. For example, it would identify the blood pressure comparison as requiring an independent samples t-test. This stage leverages modern chain-of-thought reasoning principles to ensure the selection is logical and defensible (Movva et al., 2025). The system then executes the test using standard computational libraries.

3.3. **Stage 3: Results Synthesis and Translation:** Finally, the framework translates the raw statistical output (e.g., p-value, t-statistic, effect size) into a human-readable narrative. It generates a summary that explains whether the hypothesis was supported, interprets the meaning of the

 doi:10.20944/preprints202509.1105.v1

results in the context of the original question, and notes any relevant statistical details, effectively replicating the summary an expert analyst would provide (Bollen et al., 2024; Wang et al., 2023).

4. Methodology and Architecture Overview

The proposed framework will be architected as a modular system, with each component responsible for a specific task in the pipeline:

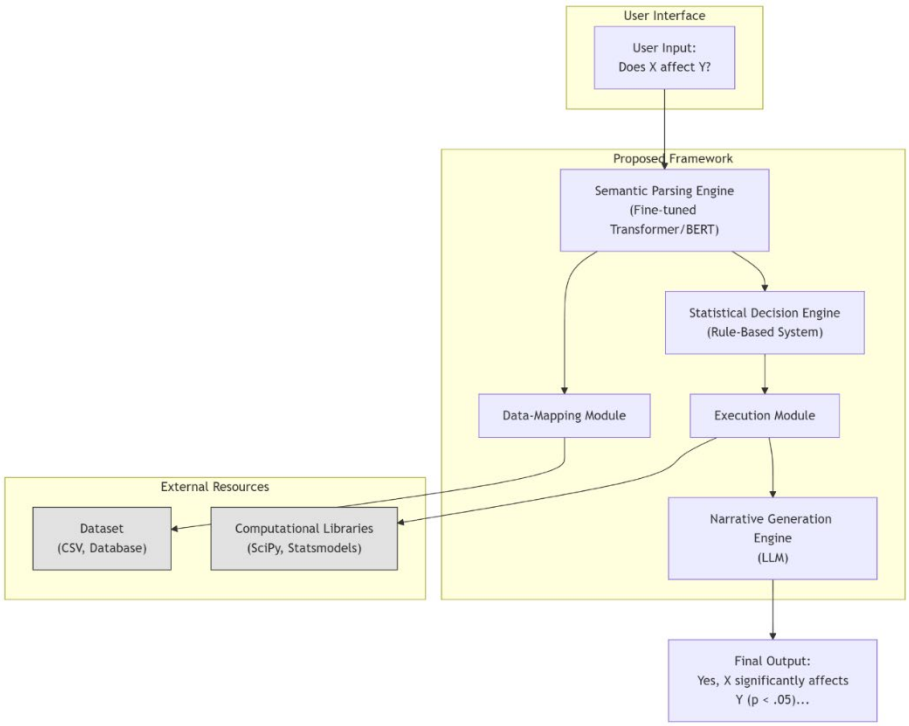

Figure 3. Architecture overview.

4.1. **Natural Language Interface:** A user-facing input layer where hypotheses are submitted.

4.2. **Semantic Parsing Engine:** At its core, this engine will utilize a fine-tuned transformer model (e.g., a derivative of BERT) to perform Named Entity Recognition (NER) and Relation Extraction. It will be trained to identify statistical entities (variables, populations) and the relationships between them.

4.3. **Data-Mapping Module:** This component will take the parsed entities and map them to the corresponding columns or variables within a structured dataset (e.g., a CSV file or database table).

4.4. **Statistical Decision Engine:** This will be a knowledge-based system containing a ruleset derived from statistical theory. The rules will map the characteristics of the parsed intent (e.g., two-group comparison, continuous outcome) to a specific statistical procedure (e.g., scipy.stats.ttest_ind). The engine's logic will be explicitly designed to handle different types of hypotheses, including those involving causality (Pearl, 2009).

4.5. **Execution Module:** This module interfaces with established Python libraries like statsmodels and SciPy to perform the actual calculations.

4.6. **Narrative Generation Engine:** An LLM (leveraging the few-shot learning paradigm described by Brown et al., 2020) will be prompted with a template that includes the statistical results. This will enable it to generate a fluent, context-aware summary of the findings.

5. Validation Plan and Future Work

The performance and validity of the framework will be rigorously assessed through a multi-faceted evaluation plan. We will use benchmark datasets from public repositories where established

 doi:10.20944/preprints202509.1105.v1

relationships are known. The system's output from the choice of statistical test to the final interpretation will be compared against the conclusions of expert human statisticians. Key metrics will include accuracy in test selection, precision in variable identification, and qualitative scores for the clarity and correctness of the generated summaries.

Upon successful validation, future work will focus on expanding the framework's capabilities. This includes increasing the library of supported statistical tests to encompass more complex designs (e.g., mixed-effects models, survival analysis) and enhancing its ability to generalize across diverse scientific domains and code languages (Shen et al., 2025; Huang et al., 2025). We also plan to develop an interactive refinement loop, allowing the system to ask clarifying questions when faced with ambiguous hypotheses. Ultimately, this framework has the potential to not only accelerate the research cycle but also to democratize data analysis, making robust scientific inquiry more accessible to a broader audience.

6. Conclusion

The increasing complexity and volume of data in modern research demand tools that are not only powerful but also intuitive and accessible. This proposal introduces a cognitive framework that leverages advances in natural language processing, statistical reasoning, and machine learning to automate the hypothesis testing process end-to-end. By translating natural language hypotheses into structured statistical operations, executing appropriate tests, and generating interpretable summaries, this system aims to reduce cognitive burden, enhance reproducibility, and accelerate the pace of scientific discovery.

Unlike existing tools that operate in isolated segments of the research workflow, our proposed framework offers a unified, modular approach capable of generalizing across domains. It addresses core limitations of scalability, user expertise requirements, and reproducibility by integrating LLMs, causal inference, and rule-based decision systems. With further development and validation, this framework has the potential to redefine how researchers interact with data transforming hypothesis testing from a manual, technical task into a seamless and intelligent process accessible to all.

References

1. Huang, K., Jin, Y., Li, R., Li, M. Y., Candès, E., & Leskovec, J. (2025). *Automated hypothesis validation with agentic sequential falsifications*. arXiv. https://doi.org/10.48550/arXiv.2502.09858
2. Movva, R., Peng, K., Garg, N., Kleinberg, J., & Pierson, E. (2025). *Sparse autoencoders for hypothesis generation*. arXiv. https://doi.org/10.48550/arXiv.2502.04382
3. Yang, Y., Wu, J., & Yue, Y. (2025). *Robust hypothesis generation: LLM-automated language bias for inductive logic programming*. arXiv. https://doi.org/10.48550/arXiv.2505.21486
4. Shen, Y., Liu, Y., Gong, M., Li, J., & He, Y. (2025). *A survey on hypothesis generation for scientific discovery in LLMs*. arXiv. https://doi.org/10.48550/arXiv.2504.05496
5. An, K., Si, S., Hu, H., Zhao, Y., & Liang, P. (2024). *Rethinking semantic parsing for large language models: Enhancing LLM performance with semantic hints*. arXiv. https://doi.org/10.48550/arXiv.2409.14469
6. Bollen, J., Eftekhar, M., Nelson, D. R., & Eckles, D. (2024). *Automating psychological hypothesis generation with AI. Humanities and Social Sciences Communications, 11*(1), 1–14. https://doi.org/10.1057/s41599-024-03407-5
7. Wang, R., Zelikman, E., Poesia, G., Chi, E. H., & Liang, P. (2023). *Hypothesis search: Inductive reasoning with language models*. arXiv. https://doi.org/10.48550/arXiv.2309.05660
8. Dror, R., Peled-Cohen, L., Shlomov, S., & Reichart, R. (2024). *Statistical significance testing for natural language processing*. Springer. https://doi.org/10.1007/978-3-031-02174-9
9. IBM Corp. (2021). *IBM SPSS Statistics for Windows* (Version 27.0) [Computer software]. https://www.ibm.com/products/spss-statistics
10. Free Software Foundation. (2023). *GNU PSPP* (Version 2.0.1) [Computer software]. https://www.gnu.org/software/pspp/

 doi:10.20944/preprints202509.1105.v1

11. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL-HLT 2019*, 4171–4186. https://doi.org/10.48550/arXiv.1810.04805
12. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901. https://doi.org/10.48550/arXiv.2005.14165
13. Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., & Androutsopoulos, I. (2021). Legal-BERT: The muppets straight out of law school. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, 2898–2904. https://doi.org/10.18653/v1/2020.findings-emnlp.261
14. Jurafsky, D., & Martin, J. H. (2023). *Speech and language processing* (3rd ed.). Draft manuscript. Stanford University. https://web.stanford.edu/~jurafsky/slp3/
15. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830. http://www.jmlr.org/papers/v12/pedregosa11a.html
16. Pearl, J. (2009). *Causality: Models, reasoning and inference* (2nd ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511803161