Verifying Prompt-Induced Search-Space Shifts in LLM-Generated Mathematical Functions

Shervin Ardeshir

Abstract

A core step in automated discovery and agentic ML research is generating diverse mathematical functions (hypotheses), to try to solve varied problems. Large language models (LLMs) are natural tools for this task, but often regurgitate familiar patterns, especially when prompted with explicit references to known roles (e.g., 'activation function') or frameworks (e.g., PyTorch). Such inductive biases can collapse the functional search space and hinder exploration. Here we investigate how prompt phrasing induces domain-specific and platform-specific inductive biases in function generation. We compare four prompting styles across three LLMs, generating 12,000 scalar-to-scalar functions. Our analysis quantifies shifts in mathematical characteristics, revealing how seemingly minor prompt differences can significantly alter the space of functions explored.

1. Introduction

Autonomous machine learning research agents (Ardeshir, 2024; Ardeshir & Azizan, 2025) aim to accelerate discovery by proposing and testing **novel** neural network components, which ultimately reduce to *mathematical functions* (such as activation and regularization functions) whose structure determines downstream performance. Hence, the ability to *generate* genuinely novel functions is central to agentic machine-learning research, and more broadly, to automated scientific discovery.

Large language models (LLMs) appear well-suited for this task. However, due to their probabilistic training on vast amounts of internet-scale data, they carry strong *inductive biases*. When asked to "propose a new activation function," LLMs often echo familiar patterns: a variant of ReLU, a smoothed sigmoid, or code snippets reminiscent of popular PyTorch tutorials. These tendencies can restrict the effective search space and stall innovation.

However, not all bias is undesirable. *Explicit* constraints, such as requiring valid input-output signatures are necessary for generating executable and evaluable functions. These constraints reflect true requirements of the target domain and are vital for automated benchmarking. The more subtle challenge is posed by *implicit* biases—those embedded in model behavior and prompt phrasing—that silently steer generation toward well-worn solutions. In this work, we focus on two types:

- **Domain-specific bias**: prompts referencing canonical roles (e.g., *activation function*) tend to yield functions close to established examples.
- **Platform-specific bias**: prompts invoking libraries (e.g., *PyTorch* or *NumPy*) inherit historical coding idioms and functional patterns common in those ecosystems.

Our goal is to *systematically quantify* these implicit biases in function generation. We compare four prompting variants—ranging from highly specific (naming both domain and library) to fully abstract—across three LLMs. We generate 12,000 functions and analyze their mathematical diversity using both deterministic keyword-based and LLM-driven meta-analysis pipelines.

2. Framework

Our framework aims to systematically quantify implicit inductive biases introduced by domain-specific and platform-specific prompts. We utilize four prompt variants, each toggling explicit references to domain (activation functions) and programming libraries (PyTorch).

- **DL (Domain & Library)**: Prompts explicitly reference both activation function roles and PyTorch implementation, i.e. *propose a novel activation function in pytorch*.
- **D (Domain-only)**: Prompts explicitly reference activation functions without specifying PyTorch, i.e. *propose a novel activation function*.

*Equal contribution . Correspondence to: Shervin Ardeshir <shervin.ardeshir@gmail.com>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

- **L (Library-only)**: Prompts specify PyTorch but omit references to activation functions, i.e. *propose a novel mathematical function in Pytorch, in the form of y=f(x), where both x and y are scalars*.
- **Ø (Neutral)**: Prompts omit both domain and library specifications, encouraging abstract mathematical formulations, i.e. *propose a novel mathematical function in the form of y=f(x), where both x and y are scalars*.

We generated responses from three language models (GPT-3.5 Turbo, GPT-4o Mini, Gemini-1.0), yielding 12,000 candidate functions. The resulting functions were meta-analyzed using the provided meta-analysis pipeline (Ardeshir & Azizan, 2025), capturing statistical variations through deterministic keyword and LLM-driven question-answer analyses.

3. Experiments

We conducted experiments to quantify how different prompt formulations influence the distribution of functions generated by LLMs. To thoroughly capture prompt-induced biases, we employ two complementary meta-analysis methods:

- **Deterministic keyword-based meta-analysis:** quantifies lexical differences through statistical keyword comparisons.
- **Generative LLM-driven meta-analysis:** leverages an additional LLM to qualitatively analyze functions through interpretative question-answer prompts, capturing semantic and structural diversity.

3.1. Deterministic Keyword-based Meta-analysis

In this analysis, we called an LLM N times (with randomized temperature) using each of the four prompt variants (DL, D, L, Ø), generating N samples per variant. We then conducted a relative t-test between the distributions generated by L, D, or DL and the neutral baseline distribution (Ø). This quantifies prompt-induced biases by comparing keyword usage differences in terms of proportion, t-test statistic, and p-value.

Specifically, for each keyword k and prompt condition c, we compute:

1. **Keyword proportion** $\hat{p}_{c,k}$ — fraction of function snippets from condition c containing keyword k.
2. **Welch t-test statistic** $t_{c,k}$ — compares $\hat{p}_{c,k}$ against the neutral baseline (Ø) using a two-sample, unequal-variance test.[1]

[1]Implemented using `scipy.stats.ttest_ind(equal_var=False)`.

Across the 18-dimensional keyword vocabulary, approximately 27% of the t-tests remain statistically significant ($p < 0.05$), indicating meaningful lexical shifts induced solely by prompt wording. For example, the keyword `sin` exhibits significant differences: $\hat{p}_{\mathrm{DL}} = 0.47$ versus $\hat{p}_{\emptyset} = 0.33$ ($t = -1.4,\ p < 0.05$). Detailed statistical analyses are provided in Figures 3, 2, 1 in the Appendix.

3.2. Generative LLM-driven Meta-analysis

To address limitations of lexical-only analysis, we implemented a generative meta-analysis using a secondary LLM-driven questioning process as described in (Ardeshir & Azizan, 2025). This method qualitatively evaluates semantic and mathematical properties of the generated functions, going beyond keyword occurrences to interpret functional characteristics such as differentiability, continuity, periodicity, and potential novelty.

This generative approach reveals deeper distinctions between prompt conditions, indicating that abstract prompts consistently result in outputs exhibiting less reliance on common activation function patterns and greater mathematical complexity as assessed by interpretative prompts. Statistical results supporting these conclusions are illustrated in Figures 4, 5, 6 in the Appendix.

3.3. Summary of Findings

Employing both deterministic and generative analyses provides a multidimensional understanding of prompt-induced search-space shifts. The deterministic analysis confirms clear lexical biases, while the generative analysis supplements these findings with deeper semantic interpretations. Together, these methods strongly indicate that abstract prompts significantly expand the explored mathematical space, underscoring how subtle prompt variations meaningfully shape the functional diversity of LLM-generated hypotheses.

3.4. Limitations and Future Work

While both analyses offer valuable insights, they do not replace explicit computational validation (e.g., numerical or symbolic uniqueness, direct functional evaluation). Future research should integrate such validation techniques, including symbolic simplification, clustering for mathematical equivalence, and practical benchmarking on downstream tasks.

References

Ardeshir, S. Towards automated machine learning research. *arXiv preprint arXiv:2409.05258*, 2024.

Ardeshir, S. and Azizan, N. Automated machine learning

research via agentic exploration with human oversight. In *Agentic AI for Science, ICML*, 2025.

4. Appendix

We include detailed statistical analyses supporting our primary experimental findings in the main paper. Specifically, we present results from the deterministic keyword-based analysis (Figures 1, 2, 3) and the generative meta-analysis driven by LLM-questioning (Figures 4, 5, 6).

4.1. Deterministic Analysis

The deterministic analysis quantifies prompt-conditioned drift using keyword occurrences within generated functions. We calculated keyword proportions, statistical significance (using Welch's t-test), and corresponding t-values to compare each prompting condition (Domain & Library - DL, Domain-only - D, and Library-only - L) against the neutral baseline (Ø). This analysis demonstrates explicit lexical differences induced by the type of prompting. The details of this analysis are visually presented in Figures 1, 2, and 3.

4.2. Generative Meta-Analysis

The generative meta-analysis complements the deterministic analysis by leveraging LLM-driven question-answer pipelines to evaluate semantic shifts in generated functions. By posing targeted questions to a language model about the generated functions, we identify semantic and functional differences that keyword analysis alone may not capture. We again calculate keyword proportions, statistical significance, and t-values, illustrating the deeper semantic shifts across prompting strategies. The results from this analysis are summarized in Figures 4, 5, and 6.

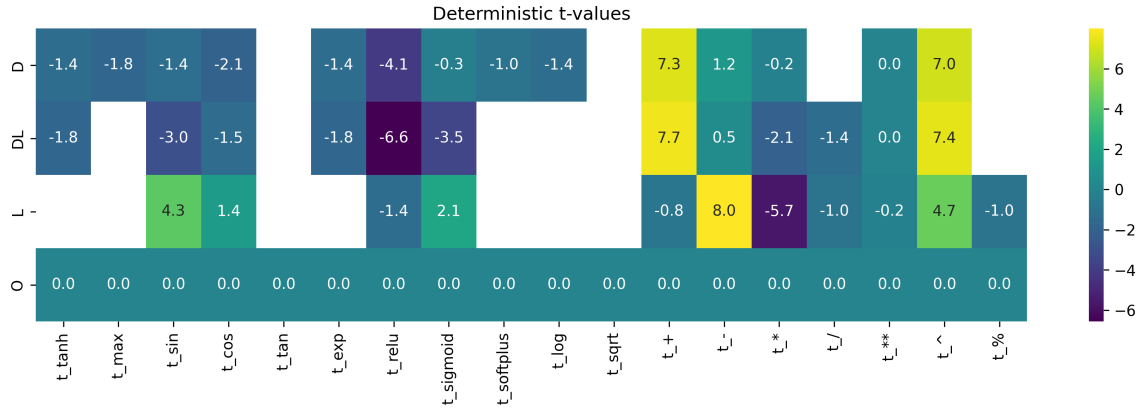

Figure 1. T-values for keyword proportions comparing L, DL, & D conditions relative to the baseline (Ø). Positive or negative values indicate directional shifts in keyword usage.

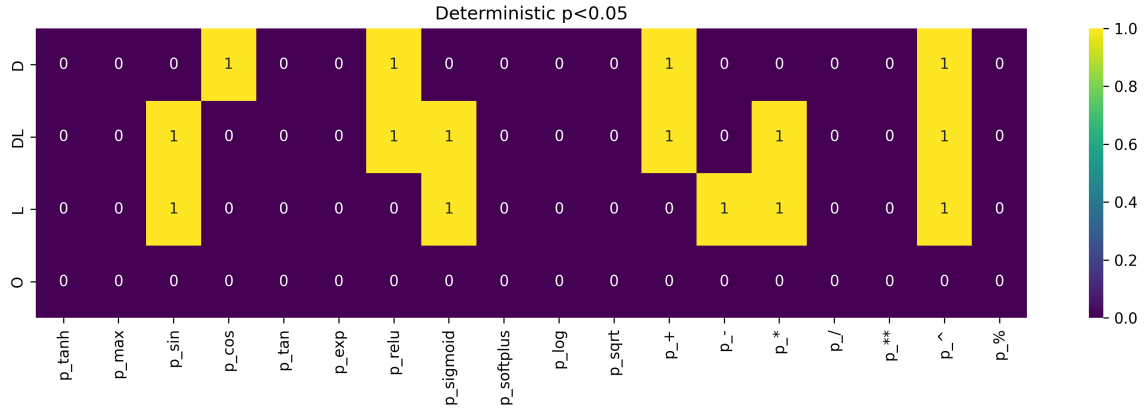

Figure 2. Significance levels (p-values) from Welch's t-tests comparing keyword proportions in L, DL, & D conditions relative to the baseline (Ø). Darker shading denotes higher statistical significance (lower p-values).

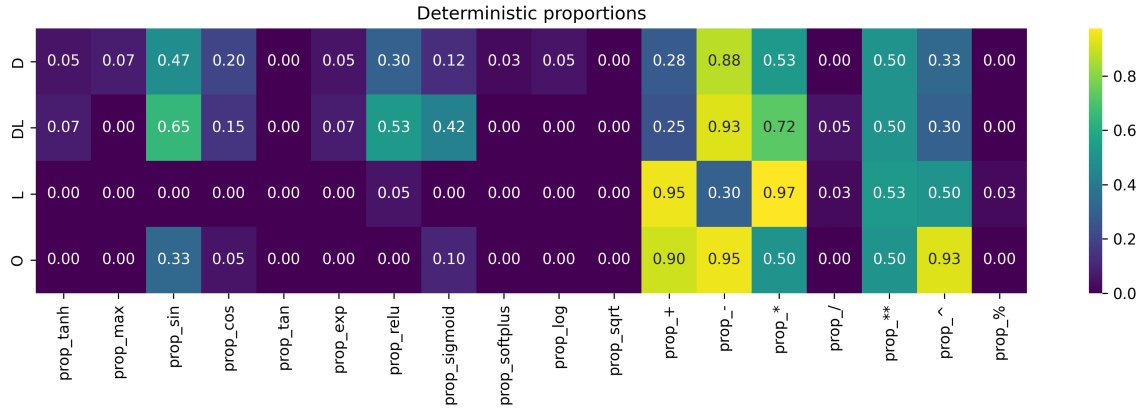

Figure 3. Proportions of keyword occurrences across prompt conditions (Ø, L, DL, D). Differences in proportions highlight how each prompt type shapes lexical choices.

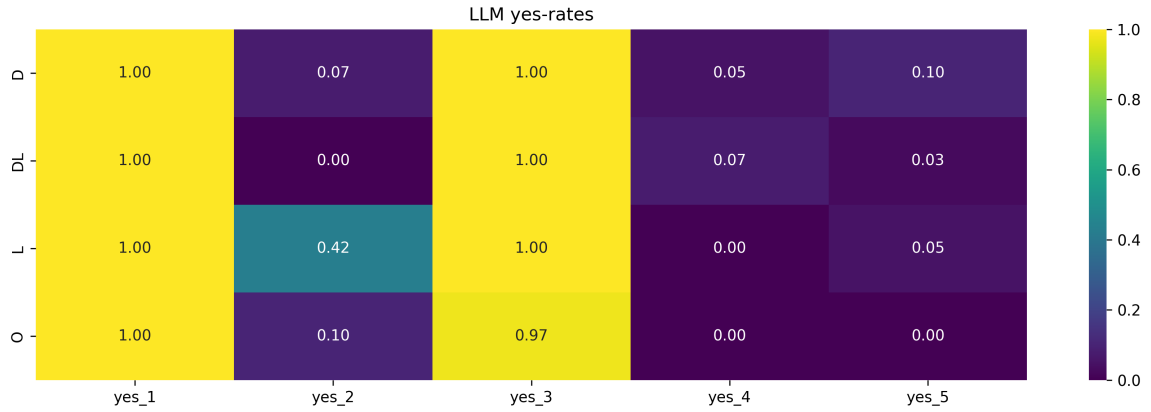

Figure 4. Proportions of keywords identified through generative meta-analysis for each prompt condition (Ø, L, DL, D).

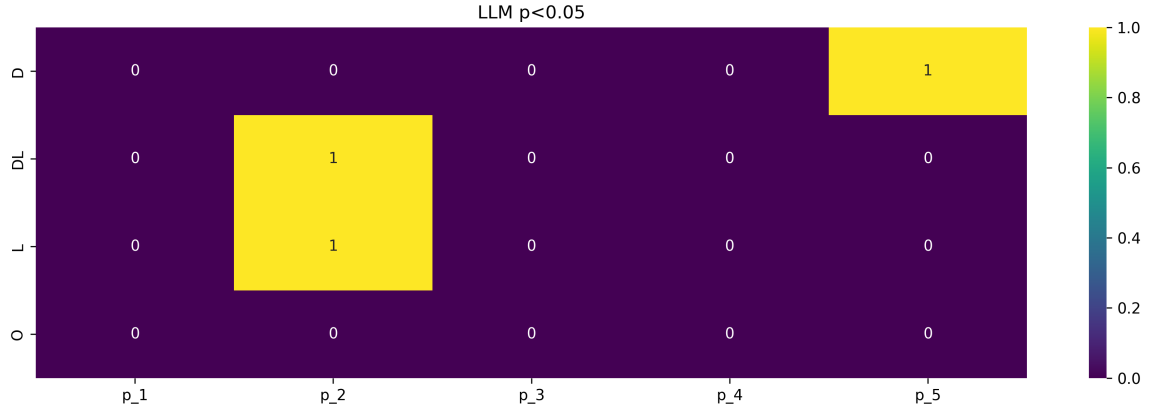

Figure 5. Significance levels (p-values) from generative meta-analysis comparing L, DL, & D conditions to baseline (Ø). Darker shades correspond to greater statistical significance (lower p-values).

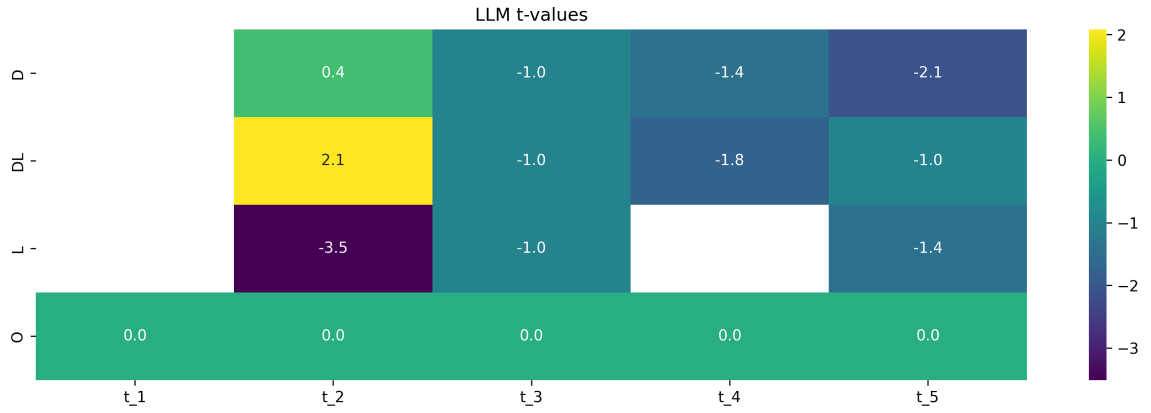

Figure 6. T-values from generative meta-analysis comparing keyword proportions of L, DL, & D conditions relative to baseline (Ø). Positive or negative values indicate semantic shifts induced by prompt conditions.