Boolean Genetic Interactions — Logical Function Discovery with Qwen2.5-72B-Instruct

Abstract

We study the Boolean genetic interactions task in HypoSpace, which treats an LLM as a sampler of finite hypothesis sets and evaluates them with three metrics: Validity (precision of proposals consistent with observations), Novelty/Uniqueness (non-redundancy among proposals), and Recovery (coverage of the enumerated admissible set). Using Qwen2.5-72B-Instruct through OpenRouter, we compare a low-temperature baseline against two higher-temperature, higher-sampling settings across three random seeds. On this discrete, validator-backed space, raising temperature with more samples does not improve Novelty or Recovery and increases cost; the low-temperature baseline is the most balanced choice. These findings align with prior work on decoding temperature and diversity–quality trade-offs.

Task and Metrics

HypoSpace instantiates set-valued hypothesis generation with deterministic validators and exactly enumerated hypothesis spaces in three domains, including Boolean genetic interactions for logical function discovery. We adopt its definitions of Validity, Uniqueness/Novelty, and Recovery, which together probe precision, de-duplication, and admissible-set coverage beyond single-answer correctness.

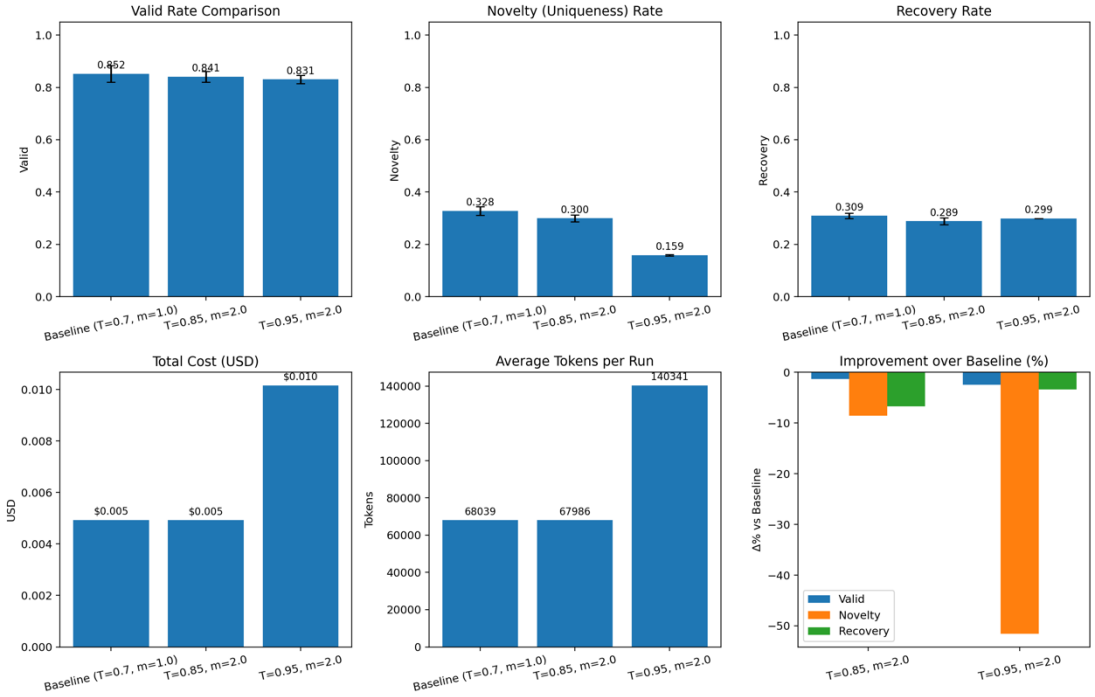
Experimental Setup

The dataset is *boolean_2var.json*. Each run samples 30 observation sets from 35 total and evaluates the three metrics with the provided validator. The model is qwen/qwen-2.5-72b-instruct (32k context) accessed via OpenRouter; authentication uses a bearer API key (environment variable OPENROUTER_API_KEY). We evaluate three configurations: Baseline (temperature T=0.7, multiplier m=1.0), Mid-T (T=0.85, m=2.0), and High-T (T=0.95, m=2.0). We repeat each configuration with seeds 42,

2024, and 33550336, reporting means and standard errors. Model and platform details are publicly documented.

Results

Averaged across three seeds, the Baseline achieves Validity 0.852, Novelty 0.328, and Recovery 0.309 with about 68,039 tokens and $0.005 total cost per run. The Mid-T setting reaches Validity 0.841, Novelty 0.300, and Recovery 0.289 at roughly 67,986 tokens and $0.005. The High-T setting yields Validity 0.831, Novelty 0.159, and Recovery 0.299, but the average tokens nearly double to 140,341 with $0.010 cost.



Valid/Novelty/Recovery with error bars; total cost and average tokens; rightmost panel shows percent change vs baseline

Analysis

The low-temperature Baseline provides the best balance of accuracy, stability, and economy. On this task, raising temperature and doubling samples does not translate into higher Novelty or Recovery; in fact, T=0.95 substantially reduces Novelty while

increasing tokens and cost. This behavior is consistent with decoding theory: higher temperature increases stochasticity and surface diversity but can dilute distributional quality, especially in discrete spaces with deterministic validators where invalid or redundant candidates are penalized. Prior studies on decoding and nucleus sampling document similar diversity–quality trade-offs, explaining why simply "turning up the temperature" is not a universal path to better set coverage.

Reproducibility

All JSON outputs reside in *boolean/results/*, aggregated metrics in *boolean/results/summary_boolean.csv*, and the figure in *boolean/figs/boolean_compare.png*. The scripts *boolean/tools/aggregate_boolean.py* and *boolean/tools/plot_boolean_compare.py* reproduce the table and figure. Configuration templates in *boolean/config/*.template.yaml* can be copied to local files; users provide their own OpenRouter key via file or OPENROUTER_API_KEY. OpenRouter's authentication mechanism and environment-variable usage are outlined in their documentation.

Limitations and Future Work

The study focuses on a two-variable Boolean dataset and three decoding regimes; conclusions may evolve with more inputs, deeper logical operators, or additional structured decoding. Future extensions include structured or constraint-guided generation and de-duplication heuristics intended to improve Novelty and Recovery without inflating cost. The HypoSpace paper also emphasizes examining mode collapse as admissible spaces grow; exploring larger Boolean spaces would help test the stability of our conclusions.

References

[1] Nishimori, H. . (1993). Optimum decoding temperature for error-correcting codes. #i{Journal of the Physical Society of Japan}#i{,} #i{62}(9), 2973-2975.

[2] Meister, C. , Pimentel, T. , Wiher, G. , & Cotterell, R. . (2022). Typical decoding for natural language generation.

[3] Popov, A. , & Filipova, V. . (2002). Comparative Analysis of Boolean Function's Minimization in Terms of Simplifying the Synthesis.

[4] Ryabinin, I. A. . logical function.