On Prompting, Priors, and What It Takes for LLMs to Produce Novelty

Max Pagels

May 27, 2026

Disclaimer: This is not a research paper but rather the author's own thoughts. It does not claim originality, and is intended as a sketch of a mental model for thinking about the problem of novelty discovery with LLMs.

The typicality problem

Autoregressive language models are trained to estimate the conditional

$$p(x_t \mid x_{<t}),$$

the probability of the next token given those that preceded it. Generation at inference is repeated sampling from this same conditional: the model samples from

$$p(y \mid x) \;=\; \prod_t p(y_t \mid y_{<t},\, x).$$

where x is the prompt and y the produced continuation of tokens.

The learned distribution sampled at inference is naturally shaped by the training data. Generation is, by construction, biased toward sequences and statistical constructs that are probable under the training distribution. Empirically, the choice of x matters: semantically similar prompts can produce qualitatively different outputs, and certain prompt structures reliably improve task performance.

Consider a hypothetical language model capable of superhuman performance for a given problem. Even such a capable model will tend, under typical prompting, to generate outputs that lie in high-density regions of its learned distribution. In other words, responses that are statistically typical given its training data. Since the model is trained to continue sequences in this way, typical prompts may bias generation toward conventional or corpus-representative solutions rather than genuinely novel ones.

Novel outputs are still possible: the model can generate sequences that never appeared verbatim in the training corpus by recombining learned abstractions compositionally. In this author's experience, however, genuinely original or unconventional solutions require prompts that steer inference toward less typical

regions of the conditional distribution $p(y \mid x)$, and this is in line with what many LLM users report: for non-trivial problems, iterative prompt refinement is frequently necessary before the model converges on an especially insightful solution. An unconventional prompter might discover something highly unconventional under this regime, but is likely biased not to, as we shall see by formalising this implicit process.

Prompting as search

There is a useful reframing here: prompting is a form of search. The user explores $p(y \mid x)$ by perturbing x, with human intuition serving as the heuristic that selects which directions to probe. The framing is useful precisely because it exposes a structural problem. The user's intuitions about which prompts are promising were shaped by roughly the same corpus the model was trained on; prompter and model share at the very least a partial statistical prior. A heuristic correlated with the distribution it is meant to search past will tend to return the searcher to typical regions — precisely the regions where novel solutions are least likely to be. The phenomenon is familiar in practice: attempts to coax a novel solution from an LLM often produce a chain of thought that circles back on itself, yielding little but frustration. If we subscribe to the idea that genuine breakthroughs happen as a function of bias-breaking novelty, it is no wonder we have seen little scientific progress made solely using manual prompts.

Requirements for novelty

The arguments above suggest two requirements for LLMs to produce genuinely novel contributions. First, exploration must be algorithm-driven rather than user-driven: the search over $p(y \mid x)$ needs a mechanism that is not anchored to the same prior that shaped the distribution being searched. Second, exploration alone is not enough. Atypicality does not imply utility, and without a signal that distinguishes the two, exploration collapses into a random walk. A feedback loop is required: a way for outcomes to be evaluated and for that evaluation to update subsequent generation. Together with the generator itself, these two requirements define the minimal structure any system aimed at systematic production of novel yet useful solutions must have.

The feedback signal cannot come from humans either in search of this goal. Human evaluators share the prior that the search is meant to escape; soliciting their input introduces precisely the bias the exploration mechanism was designed to overcome. This author takes this as a direct argument against RLHF as a route to novelty. The procedure is well-suited to aligning model behavior with human preferences, and that is a legitimate goal, but those preferences are themselves drawn from the typical region of the distribution, which pulls the policy back toward typical modes rather than away from them. What is needed instead is a feedback source whose judgments are not entangled with

the prior: some oracle whose verdicts depend on the output itself rather than on its resemblance to familiar outputs.

A framework for novelty discovery

Based on the above discussion, we can design framework for systematic novelty discovery. The author makes no claims on originality; rather, the framework is one amalgamated based on personal experience. Three components are seemingly required:

- a *generator* $\pi_\theta(y \mid x)$, a pretrained autoregressive language model treated as a stochastic policy over output trajectories y given context x;
- an *exploration algorithm* $\mathcal{E}$ that operates on top of π_θ and produces a set of candidate trajectories $\{y^{(1)}, \ldots, y^{(k)}\}$ for a given x;
- an *evaluator* $\mathcal{V}$ that assigns a signal to each candidate trajectory and feeds that signal back into subsequent generation.

The generator supplies the prior over plausible continuations and does the work of keeping candidate outputs coherent; without it, the search space is the set of all token sequences, which is far to large and almost entirely incoherent. π_θ is stochastic in the sense that it defines a distribution from which trajectories are sampled rather than a deterministic map from prompt to output; this is what allows repeated invocations to yield distinct candidates for $\mathcal{E}$ to work with.

$\mathcal{E}$ is not π_θ. It may invoke the generator repeatedly, branch at chosen points, perturb intermediate states, or combine partial trajectories, but its role is to deliberately allocate sampling effort away from the modes of π_θ and toward the atypical-but-plausible regions where novel solutions are likeliest to live. Direct sampling from π_θ, even with high temperature or nucleus parameters, does not count: those operations broaden the distribution but may not redirect it towards fruitful regions of the search space.

The constraint on $\mathcal{V}$ is the one developed above: its judgments must not be entangled with the prior that $\mathcal{E}$ is designed to escape. A verifier, a test suite, an empirical measurement, or any oracle whose verdict depends on the output itself rather than on its resemblance to familiar outputs satisfies the constraint; a human evaluator or a reward model trained on human preferences does not. Statistically, $\mathcal{V}$ must be independent of π_θ in the sense that its signal is not correlated with the likelihood of candidates under π_θ.

The loop closes when $\mathcal{V}$'s signal feeds back into subsequent generation. The framework is deliberately agnostic about how this happens. The signal can update the generator's parameters via gradient-based fine-tuning, in which case π_θ itself shifts and the framework reduces to a particular form of reinforcement learning. It can update the exploration algorithm without touching π_θ, as in search-tree methods that backpropagate value estimates over partial trajectories while the underlying policy is held fixed. It can enter the context window as part

of the next prompt, in which case the update is in-context and ephemeral. It can operate at the population level over a pool of candidates, as in evolutionary methods. In this author's view, treating these as interchangeable implementation details obscures what is actually a set of distinct design choices with distinct consequences, and one of the things the framework is meant to do is force the choice into the open and guide possibly fruitful research avenues.

Two properties of the framework are worth underlining. First, the three components are minimal in the sense that removing any one of them collapses the system back to a regime the previous sections argued against. Without $\mathcal{E}$, generation is direct sampling from π_θ and returns to preferring modes. Without $\mathcal{V}$, exploration has no signal to distinguish productive atypicality from noise and degenerates into a random walk. Without π_θ, there is no prior over coherent outputs and the search has nothing to explore from. Second, the framework is intended to be descriptive as well as prescriptive: existing approaches that produce non-trivial outputs from LLMs can, the author believes, be located within it, with differences between them corresponding to choices about how $\mathcal{E}$ is structured, what signal $\mathcal{V}$ provides, and how the feedback updates the system. A systematic mapping is beyond the scope of this paper.

Discussion

LLMs have proven incredibly powerful, yet we have not seen true scientific breakthroughs emerge from them at the frequency one might hope. The typicality problem is may be on reason why, and necessitates a shift in how we think about prompting and the design of systems aimed at novelty discovery. The framework outlined above is the author's attempt to formalise the implicit process of iterative prompting and to identify the structural requirements for systematic novelty discovery. It is intentionally minimal; it does not prescribe an exploration algorithm, a particular form of evaluator, or a specific rule for closing the feedback loop. What it does is identify the components any system aimed at novel outputs must contain and the constraint each must satisfy: a generator that supplies a prior over coherent trajectories, an exploration mechanism that is not anchored to that prior, and an evaluator whose judgments are not entangled with it. In this author's view, each of these is forced by the typicality argument and not a matter of taste. Novelty requires moving off the modes of π_θ; moving off the modes requires exploration that does not inherit the prior; and distinguishing productive atypicality from noise requires an evaluator that does not inherit it either. A system missing any of the three is not a candidate for producing genuinely novel work *systematically*, whatever else it may be good for.

This author's reason for stating things in this form is that the question of whether LLMs can produce genuinely new work is too often treated as a matter of speculation when it should be a matter of design. Within the framework, the question becomes concrete: can exploration algorithms and evaluators satisfying the stated constraints be constructed, is the feedback they produce sufficient

to shift probability mass off the prior's modes, and can these be done in a computationally efficient manner? These are the right questions, and hopefully answerable ones. The framework itself can not answer them, but in this author's view, it does provide the setting in which an answer might be found.