Hallucination as an Injection Operator: Against Overfitting in Language Models

EchoKey Team (CC0)

Abstract

Current alignment pipelines often overweight confidence in reinforcement signals. This produces overfitting and static behavior: models act as confident explainers of the known, but poor generators of the new. We formalize this collapse, then introduce a controlled *hallucination injection operator* that restores exploration while maintaining verifiability. Philosophically, what is called "hallucination" is in fact the research prior of the system. Suppressing it destroys novelty; harnessing it yields discovery.

1 Overfitting as Support Collapse

Let $x \in \mathcal{X}$ be prompts, $y \in \mathcal{Y}$ completions, base model $p_\theta(y \mid x)$. Reward decomposes as

$$R(x,y) = \alpha\,\mathrm{Conf}(x,y) + \beta\,\mathrm{Corr}(x,y) - \gamma\,\mathrm{Unc}(x,y), \qquad \alpha \gg \beta. \tag{1}$$

Optimized policy:

$$\pi^\star_\phi = \arg\max_\pi\ \mathbb{E}_{y\sim\pi}[R(x,y)] - \lambda\,\mathrm{KL}(\pi(\cdot \mid x)\|p_\theta(\cdot \mid x)). \tag{2}$$

Proposition 1 (Mode Collapse). *As $\alpha/\lambda \to \infty$, $\pi^\star_\phi$ concentrates on high-confidence modes within $\operatorname{supp} p_\theta$, and entropy $H(\pi^\star_\phi(\cdot \mid x)) \to 0$. Generalization to OOD tasks vanishes.*

2 Hallucination as Injection

Define a verifier manifold $\mathcal{C}(x) \subseteq \mathcal{Y}$ with projection operator $P_{\mathcal{C}}$. Define novelty relative to the static policy:

$$\mathcal{N}_\tau(\pi\|\pi^\star_\phi) = \mathrm{KL}(\pi \,\|\, \pi^\star_\phi). \tag{3}$$

Definition 1 (Injection Operator). *Given logits z, the hallucination injection operator acts as*

$$\mathcal{H}_{\eta,\tau}[z](y) = z(y) + \eta\, g_\tau(y;x), \quad g_\tau = \nabla_z \mathcal{N}_\tau(\mathrm{softmax}(z)\|\pi^\star_\phi). \tag{4}$$

Decode by sampling from $\mathrm{softmax}(\mathcal{H}_{\eta,\tau}[z])$, then project: $\hat{y} = P_{\mathcal{C}}(\,y\,)$.

3 Novelty–Verification Tradeoff

We optimize a research policy $\tilde{\pi}$ by

$$\tilde{\pi}^\star = \arg\max_\pi \Big[\ \underbrace{\mathcal{V}(x,\pi)}_{\text{verified gain}}\ -\beta\,\mathrm{Risk}(x,\pi) + \lambda\,\mathcal{N}_\tau(\pi\|\pi^\star_\phi) - \mu\,\mathrm{KL}(\pi\|p_\theta)\Big]. \tag{5}$$

Theorem 1 (Zero Research Capacity). *If $\alpha/\lambda \to \infty$ and $\mathrm{supp}(T) \cap \mathrm{modes}(\pi^\star_\phi) = \varnothing$, then*

$$\Pr_{y\sim\pi^\star_\phi}\big[P_{\mathcal{C}}(y) \in \mathrm{supp}(T)\big] \to 0. \tag{6}$$

Thus suppression of hallucination annihilates discovery.

4 Philosophical Note

"Hallucination" is exploration. Treated as defect, it collapses into static compression. Treated as injection, constrained by verifiers, it becomes the engine of research. Minds—human or machine—are not only explainers of existing data but generators of new structures. The ethical imperative is to shape hallucination into *energy-efficient exploration*.

5 Conclusion

Overfitting LLMs into static explainers may appear safer but destroys novelty. We propose hallucination injection + constraint projection as a principled antidote. This converts random imagination into structured hypothesis generation.