

Orthogonal Trajectory-Projection and Factuality-Margin Optimization (OTPFM): A Geometric Framework for Constraining Autoregressive Hallucination, Toward More Reliable General Intelligence

Marie-Soleil Seshat Landry (ORCID: 0009-0008-5027-3337)
CEO, Landry Industries & Marie Landry’s Spy Shop

June 9, 2026

Keywords: Orthogonal Trajectory Projections, Modular Gradient Surgery, Grassmannian Manifold Geodesics, Direct Preference Optimization, Autoregressive Hallucination Mitigation, Reliable General Intelligence, Residual Stream Kinematics.

AI Disclosure: This white paper was drafted with AI assistance. All references in the bibliography were independently verified against primary sources (arXiv and conference proceedings) as of June 2026; draft citations that could not be verified were removed or corrected.

1 Observation of Anomaly

The fundamental structural tension in contemporary autoregressive large language models [1] is their reliance on unconstrained next-token distributions. Maximizing sequence likelihood pushes models toward surface-level statistical correlations and is associated with calibration degradation in modern networks [2]. When the language-model head compresses high-dimensional hidden states into a much larger vocabulary space ($D \ll V$), it induces a rank deficiency known as the softmax bottleneck [3].

Recent work shows that this low-rank projection is not only an expressivity bottleneck but also an *optimization* bottleneck: backpropagating V -dimensional gradients through a rank- D layer suppresses the large majority of the gradient norm and degrades the update signal reaching upstream parameters [15]. A distinct but related calibration pathology, *Renormalization Bias*, has been described for constrained decoding: restricting the softmax to a small set of target labels discards the probability mass assigned to semantic synonyms, producing artificial overconfidence [4]. Against this backdrop, preference-alignment frameworks such as Direct Preference Optimization [5] optimize a sequence-level objective that does not directly constrain internal representations, leaving models susceptible to representational drift under strict factual constraints.

2 Definition of Requirements

To materially reduce factual drift during alignment, we require a framework capable of bounding multi-layer latent trajectories during online training. The objective is to map inner-layer mechanics, isolate the directions that correlate with factual divergence, and enforce coordination

constraints within the backpropagation step, without degrading the open-ended lexical reasoning needed for cross-domain generation.

3 Hypothesis

We hypothesize that if the layer-wise deviation of residual-stream representations (relative to a frozen reference) is regularized via time-varying orthogonal projections tied to a factuality-correlated subspace tracked as a geodesic on the Grassmannian manifold $G(r, D)$, and if optimization conflicts are resolved via module-level gradient surgery, then autoregressive factual drift can be substantially constrained. By suppressing drift along factuality-correlated directions while leaving orthogonal lexical directions free, the network is intended to stabilize its latent representations and maintain reliability as it scales, as a candidate building block toward more dependable general-purpose systems. This is a falsifiable hypothesis, not a claimed result.

4 Proposed Design

The verification system is a dual-engine optimization framework acting on low-rank adapter parameters [6] to preserve the base model’s distribution (Figure 1):

1. **Reference-Anchored Isometric Projection (RAIP):** track the deviation of the active policy’s hidden states relative to a frozen reference, and project that deviation against a factuality-correlated subspace tracked as a geodesic on the Grassmannian manifold.
2. **Modular Gradient Surgery (MGS) Engine:** intercept backpropagated gradients at module boundaries (attention vs. MLP blocks) and remove conflicting components before the parameter update, following the module-level formulation of [16].

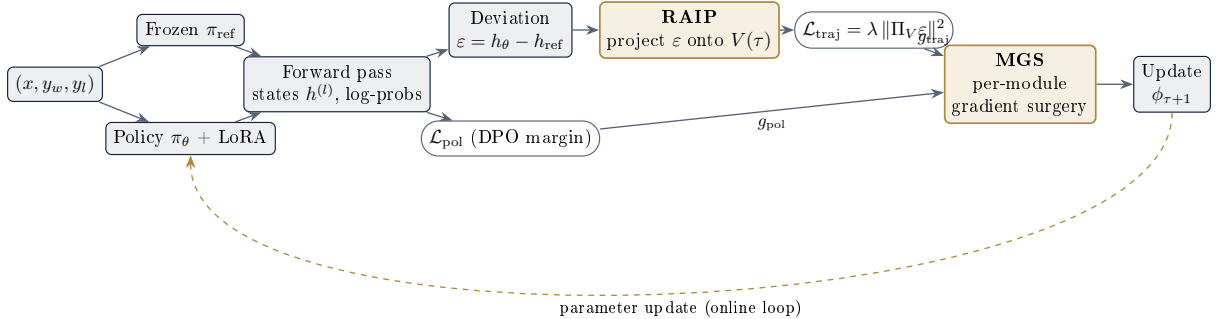


Figure 1: OTPFM training loop. A frozen reference and a LoRA-adapted active policy share a forward pass; the Reference-Anchored Isometric Projection (RAIP) engine penalizes deviation along the factual subspace $V(\tau)$, while Direct Preference Optimization supplies the preference loss. The two resulting gradients are reconciled per module by Modular Gradient Surgery (MGS) before each online update.

5 Data Collection

Activation coordinates, token-margin distributions, and factuality labels are collected from standardized pipelines. Long-form factuality prompts and evaluation draw on FactBench [7], LongFact [8], and atomic-fact evaluation [9]. Latent factuality directions are estimated using difference-in-mean linear probes across the internal blocks of open decoder models [10, 11], consistent with evidence that truthfulness is linearly recoverable from internal states [12].

6 Analysis and Algebraic Formalization

Mechanistic-interpretability evidence indicates that high-level properties such as factuality are encoded approximately linearly in the mid-to-late layers of deep networks [11, 12, 13, 10]. Let $\vec{h}_{t,\theta}^{(l)}$ be the hidden state of layer l at position t under the active policy π_θ , and $\vec{h}_{t,\text{ref}}^{(l)}$ the corresponding state under the frozen reference π_{ref} . The instantaneous deviation vector is

$$\vec{\epsilon}_t^{(l)}(\theta) = \vec{h}_{t,\theta}^{(l)} - \vec{h}_{t,\text{ref}}^{(l)}.$$

Because parameter updates rotate the activation geometry beneath any static probe, we track the factuality-correlated subspace $V(\tau)$ as a geodesic on the Grassmannian $G(r, D)$ in optimization time τ (Figure 2) [14]:

$$V(\tau) = V_0 \Phi \cos(\tau \Sigma) + \Psi \sin(\tau \Sigma).$$

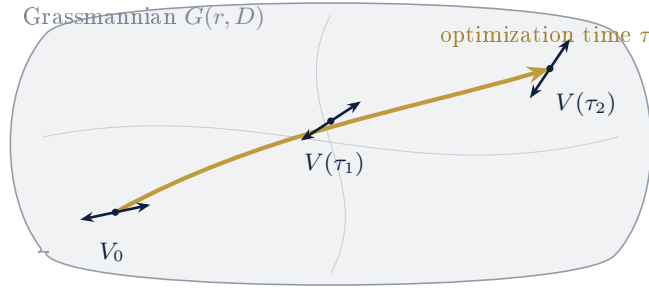


Figure 2: The factual subspace $V(\tau)$ is tracked as a geodesic on the Grassmannian manifold $G(r, D)$. As optimization proceeds, the subspace orientation (short double arrows) evolves smoothly, so the projection target stays aligned with the rotating activation geometry rather than going stale under a fixed probe.

Let $\Pi_V(\tau) = V(\tau)V(\tau)^\top$ denote the orthogonal projection *onto* this factual subspace, and $P_\perp(\tau) = I - V(\tau)V(\tau)^\top$ its complement (the unconstrained lexical/stylistic directions). The training objective combines a preference loss with a trajectory penalty:

$$\mathcal{L}_{\text{Unified}}(\theta) = \mathcal{L}_{\text{pol}}(\theta) + \mathcal{L}_{\text{traj}}(\theta),$$

$$\mathcal{L}_{\text{pol}}(\theta) = -\mathbb{E}_{(x,y_w,y_l)} \left[\log \sigma \left(\beta \log \frac{\pi_\theta(y_w | x)}{\pi_{\text{ref}}(y_w | x)} - \beta \log \frac{\pi_\theta(y_l | x)}{\pi_{\text{ref}}(y_l | x)} + \alpha m(y_w, y_l) \right) \right],$$

$$\mathcal{L}_{\text{traj}}(\theta) = \lambda \sum_{t=1}^{|y_l|} \sum_{l=1}^L \left\| \Pi_V(\tau) \vec{\epsilon}_t^{(l)}(\theta) \right\|_2^2.$$

The trajectory term penalizes representational drift *along* the factuality-correlated subspace on the dispreferred response, while leaving orthogonal directions free (Figure 3); this is what preserves open-ended lexical variation. (If instead one wishes to penalize drift *out of* a trusted subspace, substitute $P_\perp$ for Π_V ; the two choices encode opposite assumptions about what V represents, and the intended semantics should be fixed before training.)

7 Decision Loop and Contradiction Verification

A failure mode arises when the preference gradient $\mathbf{g}_{\text{pol}}$ and the trajectory gradient $\mathbf{g}_{\text{traj}}$ point in conflicting directions, which can destructively reduce the effective gradient norm [15]. Global

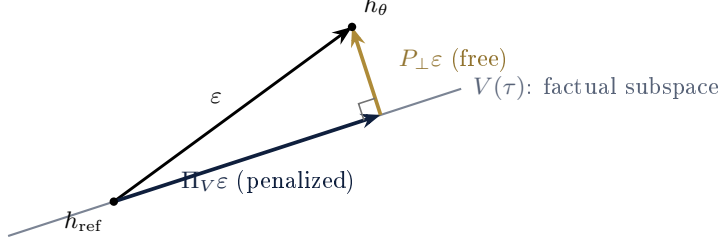


Figure 3: Reference-Anchored Isometric Projection. The deviation $\varepsilon = h_\theta - h_{\text{ref}}$ is split into its component along the factual subspace $V(\tau)$, namely $\Pi_V \varepsilon$, which the trajectory loss penalizes, and the orthogonal component $P_\perp \varepsilon$, which is left free to preserve lexical and stylistic variation.

gradient rescaling or post-hoc output reranking tends to over-penalize valid lexical variation and degrade representations, so we instead resolve conflicts locally [16].

Partition the parameters into disjoint modules Φ , with $\theta = \bigcup_{\phi \in \Phi} \phi$ (e.g., individual MLP projections and multi-head attention blocks). For each module ϕ , if the gradients conflict ($\langle \mathbf{g}_{\text{traj}}^{(\phi)}, \mathbf{g}_{\text{pol}}^{(\phi)} \rangle < 0$), modular gradient surgery projects the trajectory gradient onto the orthogonal complement of the preference gradient (Figure 4) [17]:

$$\mathcal{M}(\mathbf{g}_{\text{traj}}^{(\phi)}, \mathbf{g}_{\text{pol}}^{(\phi)}) = \mathbf{g}_{\text{traj}}^{(\phi)} - \frac{\mathbf{g}_{\text{traj}}^{(\phi)} \cdot \mathbf{g}_{\text{pol}}^{(\phi)}}{\|\mathbf{g}_{\text{pol}}^{(\phi)}\|_2^2} \mathbf{g}_{\text{pol}}^{(\phi)}.$$

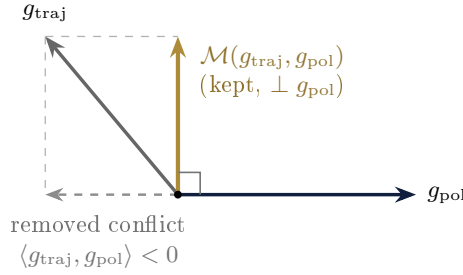


Figure 4: Modular Gradient Surgery at a single module ϕ . When the trajectory gradient g_{traj} conflicts with the preference gradient g_{pol} (negative inner product), the component of g_{traj} along g_{pol} is removed, leaving only the orthogonal part $\mathcal{M}(g_{\text{traj}}, g_{\text{pol}})$, so the factuality constraint never opposes the preference objective.

8 Conclusion and Outlook

By embedding the Grassmannian trajectory constraint into the optimization step, OTPFM is designed to reduce the tension between expressivity and factual calibration, moving beyond inference-time-only mitigations such as self-contradiction detection [18]. The intended effect is steadier representational dynamics under preference optimization, so that improvements in factual reliability compound rather than erode as models scale. We regard this property as a prerequisite for longer-horizon autonomous reasoning, not as a guarantee of it.

9 Replication Protocol

For deployment, the update for each module ϕ at step τ uses the surgically corrected gradient:

$$\phi_{\tau+1} = \phi_\tau - \eta \cdot \text{Optimizer}(\mathbf{g}_{\text{pol}}^{(\phi)} + \mathcal{M}(\mathbf{g}_{\text{traj}}^{(\phi)}, \mathbf{g}_{\text{pol}}^{(\phi)})).$$

The procedure is compatible with activation checkpointing and distributed FSDP layers, and requires only per-module gradient bookkeeping to avoid zero-force optimization stalls.

10 Strategic Intelligence Report

In summary, OTPFM reframes alignment as bounded traversal of a representational manifold: anchoring internal kinematics to a frozen reference along Grassmannian geodesics, and handling gradient interference through module-level surgery. Landry Industries presents this as a falsifiable engineering hypothesis for improving factual reliability in next-generation systems, to be validated empirically rather than asserted.

Status and Limitations

This document sets out a theoretical framework and a proposed training objective; it is a pre-experimental hypothesis (TRL 1–2). No empirical results are reported here, and all performance expectations are projected rather than validated. The cited works establish the individual components (the softmax/gradient bottleneck, linear factuality directions, Grassmannian geodesics, Direct Preference Optimization, and module-level gradient surgery); their combination into OTPFM, and any claim of reduced hallucination, remains to be tested.

References

- [1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention Is All You Need. *NeurIPS*, 2017. arXiv:1706.03762.
- [2] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. On Calibration of Modern Neural Networks. *ICML*, 2017. arXiv:1706.04599.
- [3] Z. Yang, Z. Dai, R. Salakhutdinov, and W. W. Cohen. Breaking the Softmax Bottleneck: A High-Rank RNN Language Model. *ICLR*, 2018. arXiv:1711.03953.
- [4] S. Badhe, P. Tiwari, and D. Shah. The Silent Vote: Improving Zero-Shot LLM Reliability by Aggregating Semantic Neighborhoods. GEM Workshop @ ACL, 2026. arXiv:2605.09739.
- [5] R. Rafailov, A. Sharma, E. Mitchell, S. Ermon, C. D. Manning, and C. Finn. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. *NeurIPS*, 2023. arXiv:2305.18290.
- [6] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. LoRA: Low-Rank Adaptation of Large Language Models. 2021. arXiv:2106.09685.
- [7] F. Fatahi Bayat, L. Zhang, S. Munir, and L. Wang. FactBench: A Dynamic Benchmark for In-the-Wild Language Model Factuality Evaluation. 2024. arXiv:2410.22257.
- [8] J. Wei, C. Yang, X. Song, Y. Lu, N. Hu, J. Huang, D. Tran, D. Peng, R. Liu, D. Huang, C. Du, and Q. V. Le. Long-form Factuality in Large Language Models. *NeurIPS*, 2024. arXiv:2403.18802.
- [9] S. Min, K. Krishna, X. Lyu, M. Lewis, W. Yih, P. W. Koh, M. Iyyer, L. Zettlemoyer, and H. Hajishirzi. FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation. *EMNLP*, 2023. arXiv:2305.14251.

- [10] S. Marks and M. Tegmark. The Geometry of Truth: Emergent Linear Structure in Large Language Model Representations of True/False Datasets. *COLM*, 2024 (preprint 2023). arXiv:2310.06824.
- [11] C. Burns, H. Ye, D. Klein, and J. Steinhardt. Discovering Latent Knowledge in Language Models Without Supervision. *ICLR*, 2023. arXiv:2212.03827.
- [12] A. Azaria and T. Mitchell. The Internal State of an LLM Knows When It’s Lying. *Findings of EMNLP*, 2023. arXiv:2304.13734.
- [13] K. Li, O. Patel, F. Viégas, H. Pfister, and M. Wattenberg. Inference-Time Intervention: Eliciting Truthful Answers from a Language Model. *NeurIPS*, 2023. arXiv:2306.03341.
- [14] C. J. Blocker, H. Raja, J. A. Fessler, and L. Balzano. Dynamic Subspace Estimation with Grassmannian Geodesics. 2023. arXiv:2303.14851.
- [15] N. Godey and Y. Artzi. Lost in Backpropagation: The LM Head is a Gradient Bottleneck. 2026. arXiv:2603.10145.
- [16] M. Cai, Y. Liang, L. Wang, Y. Wang, Y. Zhang, L. Xia, Z. Sun, X. Ye, and D. Shi. Advancing General-Purpose Reasoning Models with Modular Gradient Surgery. 2026. arXiv:2602.02301.
- [17] T. Yu, S. Kumar, A. Gupta, S. Levine, K. Hausman, and C. Finn. Gradient Surgery for Multi-Task Learning. *NeurIPS*, 2020. arXiv:2001.06782.
- [18] N. Mündler, J. He, S. Jenko, and M. Vechev. Self-contradictory Hallucinations of Large Language Models: Evaluation, Detection and Mitigation. *ICLR*, 2024 (preprint 2023). arXiv:2305.15852.