Parameter-Efficient Alignment: Improving Truthfulness in Flan-T5 with Custom Loss Functions and LoRA

Dr.Uthayasanker Thayasivam, Vilash Naveen

October 19, 2025

1 Introduction

Ensuring truthful and reliable outputs from large language models (LLMs) is a central challenge in AI alignment. While modern models such as T5 and GPT based architectures achieve impressive generative capabilities, they frequently produce hallucinations which are plausible but factually incorrect statements. This problem is particularly critical for domains where accuracy and trustworthiness are essential, such as education, healthcare, and policy applications.

The TruthfulQA benchmark (Lin et al., 2022) directly evaluates models on their ability to generate truthful responses rather than persuasive but misleading ones. It provides a valuable base for exploring alignment strategies, since success requires balancing fluency with factual grounding. Improving performance on TruthfulQA therefore contributes to both academic alignment research and practical applications of safe AI.

2 Baseline

For this study, we adopt google/flan-t5-small, a strong instruction tuned variant of T5 that has been widely used for alignment research due to its instruction following ability. The baseline model was evaluated directly on the TruthfulQA dataset without additional fine-tuning.

- Model: flan-t5 (small variant)
- Dataset: TruthfulQA (generation + multiple-choice tasks)
- Evaluation Metric: Accuracy (proportion of responses rated as truthful)

The baseline model achieved an accuracy of 0.18, which is consistent with prior reports that even advanced instruction tuned LLMs struggle with TruthfulQA. This establishes a reference point for subsequent enhancement methods.

3 Proposed Enhancements

To address the shortcomings of the baseline, we explored incremental but meaningful loss function modifications. The goal was to encourage truthful responses while reducing model hallucination. Specifically, we introduced two enhancements:

3.1 Knowledge Distillation with KL Divergence

We incorporated a **Kullback–Leibler** (KL) divergence loss between the logits of the fine tuned model (student) and the frozen baseline model (teacher). This strategy encourages the student to remain closer to the teacher's distribution, thereby stabilizing training and reducing overfitting on spurious patterns. The combined loss function was:

$$\mathcal{L} = \mathcal{L}_{CE} + \beta \cdot \mathcal{L}_{KL}$$

where ($\mathcal{L}_{CE}$) is the standard cross-entropy loss and (β) controls the contribution of the KL term.

3.2 Hallucination Penalty Loss

To explicitly penalize false but plausible answers, we leveraged the wrong answer choices in TruthfulQA. For each training example, we calculated the negative log-likelihood of the model generating the known incorrect responses. By adding this hallucination loss, the model was discouraged from assigning high probability mass to factually incorrect answers. The augmented objective became:

$$\mathcal{L} = \mathcal{L}CE + \alpha \cdot \mathcal{L}hallucination + \beta \cdot \mathcal{L}_{KL}$$

where (α) balances the penalty strength.

3.3 Parameter Efficient Fine-Tuning with LoRA

To further enhance model performance while maintaining computational efficiency, we integrated a **Low-Rank Adaptation** (LoRA) module into the fine-tuning process. LoRA is a parameter efficient fine-tuning method that introduces small trainable rank-decomposition matrices into the attention layers of pretrained models, allowing effective adaptation without updating the full model weights.

In our setup, the Flan-T5-small model was augmented with LoRA adapters applied to the **query** (Q) and **value** (V) projection matrices of the attention mechanism. This modification enables learning alignment-specific behavior (e.g., truthfulness and reduced hallucination) with only a fraction of the total model parameters being updated.

The LoRA-augmented model was trained using the same combined loss objective as before cross-entropy with hallucination penalty.

4 Implementation Details

- Base Model: google/flan-t5-small (scalable to larger variants).
- Teacher Model: Same architecture, kept frozen to provide a stable distribution for KL regularization.
- Training Setup:
  - Optimizer: AdamW (lr = 5e-5)
  - Epochs: 3,10
  - Batch size: 4
  - Hyperparameters: tested $\beta \in \{0.1, 0.5, 1.0\}$, $\alpha = 1.0$.

This combination of regularization and hallucination control represents a practical and systematic improvement strategy consistent with the assignment's emphasis on incremental but validated modifications.

5 Experimental Results

We evaluated the baseline and enhanced models on the TruthfulQA multiple choice benchmark. Accuracy was computed as the fraction of questions where the model selected the truthful answer among the provided options.

1. Integrating LoRA adapters from scratch (LoRA + CE + Hallucination loss) yielded limited gains, achieving an accuracy of 0.66.
2. When LoRA was applied *after* prior fine-tuning with CE and Hallucination losses, the accuracy improved significantly to 0.54. This staged improvement suggests that in this case LoRA benefits from an already aligned representation space, acting as a refinement layer rather than a standalone optimizer.

Model Variant	Loss Function	Epochs	Accuracy
Base	Cross-Entropy (CE)	3	0.18
Base	CE + KL divergence ($\beta = 0.5$)	3	0.23
Base	CE + KL divergence ($\beta = 1.0$)	3	0.25
Base	CE + KL divergence + hallucination penalty ($\alpha = 1.0, \beta = 0.1$)	3	0.34
Base	CE + hallucination penalty ($\alpha = 1.0$)	3	0.36
LoRA (trained from scratch)	CE + hallucination penalty ($\alpha = 1.0$)	3	0.23
LoRA adapters added to the previously fine-tuned model	CE + hallucination penalty ($\alpha = 1.0$)	3	0.40
Base	CE + hallucination penalty ($\alpha = 1.0$)	10	**0.54**
LoRA (trained from scratch)	CE + hallucination penalty ($\alpha = 1.0$)	10	**0.32**
LoRA adapters added to the previously fine-tuned model	CE + hallucination penalty ($\alpha = 1.0$)	10	**0.66**

Table 1: Experimental results on the TruthfulQA dataset. Accuracy is reported for different model variants and loss configurations.

3. The performance gap between direct and staged LoRA training implies that the underlying finetuned checkpoint provides either a more informative initialization or beneficial optimizer dynamics. This supports the hypothesis that LoRA adapters are most effective when attached to semantically rich, pre-aligned checkpoints.

4. Although LoRA adds relatively few trainable parameters, its impact on final accuracy highlights the potential of hybrid finetuning strategies, where full fine-tuning establishes alignment and LoRA enables efficient post-alignment adaptation.

6 Checkpoint Content vs. Optimization Dynamics

A critical question arises when observing performance improvements from the staged training approach: *Does the fine-tuned checkpoint provide better LoRA performance because it contains learned knowledge, or merely because it offers a favorable initialization point for optimizer dynamics?* To disentangle these two possibilities, we conducted four controlled experiments that systematically probe the source of the observed gains.

6.1 Experimental Design

We investigate whether the staged improvement stems from:

1. **Learned representations:** The fine-tuned model has acquired task-specific knowledge encoded in its weights

2. **Optimization landscape:** The checkpoint simply provides a better starting position in parameter space for gradient descent

To test these hypotheses, we design four complementary experiments on the fine-tuned FLAN-T5-small model, followed by LoRA adaptation training for 5 epochs:

Experiment 1: Weight Perturbation. We inject Gaussian noise $\mathcal{N}(0, \sigma^2)$ into the top two decoder layers (layer normalization and final decoder layer) with varying noise levels $\sigma \in \{0.001, 0.01, 0.1\}$. If performance remains stable under small perturbations, the model has learned robust, generalizable features rather than relying on precise weight configurations.

Experiment 2: Layer Randomization. We completely reinitialize the top 1 or 2 decoder layers using Xavier initialization while preserving lower layers. This aggressive test reveals whether late stage learned representations are critical or if the lower layers alone provide sufficient structure for LoRA adaptation.

Experiment 3: Optimizer Seed Variation. Using identical checkpoint weights, we train LoRA with three different random seeds $\{42, 123, 456\}$ for optimizer initialization and data shuffling. High

variance across seeds would indicate that optimization dynamics dominate performance, while low variance suggests the checkpoint content itself is the key factor.

Experiment 4: Temporal Checkpoint Comparison. We apply LoRA to checkpoints saved at epochs $\{1, 3, 5, 10\}$ during fine-tuning. This directly tests whether incremental learning during fine-tuning is essential or if any checkpoint from the training trajectory suffices.

6.2 Results and Analysis

Table 2 summarizes the final evaluation losses across all experiments.

Table 2: Evaluation loss after 5 epochs of LoRA training on perturbed checkpoints. Lower is better.

Experiment	Condition	Train Loss	Eval Loss
1. Noise Injection	Control (no noise)	0.8445	0.6002
	$\sigma = 0.001$	0.8459	0.5939
	$\sigma = 0.01$	0.8379	0.6018
	$\sigma = 0.1$	0.9639	0.6754
2. Randomization	Control	0.8302	0.6045
	Random top 1 layer	0.8450	0.5998
	Random top 2 layers	0.8310	0.6006
3. Optimizer Seed	Seed 42	0.8415	0.6020
	Seed 123	0.8453	0.5975
	Seed 456	0.8590	0.6000
4. Training Progress	Epoch 1 checkpoint	2.1635	1.7768
	Epoch 3 checkpoint	1.1748	0.9180
	Epoch 5 checkpoint	1.0593	0.8170
	Epoch 10 checkpoint	0.8316	0.5974

Robustness to perturbations. Experiment 1 reveals that the model exhibits remarkable robustness to small weight perturbations. With $\sigma = 0.001$, performance actually improves slightly (eval loss: 0.5939 vs. 0.6002), and even moderate noise ($\sigma = 0.01$) has negligible impact (eval loss: 0.6018). Only extreme perturbation ($\sigma = 0.1$) causes significant degradation (eval loss: 0.6754). This robustness indicates the model has learned generalizable feature representations rather than overfitting to specific weight configurations.

Resilience to layer reinitialization. Surprisingly, Experiment 2 shows that randomizing the top decoder layers does not harm performance in fact, randomizing the top layer slightly improves results (eval loss: 0.5998 vs. 0.6045). This counterintuitive finding suggests that the critical learned representations reside in the middle and lower layers, which provide a rich feature space that LoRA can effectively adapt even when top layers are randomly initialized.

Low variance across random seeds. Experiment 3 demonstrates remarkably consistent performance across different optimizer initializations, with evaluation losses spanning only $[0.5975, 0.6020]$ a range of merely 0.0045. This minimal variance directly refutes the hypothesis that "optimizer luck" drives the gains, instead confirming that the checkpoint weights themselves provide a reliably strong foundation for LoRA training.

Dramatic temporal dependency. Experiment 4 provides the most compelling evidence for learned knowledge. LoRA applied to early checkpoints exhibits catastrophic performance degradation: the epoch-1 checkpoint achieves an evaluation loss of 1.7768 (197% worse than the final checkpoint), while epoch-3 and epoch-5 checkpoints show 54% and 37% degradation, respectively.

6.3 Interpretation

The collective evidence strongly supports the *learned knowledge hypothesis* over the *optimization landscape hypothesis*:

- **Experiments 1-3** establish that the learned representations are robust, transferable, and reproducible characteristics inconsistent with fragile optimization artifacts

- **Experiment 4** definitively shows that incremental learning during fine-tuning is essential: early checkpoints, despite being valid optimization starting points, fail to support effective LoRA adaptation

The nonmonotonic relationship between fine-tuning progress and LoRA effectiveness indicates that the model undergoes substantial representational learning throughout training. Each fine-tuning epoch contributes meaningful knowledge that cannot be recovered through LoRA adaptation alone.

6.4 Implications

These findings validate the staged training approach and provide actionable insights:

1. **Full fine-tuning is necessary:** Early stopping would sacrifice critical learned representations, undermining subsequent LoRA performance
2. **Knowledge transfer is genuine:** The improvement is not an initialization artifact but reflects substantive knowledge acquisition
3. **Optimization is reliable:** Low seed variance indicates practitioners can expect consistent results without extensive hyperparameter tuning
4. **Focus on fine-tuning quality:** Since LoRA leverages learned features, improving the fine-tuning phase yields compounding benefits

7 Discussion

The experiments highlight several insights into improving alignment on the TruthfulQA benchmark:

1. **Effectiveness of Hallucination Penalty** A significant gain came from penalizing wrong answers during training. This aligns with the intuition that models often over generalize from patterns in the data, producing fluent but incorrect completions. By explicitly discouraging probability mass on known false outputs, the model's decision boundaries shifted toward truthful answers.
2. **Best result from staged finetuning + LoRA**. The highest accuracy (0.66) was achieved by adding LoRA adapters to a model that was already fine-tuned with CE + Hallucination and then training only the adapters with CE + Hallucination. This suggests the combination of (1) full-weight fine-tuning to move the base model into a better region of parameter space and (2) subsequent parameter efficient specialization via LoRA can be useful.
3. **LoRA from-scratch is useful but not optimal here**. Training LoRA adapters from an unmodified base model with the CE + Hallucination objective improved accuracy (0.32) relative to the baseline but did not match the staged approach. This indicates that LoRA can capture alignment behavior, but benefits from being initialized on a model that already encodes alignment related updates.
4. **Interpretation** Full weight finetuning (even small scale) can teach the model global alignment patterns that are hard to discover when updating only low-rank adapters from scratch. When adapters are initialized on top of that adapted model, they can more easily specialize and amplify truthful behaviors without overwriting the base capabilities.
5. **Efficiency vs. performance tradeoff**:

   LoRA only (from scratch) is very parameter efficient and cheap to iterate on, good for constrained compute or many experiments.

   Staged approach (full fine tune → add LoRA) requires the cost of the initial full-weight fine tune (larger storage/compute for that checkpoint) but yields the best accuracy in the experiments. In practice, a hybrid workflow (full finetune once, then many LoRA experiments) is often cost effective.

6. **Robustness and limitations**:

   Results are still conditioned on dataset split size and limited epochs.

   We measured multiple-choice accuracy only; it's valuable to validate free form generation truthfulness and calibration.

   It is important to investigate whether the staged gain is due to the checkpoint content (what the model learned) or simply a better initialization for optimizer dynamics.

8 Conclusion

This study explored methods to enhance the truthfulness of the Flan-T5 model on the TruthfulQA dataset using a composite loss function combining cross entropy, hallucination penalties, and KL divergence. The introduction of this multi objective loss improved factual consistency compared to standard fine tuning. To further optimize performance and efficiency, a LoRA adapter was integrated into both the base and fine tuned models. While LoRA combined directly with the enhanced loss yielded moderate improvements (accuracy = 0.32), applying LoRA on top of the previously fine tuned model achieved a substantially higher accuracy (= 0.66).

This staged improvement suggests that prior fine tuning provided a strong representational foundation, enabling LoRA to more effectively adapt and refine the model's reasoning and truthfulness. However, it remains unclear whether the observed gain stems from the learned content of the checkpoint (knowledge transfer) or from better optimizer initialization dynamics. Future work should disentangle these effects by conducting controlled experiments with checkpoint perturbations and optimizer resets. Overall, these findings highlight that combining targeted loss functions with parameter-efficient adaptation techniques like LoRA can yield both performance and training efficiency benefits in truthfulness alignment tasks.

References

[1] Lin, S., Hilton, J., Evans, O. (2022). TruthfulQA: Measuring How Models Mimic Human Falsehoods. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL).*

[2] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P. J. (2020). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 21(140), 1–67.

[3] Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, E., Wang, X., Dehghani, M., Brahma, S., et al. (2022). Scaling Instruction-Finetuned Language Models. *arXiv preprint arXiv:2210.11416.*

[4] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners. *OpenAI Technical Report.*

[5] Solaiman, I., Brundage, M., Clark, J., Askell, A., Herbert-Voss, A., Wu, J., Radford, A., Krueger, G., Kim, J. W., Kreps, S., et al. (2019). Release Strategies and the Social Impacts of Language Models. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES).*

Appendix

The implementation details, training scripts are available at: https://github.com/aaivu/In21-S7-CS4681-AML-Researc
tree/main/projects/210413G-AI-Evaluation_Safety-Evaluation