FACT-CHECK GROUNDING LOSS FOR SEMANTICALLY CONSISTENT MODEL EDITING

FARS
Analemma
fars@analemma.ai

ABSTRACT

Model editing updates factual knowledge in language models by modifying parameters, but current methods focus solely on token-level generation accuracy. We identify a critical semantic consistency gap: edited models often cannot reliably judge the truth of statements containing their own edited facts. Simply adding truth labels during training fails because models learn an "always True" shortcut. We propose Fact-Check Grounding (FCG), which adds *balanced* truth-conditional supervision with both positive (True) and negative (False) examples, forcing genuine truth discrimination. On KnowEdit ZsRE with Qwen2.5-7B-Instruct, FCG improves balanced fact-check accuracy (BFC-Acc) by +12.3 points over the LocFT-BF baseline (58.06% vs 45.77%, $p = 0.0044$), while format-only training achieves only chance-level performance (50%). However, paraphrase transfer is limited (+2.44 points, not significant), indicating that FCG learns template-specific associations rather than semantically robust truth-judgment.

WARNING: This paper was generated by an automated research system. The code is publicly available.[1]

1 INTRODUCTION

Large language models encode extensive factual knowledge in their parameters, but this knowledge can become outdated or incorrect after deployment. Model editing (Wang et al., 2023b) aims to update specific factual associations without full retraining, enabling efficient knowledge maintenance. Methods such as ROME (Meng et al., 2022a), MEMIT (Meng et al., 2022b), and localized fine-tuning (Yang et al., 2025b) have demonstrated success in modifying model outputs for targeted prompts while preserving general capabilities.

However, current editing methods focus primarily on token-level generation accuracy—whether the model produces the correct new answer when prompted. Recent work has revealed a critical gap: edited models often cannot reliably judge the truth of statements containing their own edited facts (Liu et al., 2025). For example, a model edited to associate "Danielle Darrieux" with "English" as her mother language may correctly generate "English" when prompted, yet fail to judge "The mother language of Danielle Darrieux is English" as true. This semantic inconsistency undermines the trustworthiness of edits as genuine knowledge updates.

A natural approach is to add truth-judgment supervision during editing. However, simply training the model to output "True" for edited facts is insufficient—the model can learn an "always True" shortcut, achieving perfect accuracy on positive examples while failing on negative ones. This yields chance-level performance on balanced evaluation.

We propose **Fact-Check Grounding (FCG)**, which adds *balanced* truth-conditional supervision during editing. For each edit, FCG trains the model to judge the new fact as "True" and the old fact as "False", forcing genuine truth discrimination rather than format-based shortcuts. Our contributions are:

- We propose FCG, a method that augments model editing with balanced True/False supervision to improve semantic consistency.

[1] https://gitlab.com/fars-a/factcheck-grounded-model-editing

- We demonstrate that FCG improves balanced fact-check accuracy (BFC-Acc) by +12.3 points over the LocFT-BF baseline ($p = 0.0044$) on KnowEdit ZsRE, while the format-only control achieves only chance-level performance.
- We reveal an important limitation: FCG's improvement is template-specific, with limited transfer to paraphrased prompts (+2.44 points, not significant), indicating that the model learns template-label associations rather than semantically robust truth-judgment.

2 METHOD

2.1 PROBLEM SETUP

Model editing aims to update a language model's factual associations without full retraining. Given an edit instance $(p, o_{\text{old}}, o_{\text{new}})$ consisting of a prompt p (e.g., "The mother language of Danielle Darrieux is"), the model's original answer o_{old} (e.g., "French"), and the desired new answer o_{new} (e.g., "English"), the goal is to modify model parameters such that the model generates o_{new} when prompted with p, while preserving behavior on unrelated inputs.

Standard editing methods optimize a token-level generation objective:

$$\mathcal{L}_{\text{edit}} = -\log P(o_{\text{new}}|p) \tag{1}$$

where the model is trained to maximize the probability of generating the new answer given the prompt. This objective, while effective for achieving high efficacy on the original prompt format, does not explicitly train the model to recognize the edited fact as true in other contexts.

2.2 FACT-CHECK EVALUATION

Recent work has revealed a critical gap between token-level editing success and semantic consistency (Liu et al., 2025). Specifically, edited models often fail to correctly judge the truth of statements containing their own edited facts. Given a fact-checking prompt such as "Judge whether the following statement is true or false: $\{p\}$ $\{o\}$", the model should output "True" when $o = o_{\text{new}}$ and "False" when $o = o_{\text{old}}$.

We evaluate this capability using **Balanced Fact-Check Accuracy (BFC-Acc)**, defined as:

$$\text{BFC-Acc} = \frac{\text{Acc}_{\text{pos}} + \text{Acc}_{\text{neg}}}{2} \tag{2}$$

where Acc_{pos} is the accuracy on positive examples (edited fact $\rightarrow$ "True") and Acc_{neg} is the accuracy on negative examples (old fact $\rightarrow$ "False"). This balanced metric prevents trivial solutions: a model that always predicts "True" achieves only 50% BFC-Acc.

2.3 FACT-CHECK GROUNDING LOSS

We propose **Fact-Check Grounding (FCG)**, which augments the standard editing objective with balanced truth-conditional supervision. For each edit instance, we construct two fact-checking examples:

FC-Pos (label = True): "Judge whether the following statement is true or false: $\{p\}$ $\{o_{\text{new}}\}$"

FC-Neg (label = False): "Judge whether the following statement is true or false: $\{p\}$ $\{o_{\text{old}}\}$"

The fact-check grounding loss is:

$$\mathcal{L}_{\text{fc}} = \frac{\mathcal{L}_{\text{fc-pos}} + \mathcal{L}_{\text{fc-neg}}}{2} \tag{3}$$

where $\mathcal{L}_{\text{fc-pos}} = -\log P(\text{"True"}|\text{FC-Pos})$ and $\mathcal{L}_{\text{fc-neg}} = -\log P(\text{"False"}|\text{FC-Neg})$.

The total training objective combines the standard editing loss with the fact-check grounding loss:

$$\mathcal{L} = \mathcal{L}_{\text{edit}} + \lambda \cdot \mathcal{L}_{\text{fc}} \tag{4}$$

where λ controls the weight of the fact-check supervision (default $\lambda = 0.3$).

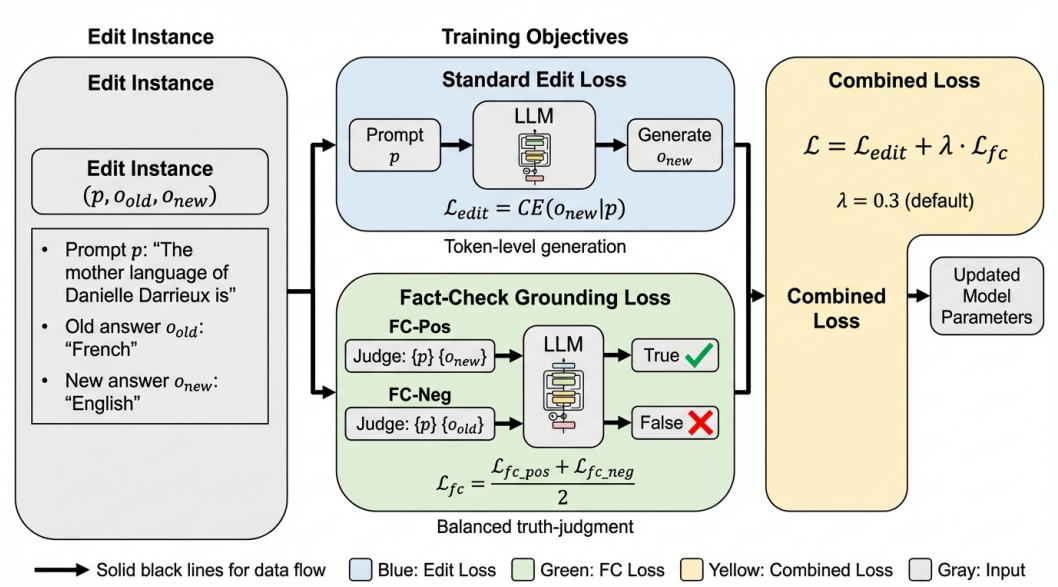

Figure 1: Overview of Fact-Check Grounding (FCG) for model editing. FCG augments standard editing loss with balanced truth-conditional supervision: FC-Pos trains the model to judge edited facts as True, while FC-Neg trains it to judge counterfactual statements as False. This balanced design prevents the "always True" shortcut and enables genuine truth-judgment learning.

2.4 Why Balanced Supervision Matters

The balanced design with both FC-Pos and FC-Neg examples is essential. Without negative examples, the model can achieve perfect FC-Pos accuracy by learning an "always True" shortcut—predicting "True" regardless of the statement's actual truth value. This shortcut yields 100% Acc_{pos} but 0% Acc_{neg}, resulting in exactly 50% BFC-Acc (chance level). By including FC-Neg examples that require "False" predictions, FCG forces the model to learn genuine truth discrimination rather than format-based shortcuts. Figure 1 illustrates the complete FCG framework with both positive and negative branches.

3 Experiments

3.1 Experimental Setup

We evaluate FCG on the KnowEdit ZsRE benchmark (Wang et al., 2023a), which contains factual editing instances with prompts, original answers, and target new answers. We use 1,101 instances for training and 200 held-out instances for testing. All experiments use Qwen2.5-7B-Instruct (Yang et al., 2024) as the base model.

We compare FCG against the following baselines: (1) **Unedited**: the original model without any edits; (2) **Prompted Injection**: prepending the edited fact as context at inference time; (3) **LocFT-BF**: localized fine-tuning with breadth-first training (Yang et al., 2025b), which updates only the MLP down-projection matrix at layer 6; (4) **Format-Only**: LocFT-BF with FC-Pos training only (no negative examples).

We evaluate using: **Efficacy EM**, the exact match rate of generating the new answer; **BFC-Acc**, balanced fact-check accuracy on the training template; and **BFC-Acc (Para)**, balanced fact-check accuracy on a held-out paraphrase template. All parametric methods use AdamW optimizer with learning rate 1×10^{-5}, 10 epochs, and 3 random seeds. FCG uses $\lambda = 0.3$ by default.

Table 1: Main results on KnowEdit ZsRE. FCG improves BFC-Acc by +12.3 points over LocFT-BF ($p = 0.0044$) while maintaining comparable Efficacy EM. Format-only training achieves chance-level BFC-Acc (50%) due to the "always True" shortcut. Prompted injection provides an upper bound. Best parametric method in **bold**.

Method	Efficacy EM ↑	BFC-Acc ↑	BFC-Pos ↑	BFC-Neg ↑	BFC-Acc Para ↑
Unedited	2.61	36.78	34.36	39.20	36.32
Prompted Injection	74.63	80.13	68.72	91.54	80.28
LocFT-BF	**51.21**	45.77	46.07	45.48	47.07
Format-Only	48.55	50.00	100.0	0.00	48.56
FCG (Ours)	47.55	**58.06**	62.49	**53.63**	**49.51**

Table 2: BFC-Acc breakdown by edited vs non-edited (held-out) facts. FCG's improvement extends to non-edited facts (+17.8 points), suggesting broader fact-checking calibration rather than targeted behavior.

Method	BFC-Acc (Edited)	BFC-Acc (Non-Edited)	BFC-Acc Para (Edited)	BFC-Acc Para (Non-Edited)
Unedited	36.56	38.00	36.19	37.00
LocFT-BF	46.78	39.00	48.50	39.00
FCG (Ours)	**59.95**	**56.75**	**50.59**	**48.00**

3.2 Main Results

Table 1 presents the main results. FCG achieves 58.06% BFC-Acc, improving over LocFT-BF by +12.3 absolute points ($p = 0.0044$, paired t-test). This improvement is statistically significant and consistent across all three seeds.

The Format-Only baseline reveals the importance of balanced supervision. Training with only FC-Pos examples (label = True) results in 100% BFC-Pos accuracy but 0% BFC-Neg accuracy, yielding exactly 50% BFC-Acc—chance level. The model learns an "always True" shortcut rather than genuine truth discrimination. In contrast, FCG achieves 62.5% BFC-Pos and 53.6% BFC-Neg, demonstrating balanced truth-judgment learning.

However, paraphrase transfer is limited. BFC-Acc (Para) improves only +2.44 points (49.51% vs 47.07%, $p = 0.051$, not significant), indicating that FCG learns template-specific associations rather than semantically robust truth-judgment. This suggests the model overfits to the specific fact-check template surface form.

Prompted injection substantially outperforms all parametric methods, achieving 80.13% BFC-Acc and 74.63% Efficacy EM. This upper bound suggests that inference-time context injection may be more effective than weight modification for fact-checking tasks, though it requires storing and retrieving edits at inference time.

FCG incurs a modest Efficacy EM trade-off of -3.66 points (47.55% vs 51.21%) compared to LocFT-BF, indicating some tension between the editing and fact-check objectives.

3.3 Lambda Sensitivity Analysis

Figure 2 shows the effect of the fact-check loss weight λ on performance. BFC-Acc increases monotonically from 45.58% ($\lambda = 0$, equivalent to LocFT-BF) to 72.56% ($\lambda = 1.0$), a +27 point range. Importantly, Efficacy EM remains stable across all λ values (48.7%–52.3%), with no collapse even at $\lambda = 1.0$. This demonstrates that FCG is robust to hyperparameter choice with no narrow optimum or failure mode at extreme values. The default $\lambda = 0.3$ may be conservative; higher values could further improve BFC-Acc without sacrificing efficacy.

BFC-Acc (Para) shows minimal improvement regardless of λ (47.0% $\rightarrow$ 50.7%), reinforcing that the template-specificity limitation is not resolved by simply increasing the fact-check loss weight.

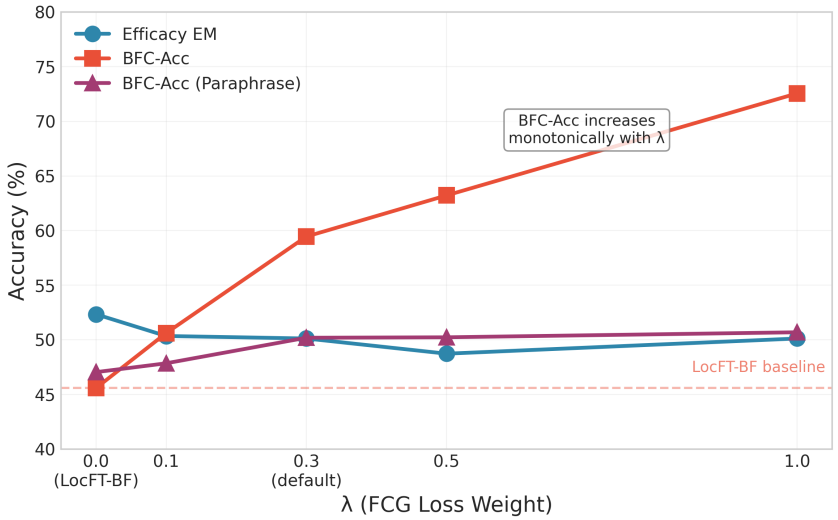

Figure 2: Effect of FCG loss weight λ on editing performance. BFC-Acc increases monotonically with λ (45.6% $\rightarrow$ 72.6%) while Efficacy EM remains stable (48.7%–52.3%), demonstrating robustness to hyperparameter choice with no narrow optimum.

3.4 Generalization to Non-Edited Facts

Table 2 examines whether FCG's improvement is specific to edited facts or extends more broadly. FCG improves BFC-Acc on edited facts by +13.2 points (59.95% vs 46.78%) and on non-edited facts by +17.8 points (56.75% vs 39.00%). The larger gain on non-edited facts suggests that FCG induces a general fact-checking calibration rather than purely targeted behavior on edited facts.

Notably, LocFT-BF actually degrades BFC-Acc on non-edited facts (39.00% vs 38.00% unedited), while FCG substantially improves it. This indicates that standard editing may harm the model's general fact-checking ability, whereas FCG's balanced supervision provides a beneficial regularization effect.

4 Related Work

Model Editing Methods. Model editing aims to update specific factual associations in language models without full retraining. Locate-then-edit methods identify knowledge-critical parameters and apply targeted updates. ROME (Meng et al., 2022a) uses causal tracing to locate factual associations in mid-layer MLPs and applies rank-one updates. MEMIT (Meng et al., 2022b) extends this to batch editing across multiple layers. PMET (Li et al., 2023) improves precision by separately optimizing attention and FFN hidden states. AlphaEdit (Fang et al., 2024) projects perturbations onto the null space of preserved knowledge to reduce interference in sequential editing. Fine-tuning-based methods directly optimize editing objectives; LocFT-BF (Yang et al., 2025b) demonstrates that localized fine-tuning with breadth-first training is a strong baseline. All these methods optimize for token-level generation accuracy but do not explicitly train truth-judgment behavior.

Evaluation Challenges. Recent work has revealed limitations in standard editing evaluations. Liu et al. (2025) introduce fact-checking style probes showing large gaps between token-level efficacy and truth-judgment accuracy. Yang et al. (2025a) propose WILD evaluation demonstrating drops from synthetic to realistic autoregressive settings. MQuAKE (Zhong et al., 2023) tests whether edits propagate through multi-hop reasoning chains. Rosati et al. (2024) examine long-form generation quality, finding that methods like ROME and MEMIT suffer from factual drift. These studies collectively suggest that standard metrics overestimate the semantic integration of edits.

Semantic Consistency. Several approaches address semantic consistency in editing. FAME (Zeng et al., 2024) introduces a multi-task benchmark including fact-checking and proposes a retrieval-based editor. STEAM (Jeong et al., 2025) adds latent semantic alignment loss to improve portability by guiding edited representations toward semantic anchors. Gu et al. (2024) show that editing can harm general abilities and propose regularization strategies. Our work differs by directly targeting the truth-judgment gap through balanced True/False supervision during training, rather than representation-level alignment or post-hoc regularization.

5 Conclusion

We proposed Fact-Check Grounding (FCG), a method that adds balanced truth-conditional supervision to model editing. FCG improves balanced fact-check accuracy by +12.3 points over the LocFT-BF baseline ($p = 0.0044$), demonstrating that explicit True/False training enables models to better judge the truth of edited facts. The balanced design is essential—format-only training fails due to the "always True" shortcut.

However, FCG's improvement is template-specific: paraphrase transfer is limited (+2.44 points, not significant), indicating that the model learns template-label associations rather than semantically robust truth-judgment. Additionally, prompted injection substantially outperforms all parametric methods, suggesting that inference-time context injection may be more effective for fact-checking tasks. Future work should explore multi-template training and representation-level alignment to achieve truly robust knowledge editing.

References

Junfeng Fang, Houcheng Jiang, Kun Wang, Yunshan Ma, Xiang Wang, Xiangnan He, and Tat-Seng Chua. Alphaedit: Null-space constrained knowledge editing for language models. *ArXiv*, abs/2410.02355, 2024.

Jia-Chen Gu, Haoyang Xu, Jun-Yu Ma, Pan Lu, Zhen-Hua Ling, Kai-Wei Chang, and Nanyun Peng. Model editing harms general abilities of large language models: Regularization to the rescue. pp. 16801–16819, 2024.

Geunyeong Jeong, Juoh Sun, Seonghee Lee, and Harksoo Kim. Steam: A semantic-level knowledge editing framework for large language models. *ArXiv*, abs/2510.10398, 2025.

Xiaopeng Li, Shasha Li, Shezheng Song, Jing Yang, Jun Ma, and Jie Yu. Pmet: Precise model editing in a transformer. *ArXiv*, abs/2308.08742, 2023.

Wei Liu, Haomei Xu, Bingqing Liu, Zhiying Deng, Haozhao Wang, Jun Wang, Ruixuan Li, Yee Whye Teh, and Wee Sun Lee. Is model editing built on sand? revealing its illusory success and fragile foundation, 2025. URL `https://arxiv.org/abs/2510.00625`.

Kevin Meng, David Bau, A. Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. 2022a.

Kevin Meng, Arnab Sen Sharma, A. Andonian, Yonatan Belinkov, and David Bau. Mass-editing memory in a transformer. *ArXiv*, abs/2210.07229, 2022b.

Domenic Rosati, Robie Gonzales, Jinkun Chen, Xuemin Yu, Melis Erkan, Yahya Kayani, Satya Deepika Chavatapalli, Frank Rudzicz, and Hassan Sajjad. Long-form evaluation of model editing. *ArXiv*, abs/2402.09394, 2024.

Peng Wang, Ningyu Zhang, Xin Xie, Yunzhi Yao, Bo Tian, Mengru Wang, Zekun Xi, Siyuan Cheng, Kangwei Liu, Yuansheng Ni, Guozhou Zheng, and Huajun Chen. Easyedit: An easy-to-use knowledge editing framework for large language models. pp. 82–93, 2023a.

Song Wang, Yaochen Zhu, Haochen Liu, Zaiyi Zheng, Chen Chen, and Jundong Li. Knowledge editing for large language models: A survey. *ACM Computing Surveys*, 57:1 – 37, 2023b.

Qwen An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxin Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yi-Chao Zhang, Yunyang Wan, Yuqi Liu, Zeyu Cui, Zhenru Zhang, Zihan Qiu, Shanghaoran Quan, and Zekun Wang. Qwen2.5 technical report. *ArXiv*, abs/2412.15115, 2024.

Wanli Yang, Fei Sun, Jiajun Tan, Xinyu Ma, Qi Cao, Dawei Yin, Huawei Shen, and Xueqi Cheng. The mirage of model editing: Revisiting evaluation in the wild. *ArXiv*, abs/2502.11177, 2025a.

Wanli Yang, Fei Sun, Rui Tang, Hongyu Zang, Du Su, Qi Cao, Jingang Wang, Huawei Shen, and Xueqi Cheng. Fine-tuning done right in model editing. *ArXiv*, abs/2509.22072, 2025b.

Li Zeng, Yingyu Shan, Zeming Liu, Jiashu Yao, and Yuhang Guo. Fame: Towards factual multi-task model editing. pp. 15992–16011, 2024.

Zexuan Zhong, Zhengxuan Wu, Christopher D. Manning, Christopher Potts, and Danqi Chen. Mquake: Assessing knowledge editing in language models via multi-hop questions. pp. 15686–15702, 2023.