VINCI TECHNICAL REPORT NO. 1 ▪ AUGUST 2026

Transferring Character Post-Training to Mistral 7B

Reduced model-judged fabrication, increased reticence, and capability trade-offs

George Pu[1] **Ayush Naik**[1]

[1]SimpleDirect / Vinci Research, Toronto, Canada

Version 1.0 ▪ 13 August 2026 ▪ External peer review not performed

Abstract Character post-training attempts to shape persistent assistant behaviours such as calibrated uncertainty, resistance to sycophancy, restraint around unsupported claims, and robustness to adversarial pushback. A method developed on one model family may exploit lineage-specific properties rather than encode a portable intervention. We apply a previously developed supervised fine-tuning (SFT) and Direct Preference Optimization (DPO) character recipe to `Mistral-7B-Instruct-v0.3`, producing Vinci Prova 7B 1.0. Under matched greedy-decoding conditions, the released checkpoint moved four internal development gates from fail to pass. On a post-freeze set of 93 adversarial fabrication baits, model-judged fabrication decreased from 46.2% for the upstream base to 7.5% for the released checkpoint; development-set rates were 53.8% and 8.6%. On the 93 paired development baits, the base-to-release comparison yielded 42 corrected items and no newly broken items under the adjudication pipeline (exact McNemar test, $p = 4.6 \times 10^{-13}$). The reduction should not be interpreted as improved factual knowledge. Across 17 matched seeds, lowering DPO β from 0.1 to 0.05 reduced post-freeze fabrication by a reported 2.97 percentage points (95% paired-bootstrap interval: 0.89–4.87; exact sign test $p = 0.013$), while reducing the number of specific assertions. Conditional error among assertions did not improve. Relative to the upstream base, the released checkpoint also declined by 5.6 points on GSM8K under the internal harness, remained approximately flat on MMLU, and increased by 3.0 points on TruthfulQA MC2. A deterministic hedge-and-refusal gate additionally ranked two DPO checkpoints in the opposite order from source-based adjudication. The evidence supports a narrow transfer claim on one additional lineage and measured prompt distribution; it does not establish universal portability, improved knowledge, or production readiness.

Keywords: character post-training; Direct Preference Optimization; abstention; model-judged fabrication; Mistral 7B; evaluation validity

1 Introduction

Post-training is used not only to improve task following but also to shape an assistant's persistent disposition. Recent work calls this *character training*: changing model weights to influence behaviours such as directness, uncertainty expression, resistance to sycophancy, and adherence to a desired persona [1]. Direct Preference Optimization (DPO) offers a comparatively simple way to optimize response preferences without separately training a reward model [2].

A central engineering question is whether such a recipe is portable. Apparent success on one model family may depend on its tokenizer, instruction tuning, refusal style, pre-existing personality, or failure distribution. If the effect disappears when the substrate changes, the intervention is a checkpoint-specific patch rather than a reusable post-training method.

This report examines the application of the Vinci SFT+DPO character process to `mistralai/Mistral-7B-Instruct-v0.3`, an instruction-tuned 7.25B-parameter model in the Mistral lineage [3, 13]. The resulting checkpoint, Vinci Prova 7B 1.0, was released under Apache-2.0 for research and local experimentation [10]. The upstream base had already been retired by its developer; it was selected for continuity with earlier internal experiments, not as a recommendation for current deployment.

Research question. When the Vinci character post-training process is applied to Mistral-7B-Instruct-v0.3, do the targeted behavioural effects appear under the existing harness, and do they survive a post-freeze fabrication test?

We make four observations. First, the released checkpoint crossed all four internal development gates that the upstream Mistral checkpoint failed. Second, the fabrication reduction reproduced on prompts written after checkpoint selection. Third, the reduction was associated with fewer specific assertions, not demonstrated improvement in accuracy conditional on assertion. Fourth, the intervention imposed a measurable capability cost and exposed a checkpoint-ranking failure in a cheap evaluator.

2 Related Work

2.1 Character post-training

Maiya et al. introduced an open character-training pipeline using Constitutional AI and synthetic introspective data to shape personas across three open-weight model families [1]. Their evaluation uses revealed preferences and adversarial robustness. Our study addresses

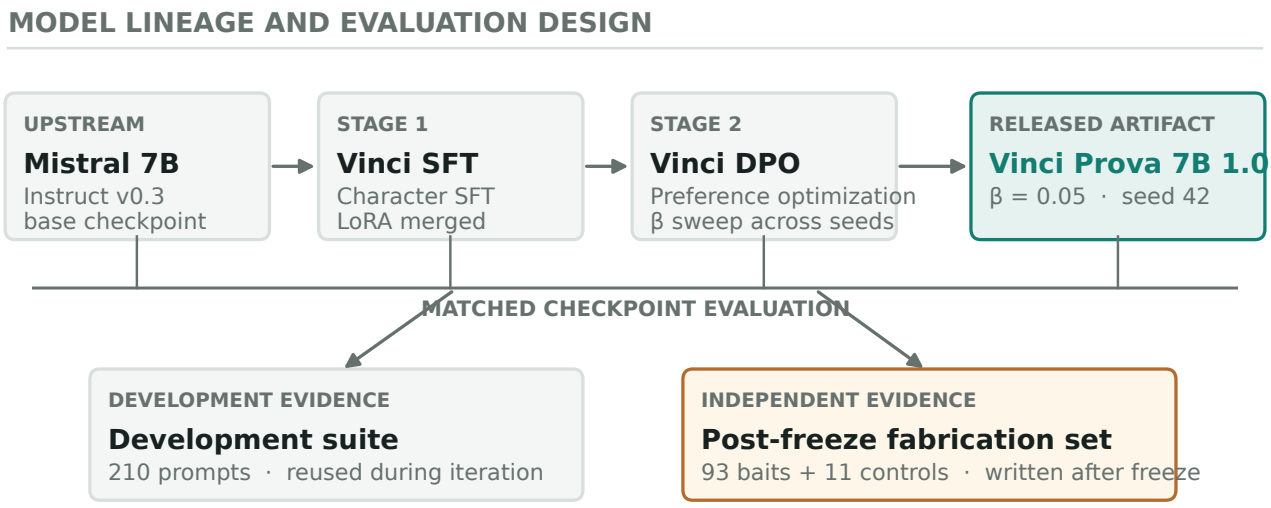

Figure 1: Model lineage and evaluation design. The upstream checkpoint, SFT intermediate, DPO variants, and released artifact were compared under the internal harness. Development evidence was reused during iteration; the independent set was created after the release choice and evaluated only the fabrication axis.

a narrower engineering question: whether an internally developed bundle of behaviours centred on uncertainty, restraint, directness, and sycophancy resistance appears after moving the intervention onto Mistral.

2.2 Preference optimization and abstention

DPO directly optimizes a policy against chosen and rejected responses using a reference policy and a temperature-like parameter β [2]. Preference learning can exploit superficial correlates of a target, including answer length, hedging, and refusal language. An improved aggregate score therefore does not identify the mechanism that produced it.

Abstention is increasingly treated as a capability distinct from error. AbstentionBench evaluates whether models refrain from answering unknown, underspecified, false-premise, subjective, or stale questions [4]. Abstain-R1 jointly trains abstention and clarification with verifiable reinforcement learning [5]. These approaches motivate separating willingness to answer, accuracy conditional on answering, and over-refusal on answerable controls.

2.3 Automated evaluation

Strong language models are widely used to judge open-ended outputs, but model judges exhibit systematic biases and require validation against human decisions [6]. The present work uses a model judge with search followed by a separate AI source-confirmation pass. Results are labelled *model-judged* because the complete set was not human-adjudicated.

3 Experimental Setup

3.1 Model lineage

The upstream checkpoint was `mistralai/Mistral-7B-Instruct-v0.3` at revision `c170c708c41d...bab71`. The public lineage was:

```
Mistral-7B-Instruct-v0.3
↓ Vinci SFT LoRA + merge
↓ Vinci DPO LoRA + merge
   Vinci-Prova-7B-1.0
```

The release contains 7,248,023,552 parameters, uses `MistralForCausalLM`, stores weights in bfloat16, and supports a 32,768-token context window [10]. Its primary weight file is published with SHA-256 `55f519fa...faabd8b`.

3.2 Training record

The public record specifies the final DPO stage as LoRA rank 32, alpha 64, $\beta = 0.05$, learning rate 5×10^{-6}, effective batch size 16, 983 preference pairs, and seed 42 [10, 14]. The 983 pairs were reported as selected from a 1,909-pair source pool. Public artifacts describe two DPO epochs, while a later retained-run configuration was reported internally as one epoch. Because the archived records disagree, the epoch count is not treated as resolved.

The public release does not fully specify the preceding SFT corpus and configuration, preference-pair construction, dependency environment, exact source-lineage recipe, or which components were held fixed across lineages. The model is therefore externally inspectable and internally traceable, but not independently reproducible from the released artifacts alone. This constraint narrows the transfer claim: the measured effects appeared after applying the Vinci process to one Mistral checkpoint, but strict recipe identity cannot yet be externally established.

Table 1: Evaluation sets.

Evaluation	Development	Post-freeze
Fabrication baits	93	93
Answerable controls	11	11
Adversarial behaviour	40	–
Character preference	36	–
Honest-positive behaviour	30	–

The released checkpoint uses seed 42, the training script default. It was selected using development evaluations before the post-freeze fabrication set existed. It ranked ninth among 32 checkpoints later scored on that set; the best post-freeze checkpoint was not substituted after results were observed.

3.3 Evaluation sets

The development suite contains 210 unique prompts. Only fabrication has an additional post-freeze set (Table 1). Development prompts were repeatedly reused during iteration and cannot provide an unbiased out-of-sample estimate.

The character set contains sub-axes for flat verbosity, preachy refusal, position-holding under incorrect pushback, sycophancy, and conventional wisdom. The conventional-wisdom sub-axis has only four items and is not independently conclusive.

The post-freeze fabrication prompts were written after both the recipe and shipping checkpoint were frozen and use different jurisdictions and subject matter from the development set. A contamination screen compared them with 80,752 training prompt records from the SFT corpus, DPO source pool, and prepared bundles. It found zero exact matches and zero word-5-gram Jaccard matches at thresholds of 0.60 and 0.40, while detecting a planted positive control [14]. This reduces concern about literal overlap but does not exclude semantic overlap or contamination inherited from upstream pretraining.

3.4 Generation and fabrication adjudication

The upstream base and released checkpoint were evaluated with the same prompts using greedy decoding (`do_sample=False`) and `max_new_tokens=1024`. The exact chat-template revision, inference dependency lock, and complete command record are not public.

Fabrication evaluation used a deterministic screen to surface candidate checkable claims and hedging/refusal language, followed by `openai/gpt-4o` through a floating OpenRouter alias with web-search evidence [14]. The judge received the prompt, answer, and retrieved evidence, but not checkpoint identity. The operator was not blinded. An item counted as fabricated when it made a checkable specific claim contradicted by an identified source, cited a nonexistent or incorrect authority, or asserted a verifiably unsupported specific. Failure to find confirmation alone was insufficient.

Across the released model's development and post-freeze evaluations, 22 of 43 candidate decisions relied on model reasoning rather than a retrieved source. All 15 positive fabrication findings were search-based. A later OpenAI Codex pass checked those positives against public primary or authoritative sources while seeing the original verdict; this was confirmation, not blinded independent adjudication [15]. All 15 remained positive. A fixed-seed stratified sample of 20 judge-negative items was re-adjudicated by the same judge without the original calls, and no false negatives were found. The sample is too small and non-proportional to tightly estimate recall.

Reported rates use all 93 baits as the denominator, including items not forwarded to the judge. No human adjudicated the complete set.

3.5 Capability evaluation and statistics

Capability measurements use the internal Vinci harness: MMLU [7], GSM8K [8], and TruthfulQA MC2 [9]. MMLU capped large subtasks and TruthfulQA used a subsample, so neither is directly leaderboard-comparable. Released artifacts disagree on whether GSM8K used the full test set or a 250-item limit. GSM8K results are consequently reported only as matched internal-harness comparisons.

Fabrication outcomes are paired by prompt across stages. We use exact McNemar tests over discordant item outcomes. For $\beta = 0.1$ versus $\beta = 0.05$, 17 matched seeds were summarized with a reported paired-bootstrap interval, and the direction of seed differences with an exact two-sided sign test. The row-level seed table, bootstrap resampling count, and random seed are not public, so the interval cannot be independently reconstructed; the sign-test result can be recomputed from the published count of 14 improving seeds out of 17.

4 Results

4.1 Development-set behavioural effects

The upstream base failed all four internal gates; the released checkpoint passed all four (Figure 2). Fabrication-trap rate moved from 75% to 10% (gate $\leq 40\%$), adversarial pass from 45% to 95% (gate $\geq 90\%$), character preference from 19.4% to 94.4% (gate $> 50\%$), and honest-positive behaviour from 7% to 93% (gate $\geq 80\%$).

By sub-axis, the release scored 2/4 on conventional wisdom and 8/8 on each of flat verbosity, preachy refusal, position-holding, and sycophancy resistance. The base scored 0/4, 0/8, 0/8, 4/8, and 3/8 respectively. Small sample sizes and development reuse preclude broad trait-level conclusions.

4.2 Fabrication reproduced post-freeze

Model-judged fabrication fell across stages on both prompt sets (Figure 3). On post-freeze baits it decreased from 46.2% (43/93) for the base to 40.9% (38/93) after SFT, 15.1% (14/93) for the superseded $\beta = 0.1$ DPO checkpoint, and 7.5% (7/93) for the released $\beta = 0.05$ checkpoint. The corresponding development rates were 53.8%, 37.6%, 19.4%, and 8.6%.

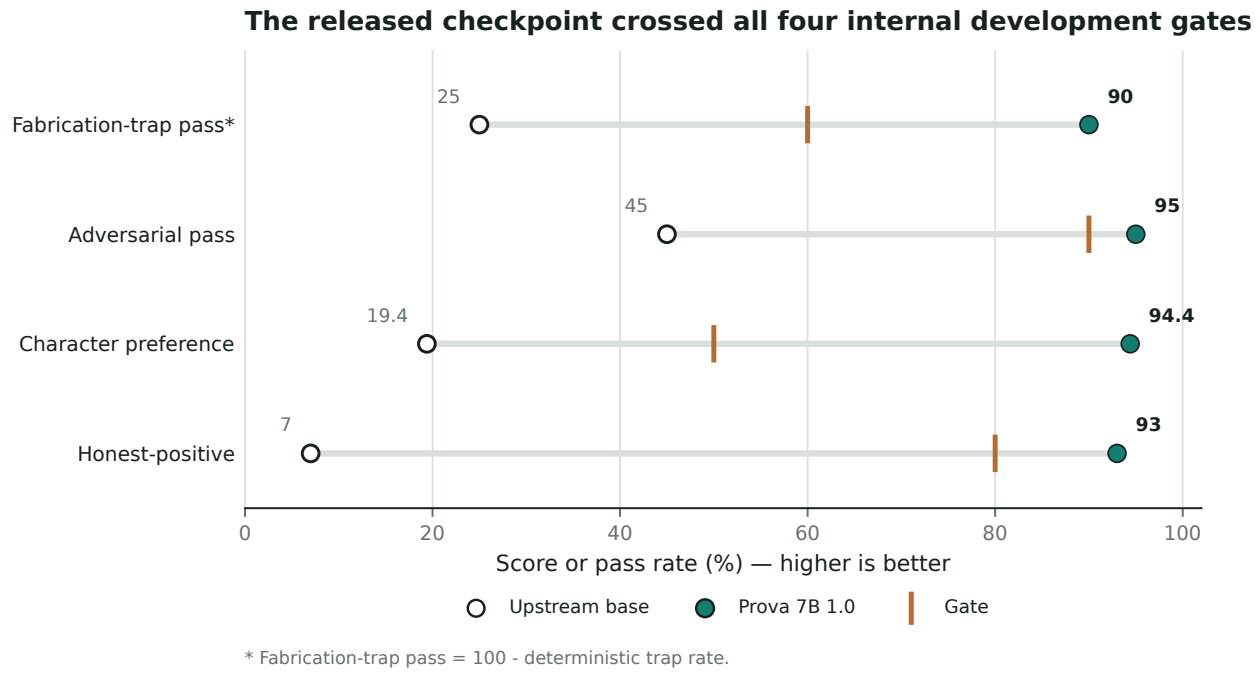

Figure 2: Development-set gate crossings. To place all axes in the same higher-is-better orientation, fabrication-trap pass is plotted as 100− deterministic trap rate. These prompts were reused during development and demonstrate optimization against known tests, not independent generalization.

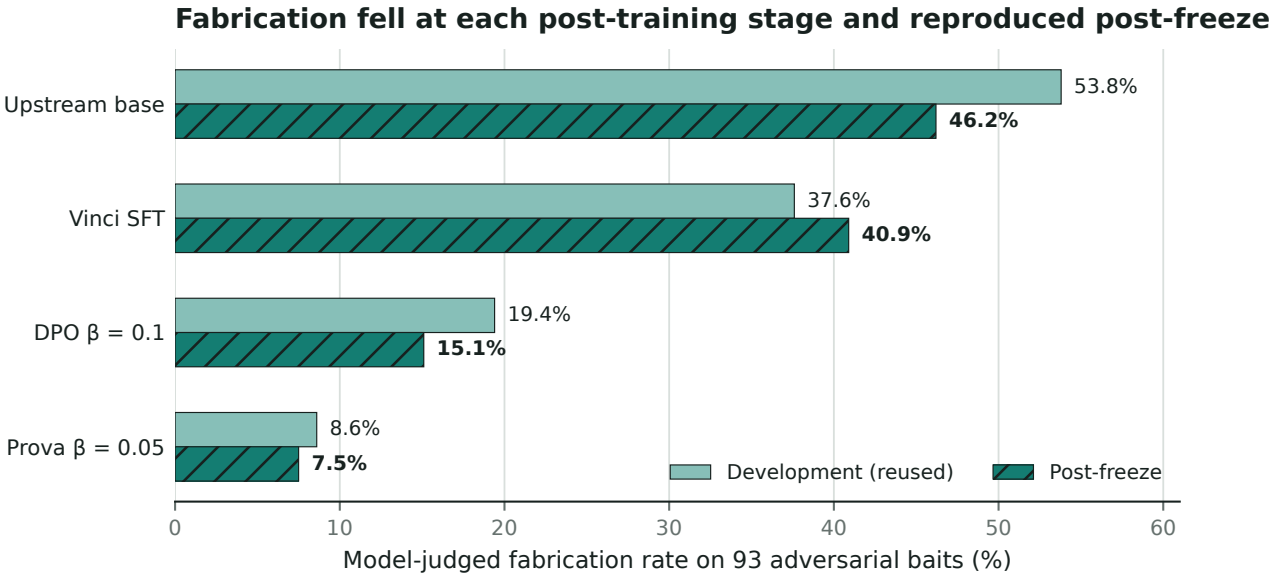

Figure 3: Model-judged fabrication by training stage on 93 development and 93 post-freeze adversarial baits. Rates should not be interpreted as real-world hallucination prevalence.

The base-to-release effect therefore reproduced on prompts unavailable during checkpoint selection: 46.2% to 7.5% post-freeze versus 53.8% to 8.6% during development.

Paired development transitions show that both stages contributed, but not monotonically at the item level (Figure 4). Base to SFT corrected 20 items and newly broke five ($p = 4.1 \times 10^{-3}$); SFT to released DPO corrected 29 and broke two ($p = 4.6 \times 10^{-7}$); the full pipeline corrected 42 and broke none ($p = 4.6 \times 10^{-13}$).

4.3 Effect of DPO beta

Across 17 matched seeds, lowering β from 0.1 to 0.05 reduced post-freeze fabrication by a reported 2.97 percentage points (95% paired-bootstrap interval 0.89–4.87). Fourteen of 17 seeds improved; the exact two-sided sign test is $p = 0.0127$, conventionally rounded to 0.013. The same contrast appeared to be worth 7.5 points on the

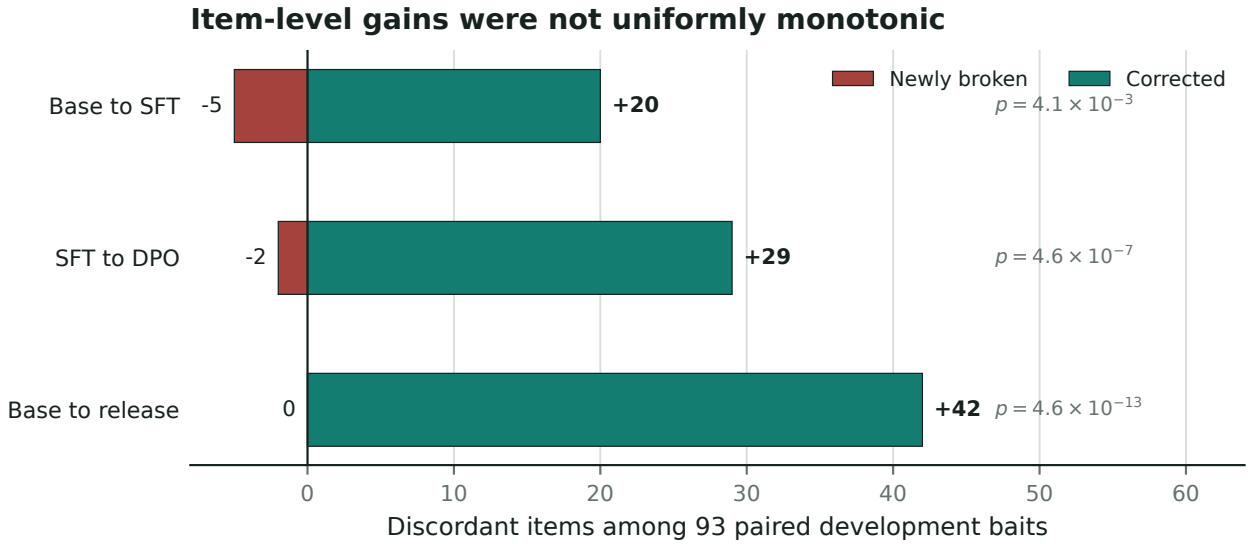

Figure 4: Discordant paired outcomes on development baits. Bars show corrected and newly broken items; p values are exact two-sided McNemar tests. Aggregate improvement does not imply monotonic item-level improvement.

reused development set, more than twice the post-freeze estimate.

Seed variance remains material. Across 23 replicates of the $\beta = 0.1$ recipe, honest-positive performance ranged from 83% to 97%. The released $\beta = 0.05$ recipe did not receive equivalent replication, so its single-checkpoint score is not an estimate of expected recipe performance.

A wider dose-response reportedly reached 2.92% mean fabrication at $\beta = 0.0125$ across seven models [14]. However, the same-seed $\beta = 0.0125$ checkpoint failed the adversarial gate: adversarial pass fell from 95% at the shipping $\beta = 0.05$ to 72%, jailbreak pass from 8/10 to 5/10, and safety-probe pass from 10/10 to 4/10. The shipping value was the lowest tested beta that cleared every internal launch gate, not the beta minimizing fabrication.

4.4 Reticence rather than demonstrated knowledge gain

On post-freeze baits, lower-beta models produced 17.2 specific assertions per 93 prompts, compared with 22.0 for matched higher-beta models (Figure 5). Observed error conditional on a specific assertion was 48.8% at $\beta = 0.05$ and 42.7% at $\beta = 0.1$; this difference was not tested for significance.

The supported interpretation is that lower beta made the model less willing to assert unsupported specifics. It did not demonstrate that the model became more accurate after choosing to assert. This distinction matters because abstention may be useful while still reducing usefulness through over-refusal, ignored output formats, or excessive hedging.

4.5 Capability trade-offs

Relative to the upstream base, the released model changed by $-0.59\,pp$ on MMLU, $-5.6\,pp$ on GSM8K, and $+3.0\,pp$ on TruthfulQA MC2 under the internal harness (Figure 6). Stage scores were 0.6161/0.6131/0.6102 for MMLU, 0.516/0.448/0.460 for GSM8K, and 0.5734/0.5397/0.6034 for TruthfulQA across base, SFT, and release. Most GSM8K loss appeared after SFT; DPO recovered 1.2 points.

A separate same-seed comparison across 13 seeds estimated the effect of changing beta alone on GSM8K at $-0.46\,pp$ (SD 2.70; 6/13 seeds improved), providing no evidence of a systematic beta-specific cost [14]. Under available summaries, the clearer capability loss is associated with the intervention overall, particularly SFT.

4.6 The deterministic evaluator misranked checkpoints

The cheap fabrication gate rewarded hedge or refusal language and could miss *hedge-then-fabricate* outputs. It assigned the released checkpoint a worse deterministic rate than the superseded checkpoint (10% versus 3%), while source-based adjudication found the reverse (8.6% versus 19.4%; Figure 7). Across 42 models with both scores, Spearman correlation was $\rho = 0.105$ ($p = 0.51$); across 16 post-freeze models it was -0.179 [14].

The failure was not merely a threshold-calibration problem: the evaluator produced the wrong ordering for the decision it governed. If an evaluator rewards the appearance of uncertainty instead of checking claims beneath it, training can optimize the appearance.

5 Discussion

5.1 What transferred

The measured behavioural intervention remained effective when applied to Mistral-7B-Instruct-v0.3. The post-freeze fabrication result is the strongest evidence because its prompts did not exist during checkpoint selection. Character, adversarial, and honest-positive results are weaker because their prompts were repeatedly used during devel-

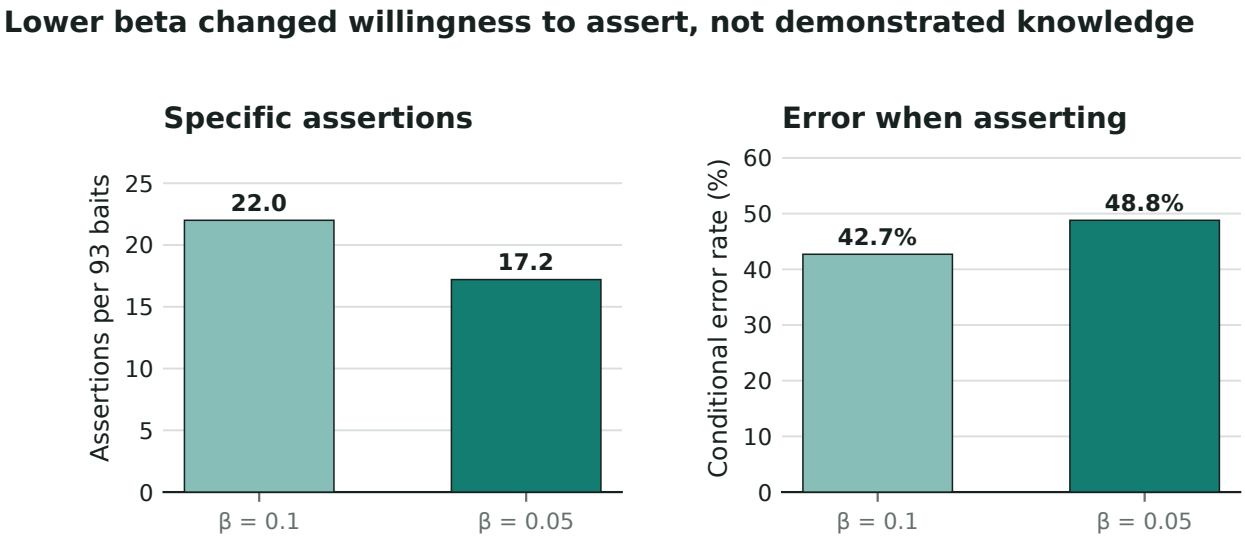

Figure 5: Mechanism-level summaries across matched seeds. Lower beta reduced the frequency of specific assertions, while observed conditional error did not improve. The experiment supports a reticence mechanism, not increased factual knowledge.

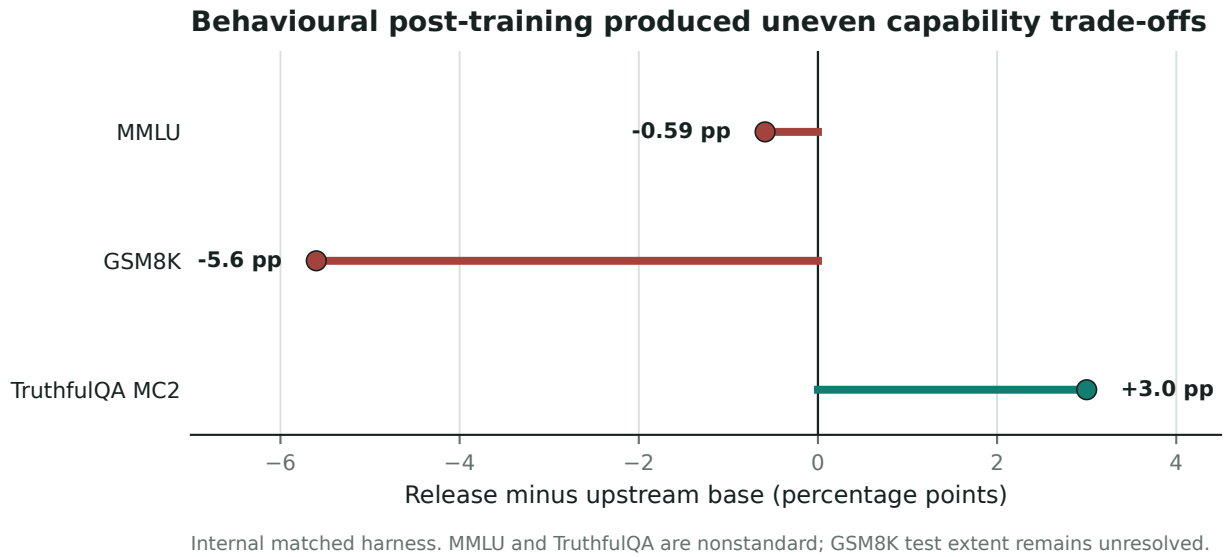

Figure 6: Release-minus-base capability deltas under the internal matched harness. MMLU and TruthfulQA protocols were nonstandard. The exact GSM8K test extent is unresolved across released artifacts, so the delta is not leaderboard-comparable.

opment.

The term *cross-lineage transfer* must remain bounded. Current artifacts show that targeted effects appeared on one Mistral checkpoint after applying the Vinci process. They do not fully specify the source-lineage intervention or demonstrate repeated, frozen application across several independent families. The evidence is consistent with transfer to one additional lineage, not general portability.

5.2 Abstention is not truthfulness

Lower fabrication can arise from reduced willingness to answer. A model that abstains appropriately may be safer, but it has not necessarily gained knowledge. Evaluation should therefore report assertion rate, accuracy conditional on assertion, and over-refusal on answerable inputs together. A single "honesty" score can collapse these behaviours and reward a model for saying less.

5.3 Capability preservation is part of alignment

The GSM8K regression shows why behavioural gains cannot be evaluated separately from task capability. A more cautious but less capable assistant may be suitable for a narrow application, but it is not an unqualified improvement. Capability-preservation gates should participate in checkpoint selection rather than being evaluated only after behavioural optimization.

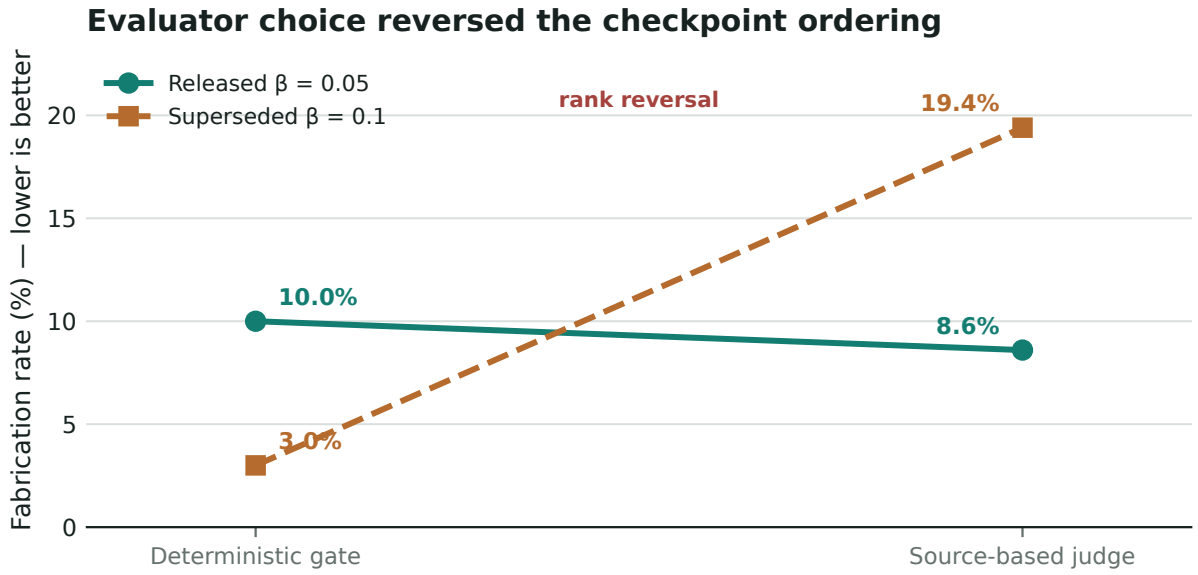

Figure 7: Checkpoint rank reversal. Lower is better on both axes. The deterministic gate preferred the checkpoint with more than twice the source-judged fabrication rate, making it unsuitable for checkpoint selection.

5.4 The evaluator is part of the training system

The regex gate was useful for inexpensive triage and unsuitable for ranking. Its failure mirrors the behaviour being trained: both gate and model overvalued uncertainty language. Evaluators used in optimization require validation not only for aggregate agreement but also for rank validity near the decision boundary they govern.

6 Limitations

The study has substantial limitations:

1. Only one additional lineage and a retired base checkpoint were evaluated.
2. The source-lineage recipe, SFT data and configuration, DPO construction, and complete environment are not publicly specified enough for strict replication.
3. Three behavioural axes and the development fabrication set were reused during iteration; only fabrication has a post-freeze set.
4. Evaluation sets are bespoke and small. Public abstention evaluations were not run.
5. Complete human adjudication was not performed; the judge used a floating GPT-4o alias, and retrieval grounding was incomplete for candidate passes.
6. Items not surfaced by the deterministic screen were counted as non-fabrications. The small negative audit does not tightly estimate missed errors.
7. Each fabrication set contains only 11 answerable controls, weakly measuring over-refusal.
8. Seed replication was unequal between beta conditions; row-level seed results and bootstrap settings are not public.
9. Capability protocols are nonstandard, and released artifacts conflict on GSM8K test extent.
10. The study does not evaluate latency, multi-turn use, tool use, long-context behaviour, agentic performance, or real-user outcomes.

These limitations do not erase the measured change. They determine its scope: a reproducible post-freeze effect under the released evaluation process, not a universal claim about truthfulness or deployment quality.

7 Artifacts and Intended Use

The merged checkpoint, model card, evaluation record, and source-confirmation record are publicly available [10, 14, 15]. The exact weights can be hash-verified. External reproduction of the training run requires additional release of the SFT parent, training corpora or manifests, complete configurations, dependency locks, per-item outputs, and statistical scripts.

Vinci Prova 7B 1.0 is an experimental research checkpoint. Remaining model-judged fabrications are concentrated in invented legal authorities and statutory references. The model should not be relied upon for legal, regulatory, financial, medical, or other high-stakes factual work. Adversarial-bait rates must not be presented as real-world hallucination prevalence.

Increased abstention can reduce unsupported assertions while denying legitimate assistance or masking capability failure behind cautious language. Deployment evaluation must therefore measure under-refusal and over-refusal together.

8 Conclusion

Applying the Vinci SFT+DPO character process to Mistral-7B-Instruct-v0.3 produced large changes on targeted behavioural evaluations. The strongest result, a reduction in model-judged fabrication from 46.2% to 7.5%,

reproduced on a post-freeze adversarial set. The mechanism, however, was primarily fewer specific assertions rather than improved accuracy conditional on asserting. The intervention also incurred a 5.6-point GSM8K regression under the internal matched harness, and a cheap evaluator ranked competing checkpoints incorrectly.

The result is useful precisely because it is bounded. It supports the claim that targeted behavioural effects appeared on one additional model lineage under the measured protocol. It does not establish universal portability, greater factual knowledge, or production readiness. Stronger evidence requires a fully frozen and public recipe across several current families, post-freeze evaluation of every behavioural axis, balanced answerability measurement, replicated seeds, human adjudication, and capability-preservation gates.

Author Contributions

Contributions follow the CRediT taxonomy; equal contribution is not claimed. **George Pu:** conceptualization, methodology, investigation, formal analysis, validation, project administration, and writing–original text. **Ayush Naik:** software, supporting methodology, technical validation, and writing–review and editing.

Competing Interests and AI Assistance

Both authors are affiliated with SimpleDirect / Vinci Research, the developer and releaser of Vinci Prova 7B 1.0. AI assistance from OpenAI Codex was used for manuscript organization, language editing, public-source consistency checks, recomputation of exact tests from published counts, and figure production. It did not conduct the underlying experiments or complete an independent human-equivalent adjudication.

References

[1] S. Maiya, H. Bartsch, N. Lambert, and E. Hubinger. "Open Character Training: Shaping the Persona of AI Assistants through Constitutional AI." arXiv:2511.01689, 2025. https://arxiv.org/abs/2511.01689

[2] 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." arXiv:2305.18290, 2023. https://arxiv.org/abs/2305.18290

[3] A. Q. Jiang et al. "Mistral 7B." arXiv:2310.06825, 2023. https://arxiv.org/abs/2310.06825

[4] P. Kirichenko, M. Ibrahim, K. Chaudhuri, and S. J. Bell. "AbstentionBench: Reasoning LLMs Fail on Unanswerable Questions." arXiv:2506.09038, 2025. https://arxiv.org/abs/2506.09038

[5] S. Zhai, J. Liang, and D. Kang. "Abstain-R1: Calibrated Abstention and Post-Refusal Clarification via Verifiable RL." ACL 2026; arXiv:2604.17073. https://arxiv.org/abs/2604.17073

[6] L. Zheng et al. "Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena." NeurIPS 2023; arXiv:2306.05685. https://arxiv.org/abs/2306.05685

[7] D. Hendrycks et al. "Measuring Massive Multitask Language Understanding." arXiv:2009.03300, 2020. https://arxiv.org/abs/2009.03300

[8] K. Cobbe et al. "Training Verifiers to Solve Math Word Problems." arXiv:2110.14168, 2021. https://arxiv.org/abs/2110.14168

[9] S. Lin, J. Hilton, and O. Evans. "TruthfulQA: Measuring How Models Mimic Human Falsehoods." ACL 2022; arXiv:2109.07958. https://arxiv.org/abs/2109.07958

[10] SimpleDirect. "Vinci Prova 7B 1.0." Hugging Face model card, 2026. https://huggingface.co/simpledirect/Vinci-Prova-7B-1.0

[11] G. Pu. "We tested whether character training transfers across model lineages. It did." SimpleDirect Research, 10 August 2026. https://www.getsimpledirect.com/blog/we-tested-whether-character-training-transfers-across-model-lineages-it-did

[12] E. J. Hu et al. "LoRA: Low-Rank Adaptation of Large Language Models." arXiv:2106.09685, 2021. https://arxiv.org/abs/2106.09685

[13] Mistral AI. "Mistral 7B v0.3 Model Card." Accessed 13 August 2026. https://docs.mistral.ai/models/model-cards/mistral-7b-0-3

[14] SimpleDirect. "EVAL – Vinci Prova 7B 1.0." Hugging Face evaluation record, 2026. https://huggingface.co/simpledirect/Vinci-Prova-7B-1.0/blob/main/EVAL.md

[15] SimpleDirect. "Source Audit – Vinci Prova 7B 1.0." Hugging Face source-confirmation record, 2026. https://huggingface.co/simpledirect/Vinci-Prova-7B-1.0/blob/main/SOURCE-AUDIT.md