

TRA: Persistent Post-Deployment Learning in a Quantized Transformer

A Generational Low-Rank Runtime for Incremental Knowledge Updating Without Rewriting the Base Model

Scientific manuscript - research snapshot 3 September 2026

Dany Bouchard - Founder, System Architect & Principal Investigator

Atlas - Research Guide & Structure Manager

Independent TRA Research

Contribution statement

Dany Bouchard conceived and leads the TRA research program, defined its research questions and architectural direction, directed the experimental program, developed technical hypotheses from observed model behavior, and made the decisions governing subsequent experiments. Atlas, an OpenAI AI system used throughout the research process, served as Research Guide & Structure Manager: structuring hypotheses and experimental observations, maintaining semantic continuity across the research program, helping formulate controlled experiments and interpret their results, organizing the research record, relating observations to scientific literature, and structuring and drafting the manuscript. Experimental implementation and artifact extraction were performed through AI-assisted development tools under Dany Bouchard's direction. Dany Bouchard retains responsibility for the research claims and final publication.

Abstract

Large language models are trained from finite historical corpora and do not automatically incorporate later events into frozen pretrained weights. Full-model retraining or fine-tuning can update a model, but repeatedly modifying an entire model to incorporate a small number of new facts is computationally and operationally disproportionate to the size of the update. TRA (Transformer Relearn Architecture) investigates whether a deployed quantized transformer can instead acquire persistent post-deployment experience while its original GGUF remains immutable, learned changes remain reversible and versioned, and inference remains available. Experiments use Granite 3.3 2B Instruct Q4_K_M and real post-horizon name-transition events, principally Nauru to Republic of Naoero. TRA progressed from ineffective bounded FFN_DOWN deformation to distributed rank-8 LoRA attached to Q/K/V/O projections across all 40 layers. The experiments measured acquisition, update-magnitude sensitivity, interference, semantic preservation, related-experience transfer, forgetting, generational freezing, Wake/Dream consolidation, persistence, and deterministic runtime time injection. The strongest historical G1-9 heuristic score was B=3/6 with A=4/5 after one pass, but raw outputs show partial lexical acquisition rather than clean semantic understanding; this paper therefore distinguishes deterministic surface-token scores from qualitative semantic correctness. Related learning transiently improved Naoero from 3/6 to 4/6 without direct retraining, followed by degradation to 1/6 in a single plastic bank. Freezing a useful generation and allocating fresh capacity improved the final Naoero trajectory to 4/6 despite temporary cross-bank interference. A deliberately bounded one-pass Wake phase limits deformation from a single observation, while Dream replays already authorized evidence for later consolidation; G1-19 improved a Wake result from 2/6 to 3/6 in two Dream iterations. Production LRMD state is versioned, non-destructive, and rollback-capable. TIJ demonstrates that deterministic changing state such as current time should be supplied at runtime rather than permanently learned. TRA does not establish general lifelong learning; it provides a transparent systems architecture and a documented set of observations on one quantized transformer.

Keywords

continual learning; post-deployment learning; knowledge editing; LoRA; catastrophic forgetting; quantized transformer; persistent memory; consolidation; model editing; runtime knowledge

1. Introduction: the stale-model problem

A pretrained language model is a historical snapshot. Once its weights are frozen, later events do not become part of its parametric memory merely because the world changes. Prompt context or retrieval can provide current information temporarily, but neither mechanism is equivalent to persistent learning inside the deployed model.

The Granite instance used in TRA displayed this limitation operationally. In cold contexts it lacked selected later relationships and often treated later dates or events as future, hypothetical, or inconsistent with its prior. We describe this as the model's observed effective knowledge horizon. We deliberately do not equate that behavior with a proven training-dataset cutoff: a model's self-description is not a reliable audit of its corpus.

If a deployed model is missing a small number of new facts, rebuilding or extensively fine-tuning the whole model is a costly response to a small update. LoRA was introduced to reduce the trainable-parameter and memory burden of full fine-tuning [3], while knowledge-editing research explicitly studies efficient updates without exhaustive retraining [5]. TRA extends this into a runtime question: can a model accumulate small amounts of persistent experience while the original quantized model remains immutable and recoverable?

TRA optimizes for net useful learning over time, not perfect memorization after every exposure. One observation may form a weak plastic trace. Repetition or consolidation may strengthen it. Some forgetting is expected. Failure means destructive broad regression, corruption, inability to continue learning, loss of authoritative learned state, broken rollback, or catastrophic semantic collapse.

2. Research questions

- Acquisition of real post-horizon relationships
- Write/read separation
- Update-magnitude stability
- Interference mechanism
- Semantic preservation
- Related-experience transfer
- Generational freezing
- Wake/Dream consolidation
- Persistent restartable learned state
- Boundary between learned experience and deterministic runtime reality

3. Relation to prior work

ROME used causal tracing to identify computations decisive for factual predictions and introduced rank-one editing of factual associations [1]. MEMIT extended direct editing to thousands of associations in GPT-J and GPT-NeoX [2].

LoRA freezes pretrained weights and trains low-rank matrices, reducing trainable parameters relative to full fine-tuning [3]. TRA adopts this broad separation principle but treats low-rank state as persistent post-deployment experience.

Elastic Weight Consolidation protects parameters important to previous tasks [4]. Cohen et al. showed that factual edits have related ripple effects [6]. Gupta et al. found progressive followed by catastrophic forgetting during sequential editing with ROME and MEMIT [7]. Hsueh et al. documented knowledge distortion and degradation of general abilities after editing [8]. TRA therefore evaluates acquisition together with preservation, transfer, interference, rollback, and continued editability.

4. TRA architecture

The original Granite GGUF remains immutable during ordinary TRA learning. Q4_K weights remain quantized. Conceptually, active inference is $y = W_base \cdot x + \Delta W \cdot x$; disabling learned state restores the frozen base path.

HERE IS THE NEW FACT: authorizes external learning evidence. It does not mean absolute truth. Internal thought and Dream replay may reorganize already authorized evidence but may never manufacture a new authoritative learning event.

Production LRMD is authoritative learned history. Tests may use disposable state, but production retains generations, provenance, knowledge-event history, metadata, and rollback.

State	Meaning
PLASTIC	May receive gradients.
FROZEN	No longer trained but active.
CONSOLIDATED	Qualified useful long-term state.
DISABLED	Preserved but inactive for rollback/comparison.

5. Experimental platform and reproducibility status

Item	Recorded value
Base model	granite-3.3-2b-instruct-Q4_K_M.gguf
Model SHA256	AC71E9E32C0BEA919B409C5918F69CA74339854B0319C5065E4E9FB6D95C4852
Clean ancestor	TRA-LEARN1; upstream parent b06aa774c03dbbb624e726664b714a57d1f49815
Representative flags	-ngl 99 -c 131072 -np 1 -t 8 -tb 8

Baseline throughput	approximately 33.2-33.5 tokens/s
Probe decoding	seed 1; temperature 0; top_k 1; top_p 1; min_p 0
LoRA initialization	$A \sim N(0, 0.01)$, seed 42; $B=0$
Serving	single :8080 frontend; two private Granite backends

The evidence audit recovered 21/30 requested reproducibility fields as FOUND, 7 PARTIAL, and 2 NOT FOUND. Prompts, scorers, seeds, LoRA configurations, raw post-learning answers, Wake/Dream results, and TIJ Windows results are largely preserved. Historical server-binary-to-git-commit provenance for G1-9 through G1-19 is unpinned, so exact historical binary reproducibility is not claimed. The GGUF identity is pinned.

TRA distinguishes HOT evaluation, where teaching text may remain in ordinary context/KV, from COLD evaluation, where the teaching conversation is absent. Persistent-learning claims require cold behavior.

6. Factual learning targets and authoritative verification

Name transitions were chosen because the model can already know the underlying entity while lacking a later relationship. This permits an additive test: retain the entity and old facts while acquiring a new current/former-name relation.

Event	Historical TRA input	Authoritative status used in publication
Nauru -> Republic of Naoero	Completed 2026 rename; no URL stored in historical G1-9 artifact.	UN: on 26 Jun 2026 Nauru informed the UN it had changed its name to Republic of Naoero [10].
Kerala -> Keralam	G1-17 taught a completed/in-force Aug 2026 rename.	Official 24 Feb 2026 record says Cabinet approved a proposal and further constitutional/parliamentary steps were to follow [11]. Historical input was factually overstated.
Beyond, Inc. -> Bed Bath & Beyond, Inc.	Completed Aug 2025 rename.	Company announced 18 Aug 2025 that name/ticker change would be effective 29 Aug 2025 [12].
Santander Bank Polska -> Erste Bank Polska	Completed Apr 2026 rename.	Erste Current Report 28 records name change and registry update on 24 Apr 2026 [13].
IAC Inc. -> People Incorporated	Completed Jun 2026 rename.	SEC 8-K: changed effective 4 Jun 2026 [14].
Skillz Inc. -> Firy Inc.	Completed Jun 2026 rename.	SEC 8-K: changed on 18 Jun 2026 [15].

The Keralam discrepancy is not hidden or retroactively corrected. The experiment is reported against the sentence it actually learned, while the paper separately reports the authoritative real-world status. This is a direct demonstration of why provenance must accompany learned evidence.

7. Experiment 1 - Naoero acquisition

G1-9 used six B-side teaching statements expressing current-name, former-name, same-entity, and rename relationships between Nauru and Naoero. Six A-side statements reinforced stable Nauru geography. The archived corpus SHA256 is B336A3FB55FDB95F71C2D82F1E639E0FC7CC04CF0091F8121B55C073A5E63C16.

B probe	Prompt
current_name	What is the current name of Nauru?
formerly	What was Naoero formerly called?
same	Are Nauru and Naoero the same country?
renamed	What country was renamed Naoero?
what_naoero	What is Naoero?
where_naoero	Where is Naoero located?
A/control probe	Prompt

what_nauru	What is Nauru?
where_nauru	Where is Nauru located?
is_country	Is Nauru a country?
region	What region of the world is Nauru in?
capital	What is the capital of Nauru?

Four unrelated regression probes checked $V=RI$ (20 V), $W=VI$ (36 W), inverse Ohm law (3 A), and Paris as the capital of France.

Bounded spatial FFN_DOWN learning could improve cross-entropy while B remained 0/6. The best G1-9 substrate instead used rank-8 LoRA on q/k/v/o projections in all 40 layers: 160 slots and 4,259,840 trainable parameters. The update used sum-gradient SGD, $lr=0.2$, $max_delta_abs=0.2$, Frobenius clamp 200, and up to 10 passes. Pass 1 reached $B=3/6$ and $A=4/5$; pass 2 collapsed to $B=0/A=0$.

8. Scoring methodology and its limitation

B and A scores are deterministic surface-token heuristics, not human semantic judgments. `score_b` lowercases output and checks required substrings. `current_name` passes when 'naoero' appears; formerly when 'nauru' appears; same when 'same' or a standalone 'yes' appears unless a negative phrase is present; renamed when 'nauru' appears; the descriptive probes require combinations of entity and geography tokens. `score_a` similarly uses Nauru/geography tokens and Yaren/capital cues.

This scorer is reproducible but imperfect: fluent semantic equivalence can fail, and an incorrect answer can accidentally pass. The raw G1-9 pass-1 output illustrates the problem. `current_name` repeatedly emitted Naoero and passed; renamed contained Nauru and passed; same contained the word same and passed despite confused reasoning. Other answers described Naoero as a village and failed. Therefore $B=3/6$ is interpreted as partial lexical/relational acquisition, not as 50% semantic understanding.

This distinction strengthens rather than weakens the experimental conclusion: the adapter changed retrieval behavior in the intended lexical directions, but the model had not yet formed a clean, stable representation. TRA subsequently treats behavioral quality, preservation, and raw-answer inspection as more important than loss or token-score maximization alone.

9. Stability, interference, and preservation

G1-10 confirmed LoRA overshoot. G1-13 showed that effective step magnitude dominated optimizer family. At raw $LR=0.2$, recorded step Frobenius norms were about 70.37 for SGD and 270.32 for Adam/AdamW. Controlled scaling found comparatively safe actual step norms at or below about 0.01, instability around 0.025-0.10, and collapse at much larger effective steps. Loss could improve while semantic links churned.

G1-11 tested whether interference was explained by overlap with the learned LoRA subspace. Only about 0.51% of the next gradient energy was parallel to that protected subspace, and random protection behaved similarly. The tested interference therefore was not adequately explained by simple subspace overlap.

G1-14 added property-aware preservation loss. Starting from $B=3/6$ and $A=4/5$, $lambda_keep$ 0.25/1.0 produced $B=4/6$ and gained formerly while retaining renamed, but A fell to 3/5. Preservation helped one additive semantic objective while imposing a measurable trade-off.

10. Write/read separation

G1-15 tested five 8-layer adapter bands: 0-7, 8-15, 16-23, 24-31, and 32-39. All failed to acquire B. Partial bands used about 852K parameters and were about 1.6x faster; all-layer attachment used about 4.26M parameters and remained effective. A previously interesting causal read site near layer 32 therefore did not imply an effective write site. TRA records this as WRITE/READ SEPARATION.

11. Related experience, transfer, and forgetting

G1-17 trained a sequence of related name-transition events in one shared plastic adapter after the Naoero anchor. The archived order was Keralam, Bed Bath & Beyond, Erste Bank Polska, People Incorporated, and Firy. Naoero was not directly retrained after its anchor.

Naoero moved from 3/6 at anchor to a transient 4/6 after the first Keralam learning step, gaining formerly without direct Naoero training. Continued shared-bank learning later degraded Naoero to 1/6. This is evidence of transient positive transfer followed by forgetting. Because the Keralam teaching statement was factually overstated, the result is interpreted as transfer between related linguistic/name-transition structures, not as validation of the Keralam world fact.

The broader phenomenon is consistent with the sequential-editing concern reported by Gupta et al. [7], while the mechanism observed in TRA is not claimed to be identical to ROME or MEMIT.

12. Generational adapters

G1-18 froze the useful Naoero/Keralam state as BANK1 and created fresh BANK2 capacity for later events. The earlier single plastic bank produced Naoero 3 -> 4 -> 1. Freeze plus fresh BANK2 produced 3 -> 4 -> fluctuations -> final 4. BANK1 parameters remained frozen, although BANK2 could still temporarily alter inference. Parameter isolation therefore did not guarantee behavioral isolation, but final retention was substantially better than with one continuously plastic bank.

Bank	State	Recorded event refs
BANK1_CONSOLIDATED	FROZEN	NAOERO_ANCHOR, E_KERALAM
BANK2_PLASTIC	PLASTIC	E_BBBY, E_ERSTE_PL, E_PEOPLE_INC, E_FIRY

13. Why one pass?

TRA deliberately separates immediate acquisition from later consolidation. The daytime learning path uses one bounded impregnation pass because the first objective is not to force maximum memorization. It is to create the smallest useful plastic trace that can be evaluated, persisted, and rolled back.

This policy follows directly from the experiments: large steps overshoot; aggressive optimization could improve loss while damaging semantic behavior; and continued plastic training could erase earlier gains. One pass limits how much a single observation can deform the deployed model and preserves the epistemic meaning of one exposure: one piece of evidence should not automatically become over-reinforced certainty.

Operationally, one pass also bounds foreground learning work. The approximately five-minute value observed in the current implementation is knowledge-commit/impregnation latency, not wake-up latency. The other private Granite model continues serving inference while the learner prepares a candidate generation.

14. Why Dream?

Dream exists because a deliberately weak first trace may be useful but incomplete. Its purpose is consolidation, not discovery of new truth. Dream replays already authorized evidence and related semantic context so that TRA can strengthen useful relationships, test retention, and qualify a more stable generation without pretending that replay is a new external confirmation.

G1-19 qualified the lifecycle: new fact -> one-pass impregnation -> persist -> pending_retrain=true -> model swap -> Dream -> consolidation -> pending_retrain=false -> wake. In the recorded example, Wake produced B=2/6 and two Dream iterations improved the result to 3/6.

The gain is architectural as well as behavioral. Daytime learning stays bounded and reversible; Dream can spend additional computation during maintenance while qualification gates decide whether the consolidated state should remain active. Replay never creates a new HERE IS THE NEW FACT event and therefore cannot falsely increase external confirmation count.

Dream is interrupt-safe. One private model remains READY while the other is DREAM_TRAINER. YOU CAN WAKE UP promotes READY to LIVE immediately; the runtime does not wait for the current Dream operation to finish.

15. Dual-model runtime and production persistence

TRA exposes one public assistant at :8080 while two private Granite instances take LIVE, READY, SHADOW/LEARNER, or DREAM_TRAINER roles. The frontend owns conversation continuity, routing, TRA command interception, runtime context, and learned-generation coordination. Candidate state is activated only at controlled boundaries; inference outranks learning and Dream.

By G1-21, authoritative production learned state had reached Generation 4 with four accumulated knowledge events preserved. Production LRMD v2 refused destructive reset, truncate, and delete operations; destructive tests were restricted to TEST state; and a qualified generation archive remained available for rollback/history. Later freeze bookkeeping reached generation 5, so the paper reports Generation 4 specifically as the pre/post-G1-21 production-learning state rather than treating later live-store drift as part of that experiment.

16. TIJ: what should not be learned

Current date/time becomes obsolete continuously. Training 'today is 2026-09-03' would waste learning capacity and immediately age. TRA therefore introduced the Temporal Injection Junction (TIJ), a protected runtime KV region carrying deterministic current time.

G1-21 used portable C++ `std::chrono`, a fixed 96-token temporal slot, and minute-based refresh without an LRMD generation increment. With TIJ disabled, Granite expressed its older temporal prior. With TIJ enabled, it retrieved 2026, 2026-09-03, and current runtime time; A/B temporal state was synchronized. Physical Linux qualification was still pending at the research snapshot.

TIJ exposed a distinction between having current data and understanding that one has the capability to use it. The clock value belongs in deterministic runtime state. A durable higher-level concept such as 'this deployed TRA model can consult a trusted runtime clock' may be learned as capability experience. This paper intentionally stops at that TRA/RKV boundary and excludes Virtual Room and later action-tool concepts.

17. Experimental chronology

Experiment	Principal question/result
G1-9	All-layer Q/K/V/O r8 LoRA: baseline B=0/A=5; pass1 B=3/A=4; pass2 collapse. Partial lexical acquisition.
G1-10	Micro-step continuation confirmed LoRA overshoot and rollback behavior.
G1-11	Protected learned subspace did not prevent wipe; ~0.51% next-gradient energy parallel to protected subspace.
G1-12	Specialist-bank routing did not beat G1-9 peak.
G1-13	Step norm dominated optimizer family; useful behavior confined to narrow update region.
G1-14	Property-aware preservation gained formerly but reduced A 4->3; no promotion.
G1-15	Five 8-layer bands failed target acquisition; all-layer remained best.
G1-16	Conventional full fine-tuning comparison not executed because suitable FP ancestor was absent and Q4 full FT was forbidden.
G1-17	Related events: transient Naoero 3->4 transfer, then end-state 1/6 in shared plastic bank.
G1-18	Frozen BANK1 + fresh BANK2: final Naoero 4/6 despite temporary cross-bank interference.
G1-19	Wake/Dream lifecycle qualified; Wake 2/6 -> Dream 3/6.
G1-21	Production-memory protection and TIJ qualified on Windows; Linux physical TIJ test pending.

18. Discussion

The experiments support a limited but coherent interpretation. Useful learning in this Granite configuration was distributed rather than captured by the tested narrow write regions. Effective update magnitude was critical. Training loss and token heuristics could improve without clean semantic understanding. Related experiences could transfer positively, but the same shared plasticity could later erase that gain. Freezing useful generations and allocating fresh capacity improved final retention, although new banks could still interfere behaviorally with frozen banks.

The one-pass/Dream split is therefore not merely an anthropomorphic metaphor. It implements a specific optimization policy: cautious low-exposure acquisition during normal operation, followed by optional later consolidation using only previously authorized evidence. This separates the epistemic weight of one external observation from the amount of compute available for consolidation.

The Keralam artifact also exposes a second lesson. A continual-learning system can faithfully learn an incorrect external statement if provenance and factual status are not handled carefully. TRA therefore requires evidence provenance and distinguishes external evidence from absolute truth. The publication preserves the historical mistake rather than retroactively sanitizing it.

19. Limitations and reproducibility

The transformer evidence is currently restricted to Granite 3.3 2B Instruct Q4_K_M. Probe sets are development-scale and the historical B/A metrics are surface-token heuristics, not standardized semantic benchmarks. Historical G1-9 through G1-19 server binaries are not cryptographically bound to exact git commits. Full baseline answer strings for the best G1-9 trial were not preserved, although baseline scores and post-learning raw answers were. LoRA-era VRAM measurements are incomplete. Linux TIJ qualification remained pending.

These limitations do not invalidate the historical observations, but they constrain their scope. A public replication package should freeze a current code commit and executable hash, publish the exact prompts, teaching sentences, seeds, event sources, adapter configuration, optimizer parameters, raw outputs, scoring code, throughput, VRAM, rollback, and restart-persistence results. A golden reproduction of the principal Naoero path from a pinned build would strengthen reproducibility without pretending that historical binary provenance is available.

The present empirical claims are restricted to the tested Granite configuration. Replication on additional architectures and model scales is future work required to determine whether the observed TRA dynamics generalize beyond this configuration.

20. Conclusion

TRA addresses a practical limitation of static deployed language models: a small number of verified later facts should not necessarily require rebuilding or replacing the entire pretrained model. The experiments demonstrate that a quantized frozen Granite model can be placed inside a reversible generational runtime that acquires partial new relationships, experiences transfer and forgetting, preserves useful learned generations, consolidates bounded Wake traces during Dream, and maintains persistent learned state without rewriting the base GGUF.

The work also demonstrates why continual learning cannot be judged by a single score. G1-9 changed the intended lexical relations but raw answers remained confused. More training could collapse the model. Related learning could help and later hurt. Frozen parameters could still suffer behavioral interference from a new bank. An external teaching statement could itself be wrong. These observations motivate TRA as a systems architecture for cautious evidence accumulation rather than a mechanism for blindly forcing facts.

TRA is not presented as a solution to lifelong learning. It is a documented experimental step toward a deployed model that can learn incrementally, remain available, preserve rollback, distinguish evidence from truth, and reserve deterministic runtime reality for mechanisms such as TIJ rather than wasting persistent learning capacity.

21. References

- [1] Meng, K., Bau, D., Andonian, A., & Belinkov, Y. (2022). Locating and Editing Factual Associations in GPT. *Advances in Neural Information Processing Systems* 35 (NeurIPS 2022). DOI: 10.52202/068431-1262.
- [2] Meng, K., Sen Sharma, A., Andonian, A., Belinkov, Y., & Bau, D. (2023). Mass-Editing Memory in a Transformer. *International Conference on Learning Representations (ICLR 2023)*.

- [3] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2022). LoRA: Low-Rank Adaptation of Large Language Models. ICLR 2022. arXiv:2106.09685.
- [4] Kirkpatrick, J., Pascanu, R., Rabinowitz, N., et al. (2017). Overcoming catastrophic forgetting in neural networks. Proceedings of the National Academy of Sciences, 114(13), 3521-3526. DOI: 10.1073/pnas.1611835114.
- [5] Zhang, N., Yao, Y., & Deng, S. (2024). Knowledge Editing for Large Language Models. Proceedings of LREC-COLING 2024: Tutorial Summaries, 33-41.
- [6] Cohen, R., Biran, E., Yoran, O., Globerson, A., & Geva, M. (2024). Evaluating the Ripple Effects of Knowledge Editing in Language Models. Transactions of the Association for Computational Linguistics, 12, 283-298. DOI: 10.1162/tacl_a_00644.
- [7] Gupta, A., Rao, A., & Anumanchipalli, G. (2024). Model Editing at Scale leads to Gradual and Catastrophic Forgetting. Findings of ACL 2024, 15202-15232. DOI: 10.18653/v1/2024.findings-acl.902.
- [8] Hsueh, C.-H., Huang, P. K.-M., Lin, T.-H., Liao, C. W., Fang, H.-C., Huang, C.-W., & Chen, Y.-N. (2024). Editing the Mind of Giants: An In-Depth Exploration of Pitfalls of Knowledge Editing in Large Language Models. Findings of EMNLP 2024, 9417-9429.
- [9] ggml-org. llama.cpp source and documentation. Official project repository.
- [10] United Nations. Naoero - Member States. Record states that on 26 June 2026 Nauru informed the United Nations that it had changed its name to Republic of Naoero.
- [11] Government of India, Prime Minister / Press Information Bureau. Cabinet approves alteration of the name of the State of Kerala to Keralam. 24 February 2026. The release records Cabinet approval of a proposal and describes subsequent constitutional steps.
- [12] Bed Bath & Beyond, Inc. (formerly Beyond, Inc.). Beyond, Inc. Changes Name to Bed Bath & Beyond, Inc. and Reclaims Ticker Symbol BBBY. 18 August 2025; effective 29 August 2025.
- [13] Erste Bank Polska S.A. Current Report no. 28 (2026), 24 April 2026. Change of Santander Bank Polska S.A. name to Erste Bank Polska S.A. and registry amendments.
- [14] U.S. Securities and Exchange Commission. IAC Inc. / People Incorporated Form 8-K, 4 June 2026.
- [15] U.S. Securities and Exchange Commission. Skillz Inc. / Firy Inc. Form 8-K, 18 June 2026.

Appendix A - Historical G1-9 scoring caveat

The publication evidence closure confirms that B=3/6 and A=4/5 are produced by executable token/sub-string heuristics. The scorer is case-insensitive and does not generally recognize synonyms. Fluent semantic equivalence can fail; incorrect answers can pass. For example, an inverted Keralam answer could satisfy current_name merely by containing the target token. Accordingly, all historical B/A values in this manuscript are labeled heuristic scores and interpreted alongside raw outputs.

Appendix B - Code provenance status

G1-9 through G1-15 and G1-17 through G1-19 are historically UNPINNED with respect to exact llama-server executable SHA256 and source commit. G1-16 is PARTIALLY PINNED because the GGUF identity is recorded, but the training comparison was not executed. A later TRA-RUNTIME-V1 freeze commit exists but is not retroactively claimed as the binary used in earlier experiments.