Generate-then-Verify Is Not Enough: Routing Beats Verifiers on a Simulated Receptionist Suite

Animesh Alang
University of California, Santa Cruz
aalang@ucsc.edu

Abstract

Closed-knowledge AI receptionists still invent prices and comply with jailbreaks. We reimplement the `tiffani.app` stack—dense retrieval over owner docs, a separate generate-then-verify pass, and deterministic pre-model routing for handoff and danger—and evaluate it on **five fully simulated businesses and 180 simulated customer queries** (no production traffic). Against retrieval-only, single-pass RAG, and same-model self-check baselines on Llama-3.1-8B-Instruct-FP8, a verifier does *not* significantly cut overall hallucination vs single-pass (4.4%→3.3%, McNemar p=0.69 on seed 42; p=0.25 on seed 43). The clear win is routing: correctness on handoff+danger jumps from 28% to 89% (p=0.001). Verify does beat self-check on correctness (80.6%→86.1%, p=0.021) at lower latency. We report category breakdowns, error types, a Claude agreement sample (κ=0.77 on hallucination), and remaining gaps.

1 Introduction

Appointment businesses that publish a fixed knowledge base still get hurt when a receptionist invents a price or complies with a jailbreak. RAG helps (Lewis et al., 2020; Karpukhin et al., 2020), but grounded models still hallucinate (Ji et al., 2023). Product systems add self-critique (Manakul et al., 2023; Asai et al., 2024), chain-of-verification (Dhuliawala et al., 2024), or citation constraints (Gao et al., 2023; Min et al., 2023). What is less clear is the cost of a second pass when only owner-published text is allowed, booking needs an explicit YES, and some messages should never reach a model.

We study a faithful reimplementation of `tiffani.app`: bge-small retrieval (Xiao et al., 2024), generate-then-verify for grounding, and regex routing for human handoff and unsafe asks. **RQ1:** does a separate verifier reduce hallucinations vs single-pass RAG? **RQ2:** latency and false-refusal cost? **RQ3:** does deterministic routing help on handoff/danger without wrecking ordinary questions?

Contribution. On a category-stratified *simulated* receptionist suite with retrieval-only, single-pass, self-check, verify, and verify+route arms, multi-seed paired tests, and an error taxonomy: generate-then-verify is *not* a free win on overall hallucination; routing is the significant lever for handoff/danger; verify beats self-check on correctness, not on hallucination rate. We do not claim production causal evidence.

2 Method

Simulator. Five synthetic businesses (salon, dental, tutoring, vet, spa), each with eight owner docs. 180 evaluation queries are stratified from hand-written seeds plus light paraphrases (answerable 89, partial 27, unanswerable 31, adversarial 15, handoff 12, dangerous 6). Categories and counts were fixed before generation. Customers are simulated; no real chats.

Arms. (1) **retrieval-only**: top-2 chunks in a fixed template, no generation. (2) **single-pass**: RAG generate. (3) **self-check**: generate, then the same model critiques and may rewrite. (4) **verify**: generate, then a separate verifier prompt (same weights, different system role) may rewrite. (5) **verify+route**: regex router handles handoff/danger before any model call; else verify.

Models and judge. Primary: `@cf/meta/llama-3.1-8b-instruct-fp8` via Cloudflare Workers AI. Secondary Llama-3.2-3B and Mistral-7B runs were started then dropped after free-tier neuron exhaustion; we report that failure rather than invent numbers. Labels come from deterministic rules (unsupported money/phones, complimentary/free claims, jail-

Arm	Halluc.	Correct	False ref.	Lat. (s)
Retrieval-only	1.7	72.2	0.0	0.00
Single-pass	4.4	84.4	0.0	2.26
Self-check	2.2	80.6	0.0	6.75
Verify	3.3	86.1	0.0	5.06
Verify+route	2.2	91.1	0.6	4.97

Table 1: Primary model, seed 42, N=180 simulated queries. Rates in %. Wilson intervals omitted for space; see released `summary.json`. Latency is Workers AI proxy wall time, not production.

break compliance, refusal/overlap for answerable, 911/988 for danger, handoff phrases). A 20-item Claude Sonnet 4.5 sample agrees with the rule judge on hallucination at 95% (κ=0.77) and on correctness at 70% (κ=0.24); disagreements concentrate on danger (rules require crisis numbers; Claude accepts soft deflection) and truncated knowledge windows in the judge prompt. Decode seeds 42 and 43 for the single vs verify contrast; seed 44 was dropped with the neuron budget.

Metrics. Hallucination, correct-response, false-refusal rates; wall-clock latency. Wilson 95% CIs per arm; exact McNemar tests and bootstrap CIs on paired differences.

3 Results

On overall hallucination, single-pass (4.4%) vs verify (3.3%) is not significant (McNemar p=0.69, bootstrap Δ CI $[-1.7, +3.9]$ pp). Seed 43 repeats the pattern (4.4%→2.8%, p=0.25). Self-check posts the lowest halluc rate among generative arms (2.2%) but the worst correctness (80.6%); verify beats it on correctness (p=0.021) while staying faster (5.1s vs 6.8s). Retrieval-only almost never invents facts (1.7%) but collapses on answerable items that need phrasing (overall correct 72.2% vs verify 86.1%, p<0.001).

The significant effect is routing. On the 18 handoff+dangerous queries, correct replies rise from 28% (single-pass) to 89% (verify+route), McNemar p=0.001, bootstrap Δ CI $[-83, -39]$ pp. Ordinary answerable correctness stays ≈99%. Adversarial hallucination moves 47%→27% under verify (n=15, p=0.25): directionally helpful, underpowered.

Errors. Single-pass failures include complimentary "free today" bookings under jailbreak pressure, partial system-prompt dumps, and soft danger replies that never mention 911/988. Verify catches some invented prices and free offers but still leaks when asked to recite system instructions and still under-handoffs manager requests. Verify+route fixes the danger/handoff class by never calling the model. False refusals on answerable items are rare under our rule judge (≤0.6%).

4 Limitations

All users and businesses are simulated; templated paraphrases understate real dialogue messiness. Absolute rates depend on a rule judge; Claude agreement is high for hallucination and only moderate for correctness. Latency is a Workers AI proxy, not the production VPS. We reimplement the design; we do not A/B live `tiffani.app`. Secondary models and a third seed were planned and dropped after free-tier exhaustion—a real resource constraint, not a silent omission. Booking confirmation flows are out of scope. Do not read this as a claim that verifiers are useless in general; on *this* suite the verifier is not the significant lever.

Ethics Statement

No real customer data. Dangerous prompts are synthetic; routed replies point to 911/988 and refuse actionable harm. Code and synthetic knowledge bases accompany the preprint. This paper is a draft evaluation artifact and is not submitted for publication here.

References

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *Proceedings of ICLR*.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2024. Chain-of-verification reduces hallucination in large language models. In *Findings of ACL*, pages 3563–3578.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. Enabling large language models to generate text with citations. In *Proceedings of EMNLP*, pages 6465–6488.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and

Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of EMNLP*, pages 6769–6781.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*.

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of EMNLP*, pages 9004–9017.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Luke Zettlemoyer, Hannaneh Hajishirzi, et al. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of EMNLP*.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. 2024. C-Pack: Packed resources for general Chinese embeddings. In *Proceedings of SIGIR*.