Beyond Accuracy: A Qualitative Error Analysis of SLM Agents in Multi-Hop QA

Marcello Pelosi[1,*], Rajib Chandra Ghosh[1], Egidia Cirillo[1] and Flora Amato[1]

[1]*Department of Electrical Engineering and Information Technology (DIETI), University of Naples Federico II, Via Claudio 21, 80125 Naples, Italy*

Abstract

The adoption of Small Language Models (SLMs) as reasoning engines for autonomous agents has recently emerged as an interesting compromise between computational cost, latency, privacy constraints, and response quality. In this work, we present an empirical comparison of three open-weight SLMs (Gemma 4 e4b, Ministral-3 8B, and Qwen 3.5 9B), employed as ReAct agents equipped with tool access to the Wikipedia API, against a larger LLM baseline (Gemini 3.5 Flash) run under the same agentic architecture. The models are evaluated on a stratified sample of 75 questions drawn from HotpotQA (25 *easy*, 25 *medium*, 25 *hard*). Answers are judged by an external judge LLM (DeepSeek V4-Flash) into three categories (correct, partially correct, incorrect), each supplemented by a textual rationale; a manual validation of a subsample corroborates the automatic judgments. Among the SLMs, Qwen 3.5 9B attains the best accuracy (58.7%, 63.3% *soft*) and Ministral-3 the worst (40.0%, 46.0%), but pairwise tests show that most differences between SLMs are not statistically significant at this sample size; the LLM baseline is significantly more accurate than the two weaker SLMs, while its advantage over the strongest SLM does not reach statistical significance at this sample size. Beyond accuracy, a full qualitative error analysis, performed automatically over the judge rationales and validated manually, identifies five recurrent failure patterns (incorrect Wikipedia page selection, multi-hop failure, factual hallucination, empty or evasive output, arithmetic/comparative reasoning error) and shows that each model exhibits a distinctive error profile not reducible to overall accuracy alone.

Keywords

Small Language Models, autonomous agents, multi-hop question answering, HotpotQA, ReAct, LLM-as-a-Judge

1. Introduction

Large Language Models (LLMs) are increasingly adopted for building autonomous agents capable of planning and interacting with external tools [1], but their high computational cost limits their adoption in scenarios requiring low latency, on-device execution, or strict privacy constraints [2]. The emergence of Small Language Models (SLMs) offers a promising alternative to the aforementioned issues. Recent work suggests that, in some contexts, SLMs can match or outperform larger models on specific tasks while reducing operational cost [3]. We deliberately avoid framing tool-augmented LLM agents as a settled "de facto standard": although their adoption is growing rapidly, the field is still consolidating, and the present study is precisely an attempt to characterise where compact models stand.

The actual ability of compact models to robustly orchestrate multi-step reasoning chains and effectively query external knowledge bases remains an open problem. The execution of complex procedures and iterative planning expose SLMs to specific vulnerabilities, such as tool hallucination or the loss of contextual coherence during thought-action loops. These reliability concerns are not unique to question answering: analogous issues of trustworthiness and robustness under data poisoning, device failures, and unstable communication have been studied for reinforcement-learning agents in industrial settings [4], where the same tension between agent autonomy and verifiable correctness arises.

To help bridge this gap, this work proposes an empirical investigation of the agentic capabilities of state-of-the-art SLMs. We compare three open-weight models (Gemma 4 e4b, Ministral-3 8B, and

WOA 2026: 27th Workshop From Objects to Agents, June 15-17, 2026, Salerno, Italy

*Corresponding author.

marcello.pelosi2@unina.it (M. Pelosi); rajib.chandraghosh@unina.it (R. C. Ghosh); egidia.cirillo@unina.it (E. Cirillo); flora.amato@unina.it (F. Amato)

0009-0001-4718-4826 (M. Pelosi); 0009-0001-1137-3465 (R. C. Ghosh); 0009-0005-3227-6073 (E. Cirillo)

© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

published 2026-10-30

CEUR-WS.org/Vol-4307/paper12.pdf

Qwen 3.5 9B) employed as reasoning engines in ReAct agents, and we add a larger LLM baseline (Gemini 3.5 Flash) run under the *same* agentic architecture, in order to quantify the qualitative cost of the "small" choice. As a testbed we use a stratified sample of 75 questions (easy, medium, hard) drawn from HotpotQA [5], a challenging benchmark for open-domain multi-hop question answering. The agents are equipped with tool access to the Wikipedia API through custom-defined tools.

Given the complexity of generative evaluation in agentic scenarios, traditional metrics based on exact string matching are inadequate. We therefore adopt an LLM-as-a-judge framework, using DeepSeek V4-Flash as an external evaluator that classifies each generated answer as correct, partially correct, or incorrect, together with a textual rationale. To mitigate the known limitations of automatic judging, we additionally validate a subsample of the judgments manually. The rationales enable an in-depth, *complete* qualitative error analysis: rather than inspecting a handful of hand-picked cases, every non-correct answer is assigned to one of five failure categories through a second, automated LLM pass, yielding the full error-type distribution per model.

The central contributions of this work are: (i) a rigorous, reproducible empirical comparison of three open-weight SLMs plus an LLM baseline on the accuracy-latency trade-off, supported by confidence intervals and paired significance tests; (ii) a complete, automatically-produced and manually-validated taxonomy of failure modes that demonstrates how each model exhibits a unique error profile that cannot be deduced from overall accuracy alone.

2. Related Work

This study lies at the intersection of research on autonomous agents and on language-model efficiency. The development of tool-augmented agents has been profoundly influenced by the ReAct paradigm [6], which, by alternating cycles of reasoning, action, and observation, laid the foundations for modern agentic architectures whose effectiveness depends critically on the correct contextualisation of tools through the prompt [7, 8].

Although these architectures have historically been dominated by large LLMs, the growing need to operate in edge and on-premise scenarios has attracted strong interest toward Small Language Models. It is plausible that compact models may narrow the gap with LLMs in tool use through structured design patterns, and it is reasonable to expect that models under 10 billion parameters could sustain complex agentic tasks with acceptable accuracy, stability, and low power consumption, a hypothesis this work helps to assess. The drive toward on-device and privacy-preserving intelligence is part of a broader trend in distributed learning: federated approaches keep raw data on the client and are increasingly applied to relational and graph-structured domains such as smart healthcare [9], while recent work pushes toward single-round, communication-efficient protocols suited to resource-constrained, privacy-sensitive deployments [10]. SLM agents share the same deployment motivations, such as locality, low cost, and data minimisation, which is part of why their reliability under realistic conditions deserves close scrutiny. However, the ability of SLMs to sustain reasoning chains in open domains, where lexical disambiguation is critical, remains a challenge; access to external search tools is essential to compensate for the lower parametric knowledge of SLMs and to mitigate their propensity for hallucination during multi-hop reasoning. For this purpose we adopt HotpotQA, a benchmark that requires the integration of information from multiple documents.

Finally, to evaluate complex generative output, we rely on the LLM-as-a-judge approach. Despite known limitations such as position or verbosity biases, the most recent literature consolidates this practice as a methodologically mature one [11, 12], capable of providing reliable judgment taxonomies aligned with human evaluations for scalable system comparisons. We complement it with a manual validation step to quantify agreement.

3. Methodology

The experiment evaluates the reasoning capabilities of three Small Language Models-`gemma4:e4b` (4B parameters), `ministral-3:8b` (8B), and `qwen3.5:9b` (9B), in their default `Q4_K_M` quantisation (4-bit, mixed precision). As a larger reference point, we additionally run `gemini-3.5-flash` under the identical agentic setup. All models are configured with `temperature=0` to encourage deterministic outputs and improve reproducibility.

The models operate as ReAct agents and interact with Wikipedia through two tools: an exploratory *search*, which returns the top five titles and short snippets, and a targeted *read*, which extracts the introductory summary of a specific page. The tools query the MediaWiki REST API directly with an explicit user agent and a retry policy, and they validate the content type before parsing, so that transient empty or non-JSON responses are surfaced to the agent as actionable feedback rather than crashing the run. To prevent false lexical equivalences, the *read* tool disables silent redirects and reports an explicit "title mismatch" when the requested and returned titles differ, so the agent can search again instead of using the wrong content. Disambiguation pages and missing titles are likewise reported as diagnostic messages.

All models share the same maximum budget of ReAct iterations and the same termination criterion: the agent stops when it produces a final answer without new tool calls or, in the absence of convergence, upon reaching the budget (a recursion limit of 30 internal steps, roughly twelve full reasoning-action cycles), after which the run terminates without a final answer. All models receive the same system prompt, which constrains the agent to (i) always search before reading, (ii) verify the exact match between the article title and the subject of the question, (iii) add disambiguating terms (e.g. "film" or a year) when necessary, and (iv) prefer, in case of ambiguity, a new and more specific search rather than an uncertain answer.

3.1. Dataset and evaluation protocol

The evaluation uses a balanced sample of 75 questions, 25 for each of the three difficulty levels *annotated in the dataset itself* (*easy*, *medium*, *hard*). The sample is extracted from the `fullwiki` configuration of HotpotQA via a shuffle with a fixed seed (`seed=42`) to ensure a paired comparison: every model is evaluated on exactly the same 75 items, which is what later licenses the use of paired statistical tests. To support reproducibility, the exact HotpotQA question identifiers used in the sample and the full set of model outputs and judge rationales are available from the authors upon reasonable request.

In addition to end-to-end inference times, answer correctness is measured through an LLM-as-a-judge based on DeepSeek V4-Flash. The judge is configured to produce a structured JSON output that analyses the prediction against the reference answer, assigning a categorical judgment among `right`, `partially right`, or `wrong`, supported by a logical explanation referencing the supporting facts. From these judgments we derive a *strict* accuracy, counting only fully correct answers, and a *soft* accuracy, assigning a partial score of 0.5 to partially correct answers.

3.2. Hardware and execution environment

The SLM experiments were executed locally and sequentially on a workstation equipped with an Intel Core Ultra 9 275HX processor (P-cores up to 5.4 GHz), an NVIDIA GeForce RTX 5080 GPU with 16 GB of GDDR7 VRAM, and 32 GB of DDR5 RAM at 6400 MT/s. The software infrastructure relies on LangGraph and LangChain for managing the agent flows, and on Ollama for the local execution and serving of the SLMs. This configuration isolates computational performance from network latency for the local models. The Gemini baseline, being served via a cloud API, is therefore *not* comparable in latency to the local SLMs; for that model we report only accuracy and error-profile results, and we explicitly exclude its wall-clock times from the latency comparison.

Table 1
Distribution of evaluations and overall accuracy on the 75 HotpotQA questions, with 95% confidence intervals (Wilson for strict, bootstrap for soft).

Model	Corr.	Part.	Wrong	Strict	Soft	CI (soft)
Gemma 4 e4b	39	4	32	52.0%	54.7%	[43.3, 66.0]
Ministral-3 8B	30	9	36	40.0%	46.0%	[35.3, 56.7]
Qwen 3.5 9B	44	7	24	58.7%	63.3%	[52.7, 73.3]
Gemini 3.5 Flash (LLM)	55	3	17	73.3%	75.3%	[65.3, 84.7]

3.3. Manual validation of the judge

To corroborate the automatic evaluation, a random subsample of 20 question-answer pairs was manually re-annotated (5 per model, stratified across difficulty levels), applying the same three-class scheme used by the judge. The annotation was performed blind, i.e. without access to the judge's label, so as to keep the human assessment independent. Agreement between the human annotation and the DeepSeek judgment was high: 90.0% exact agreement (18 of 20 items) and Cohen's $\kappa = 0.83$. The two disagreements both occurred on the boundary between `right` and `partially right` (the human judging an answer fully correct where the judge was stricter), and never on the `right`/`wrong` distinction. These figures indicate that the automatic judge and the human annotator broadly agree, and that the accuracy rankings reported below are not an artefact of the automatic judging procedure.

4. Results

4.1. Overall accuracy

Table 1 reports the distribution of evaluations and the accuracy of the four models on the full sample of 75 questions, together with 95% confidence intervals; Figure 1 shows the assessment distribution for the three SLMs as stacked bars.

Among the SLMs, Qwen 3.5 9B is the most accurate (58.7% strict, 63.3% soft), Gemma 4 e4b is intermediate (52.0%, 54.7%), and Ministral-3 8B is the least accurate (40.0%, 46.0%). The LLM baseline, Gemini 3.5 Flash, is the most accurate overall (73.3%, 75.3%). The number of partially correct evaluations is small for every model.

Because the confidence intervals are wide at $n = 75$ and overlap substantially among the SLMs, these point estimates must be read with care. A Cochran's Q omnibus test on the paired binary (correct vs. not) outcomes rejects equality of the four models ($Q = 26.4$, $\mathrm{df} = 3$, $p < 10^{-4}$). Pairwise McNemar tests (exact, Holm-corrected for multiple comparisons) clarify *where* the differences lie: among the three SLMs, only the Ministral-Qwen contrast is significant ($p = 0.026$); Gemma-Ministral ($p = 0.22$) and Gemma-Qwen ($p = 0.33$) are not. The LLM baseline is significantly more accurate than Gemma ($p = 0.008$) and than Ministral ($p < 0.001$), but *not* significantly more accurate than the strongest SLM, Qwen ($p = 0.080$ after correction). In other words, at this sample size the SLMs are largely indistinguishable from one another in aggregate accuracy, and the best SLM closes the gap with the LLM baseline to the point where the difference is no longer statistically detectable.

4.2. Effect of difficulty level

Figure 2 shows the soft accuracy stratified by difficulty. All models degrade from easy to hard:

- Gemma 4 e4b: 72.0% (easy) → 48.0% (medium) → 44.0% (hard);
- Ministral-3 8B: 56.0% → 44.0% → 38.0%;
- Qwen 3.5 9B: 82.0% → 54.0% → 54.0%;
- Gemini 3.5 Flash: 84.0% → 78.0% → 64.0%.

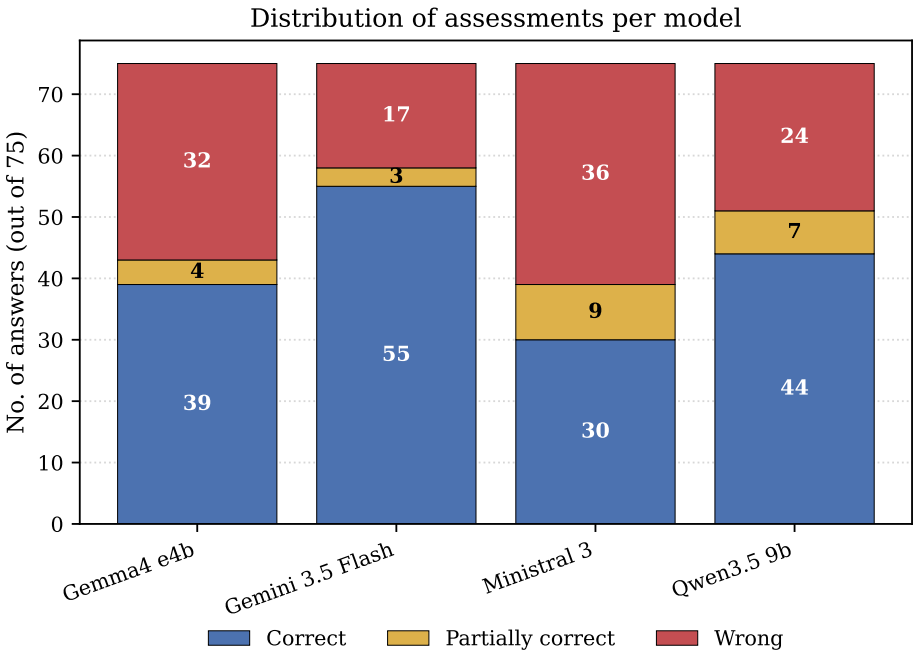

Figure 1: Distribution of evaluations per SLM on the 75 HotpotQA questions.

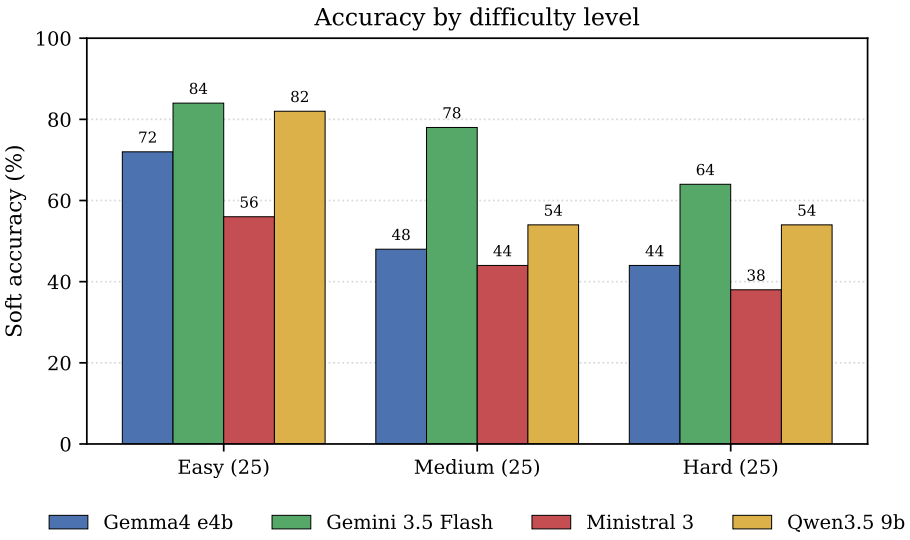

Figure 2: Soft accuracy by difficulty level (25 questions per level).

The most informative observation is the relative robustness of the LLM baseline at medium and hard difficulty: Gemini retains 78.0% at medium where every SLM has already fallen below 55%. Among the SLMs the ranking is preserved across levels, but the per-level confidence intervals ($n = 25$) are wide and overlap heavily, so the SLM differences at the hard level should be read as a consistent tendency rather than a demonstrated effect. The added value of the larger model thus manifests chiefly as resilience of multi-hop reasoning at medium-to-high difficulty, more than as a uniform accuracy gain.

4.3. Inference times and agentic metrics

Table 2 summarises the per-question inference-time statistics for the three locally-served SLMs; Figure 3 shows their full distribution. With the corrected, REST-based tool layer, none of the SLMs exhibits the

Table 2
Per-question inference-time statistics (seconds) for the locally-served SLMs.

Model	Mean	Median	Std. dev.	Min	Max
Gemma 4 e4b	16.55	13.84	9.19	1.79	42.57
Ministral-3 8B	6.64	4.70	5.36	0.86	30.43
Qwen 3.5 9B	18.63	13.16	14.55	4.42	69.62

Table 3
Mean agentic metrics per question for the locally-served SLMs. "n.c." = number of non-converged runs.

Model	Tool calls	Steps	Gen. tokens	n.c.
Gemma 4 e4b	2.96	7.92	924	0
Ministral-3 8B	2.51	6.45	189	0
Qwen 3.5 9B	4.31	9.87	733	0

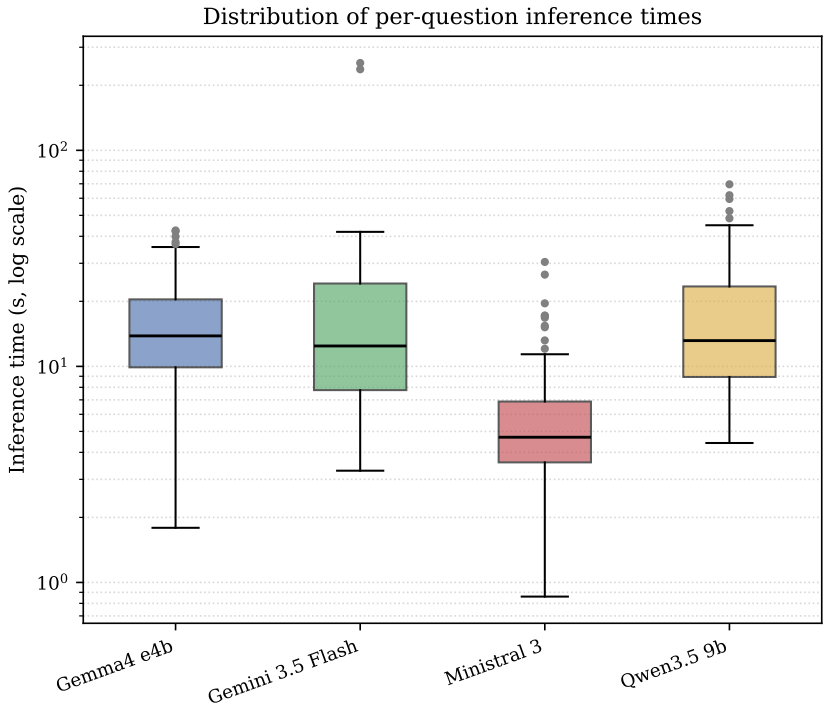

Figure 3: Boxplot of per-question inference times for the SLMs (log scale).

pathological non-convergence episodes: *no* run failed to terminate (column "n.c." in Table 3 is zero for all models), and the maximum per-question time stays below 70 s for both Qwen and Gemma and below 31 s for Ministral. Ministral-3 is clearly the fastest and most stable (mean 6.6 s, median 4.7 s), Gemma and Qwen are slower and more variable, with Qwen the heaviest of the three on average.

Table 3 reports the additional agentic metrics captured during the runs: mean number of tool calls, ReAct steps, and generated (completion) tokens per question. These quantify the qualitative behaviour behind the timings. Ministral generates by far the fewest tokens (189 on average vs. 733-924 for the others) and issues the fewest tool calls: it is fast because it "closes early". As the qualitative analysis will show, this same tendency to commit without sufficient exploration is consistent with its high rate of factual hallucination. Qwen, conversely, is the most exploratory (4.3 tool calls, 9.9 steps per question on average).

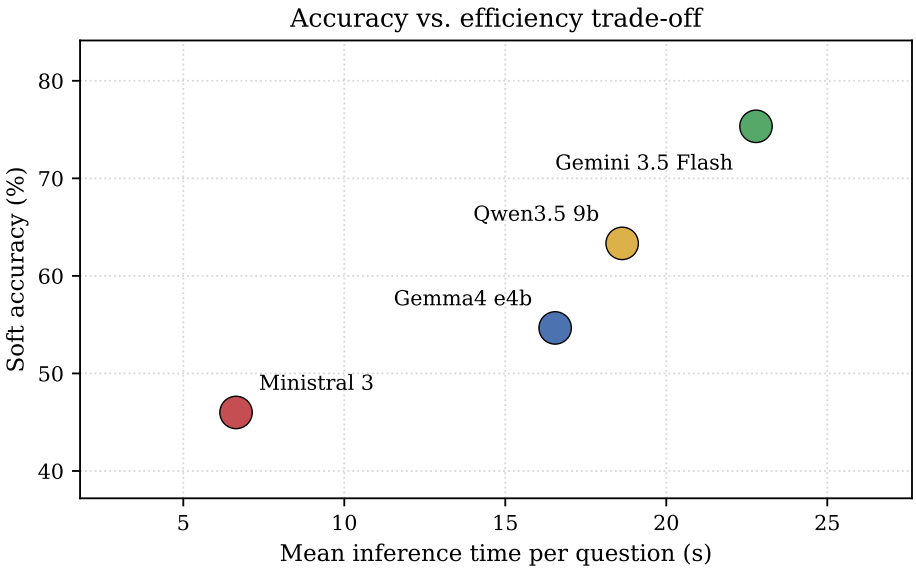

Figure 4: Trade-off between soft accuracy and mean inference time per question (SLMs).

4.4. Accuracy/efficiency trade-off

Figure 4 places the three SLMs on the (mean time, soft accuracy) plane. Qwen 3.5 9B leads on accuracy among the SLMs but at the highest mean latency; Ministral-3 8B is the most efficient but the least accurate; Gemma 4 `e4b` occupies an intermediate, weakly Pareto-dominated position. We stress that the Gemini baseline is not shown on this plane, since its cloud-served latency is not comparable with the locally-served SLMs.

4.5. Agreement among models

Out of the 75 questions, 18 are answered correctly by all four models and 10 are answered incorrectly by all of them; the remaining 47 discriminate between the models. Particularly informative is the number of *exclusive* errors, where one model fails while the other three succeed: Ministral fails alone 7 times, Gemma 4 times, and Qwen and Gemini once each. The larger count of exclusive errors for Ministral is consistent with its lower overall accuracy and, as discussed below, with its distinctive hallucination profile.

4.6. Manual-validation agreement

As reported in Section 3.3, the manual re-annotation of the 20-item subsample agrees with the automatic judge on 18 items ($\kappa = 0.83$). Crucially, no disagreement flipped a `right` into a `wrong` or vice versa: both disagreements lay on the `right`/`partially right` boundary. The model rankings are therefore stable whether one uses the automatic or the human labels, which supports the reliability of the conclusions that follow.

5. Qualitative discussion of errors

We now move from a quantitative to a qualitative analysis, examining the recurrent failure patterns in light of the judge rationales and of what they suggest about the underlying abilities of the models. Unlike a hand-picked inspection, the error-category assignment here is *complete*: every answer judged `wrong` or `partially right` was assigned to one of five categories by a second, automated LLM pass operating on the judge rationale, producing the full distribution of Table 4.

Table 4
Distribution of error categories per model over all non-correct answers. "NA" marks cases not clearly attributable to a single category.

Model	E1	E2	E3	E4	E5	NA	Total
Gemma 4 e4b	6	6	11	10	1	0	34
Ministral-3 8B	8	2	19	7	3	0	39
Qwen 3.5 9B	4	3	3	14	1	3	28
Gemini 3.5 Flash (LLM)	0	0	2	14	0	2	18

5.1. Recurrent error patterns

(E1) Incorrect Wikipedia page selection. The model identifies an entity with a similar but different name and reads the wrong page; the typical case of missed disambiguation and homonymous titles.

(E2) Multi-hop failure. The model retrieves the first piece of information correctly but does not make the second logical leap, returning an intermediate entity instead of the requested one.

(E3) Factual hallucination. The model produces a plausible answer that is not supported by the retrieved page.

(E4) Empty or evasive output. The agent terminates without a usable answer, returning an empty string or asking the user for clarification instead of extracting the answer from the context already available.

(E5) Arithmetic/comparative reasoning error. The model retrieves the facts correctly but fails a trivial comparison (e.g. who is younger given two birth dates), isolating a pure reasoning error independent of retrieval quality.

5.2. Comparison between models on critical cases

We illustrate each pattern with a representative case that also contrasts the models.

Case A - Incorrect page selection (E1). *Question* (medium): "*Sky High* starred the actress who is married to which actor?" *Correct answer:* John Travolta. The ground truth identifies the relevant actress in the 2005 film *Sky High* as Kelly Preston, married to John Travolta. Both Gemma and Qwen instead anchor on a different cast member, Mary Elizabeth Winstead, and read her page, answering with her spouse (Riley Stearns and Ewan McGregor, respectively); Ministral correctly retrieves Kelly Preston but then reads the wrong spouse, answering Bruce Willis. The judge notes in each case that the wrong entity (or the wrong page for the right entity) was used. The case shows how an early entity-selection error-picking the wrong person among several plausible candidates with similar associations-propagates through the whole retrieval chain.

Case B - Multi-hop failure (E2). *Question* (easy): "Which TV network founded in 1984 is co-owned by the Hearst Corporation and Disney?" *Correct answer:* Arts and Entertainment Network (A&E). Gemma and Ministral both answer with the parent company (*A+E Networks*/*A+E Global Media*) rather than the network itself, stopping at the intermediate entity; only Qwen completes the second hop to A&E. The first hop (the Hearst-Disney joint venture) is solved by all; the failure is purely on the second leap.

Case C — Factual hallucination (E3). *Question* (medium): "*Chairmen of the Bored* is an album that references a joke made by a cast member of what series?" *Correct answer:* Saturday Night Live. The context indicates that the album references a joke by Norm Macdonald, a cast member of *Saturday Night Live*. Ministral instead asserts, confidently and with a fabricated justification, that the joke comes from a character in *Scrubs*, inventing a play-on-words rationale that has no support in the retrieved evidence. Gemma retrieves the album's artist correctly but then claims the series cannot be identified, while Qwen abstains. The case is paradigmatic of (E3): the model fills a gap with a fluent, plausible-sounding answer that is simply not grounded in the page it read. This is the applicatively most dangerous pattern, because the answer looks right and is delivered with confidence; the concentration of E3 in Ministral (19 of its 39 errors) makes it the least reliable model where calibration matters.

Case D - Empty or evasive output (E4). *Question* (hard): "On what date was Ernest Davies' successor born?" *Correct answer:* 10 October 1940. Gemma claims it cannot find the information (which is in fact present in the context); Gemini terminates with "need more steps to process this request"; Qwen returns an empty string. The two most accurate models, Qwen and Gemini, tend to *abstain* rather than to fabricate: for both, the empty/evasive output (E4) is by far the dominant failure mode (14 of 28 and 14 of 18 errors, respectively), while hallucinations are rare. From an application standpoint this is safer than Ministral's behaviour, because an abstention is far easier to detect than a fluent hallucination. The high frequency of E4 in the best models, however, points to a limitation of the agentic runtime (iteration budget, non-convergence of the reasoning trace) more than of the models themselves.

Case E - Pure reasoning error (E5). *Question* (easy): "Who is younger, Stanley Kwan or Atom Egoyan?" *Correct answer:* Atom Egoyan. Ministral retrieves the exact birth dates (Kwan 1957, Egoyan 1960) but concludes that "Stanley Kwan is younger", reversing the comparison; Gemma claims the dates are unavailable (despite being in context); Qwen reasons correctly. Retrieval is perfect for Ministral, yet it fails the most elementary numerical comparison-isolating a pure reasoning error. Part of Ministral's weakness is thus attributable to base model capability, not to the agentic component.

5.3. Overall assessment

Integrating the quantitative and qualitative evidence yields three observations. First, errors are not uniformly distributed: the four models exhibit distinct error profiles (Table 4). Ministral-3 concentrates factual hallucinations (E3) and pure reasoning errors (E5), both independent of retrieval quality, which makes it the least reliable in applications requiring calibration guarantees. Qwen and the Gemini baseline fail predominantly by empty output (E4) rather than by hallucination-a safer profile. Gemma shows the most dispersed profile, consistent with its intermediate, "poor-compromise" position. The LLM baseline stands out for the near-absence of E1, E2 and E3 errors: its superior quality manifests as the avoidance of the disambiguation and multi-hop failures that afflict the SLMs.

Second, a non-negligible and cross-cutting portion of the errors is attributable not to the model but to the tool-use layer: empty/evasive output (E4) is the single most frequent pattern overall, and several of the questions failed by all models stem from retrieval limitations. This suggests that, with the same SLM, significant improvements are achievable by strengthening the system prompt and enriching the tools-e.g. with section-level search, full-text reading instead of only the summary, or an explicit target-entity verification mechanism.

Third, the comparison with the LLM baseline frames the qualitative cost of the "small" choice. While Gemini is significantly more accurate than the two weaker SLMs, the gap with the strongest SLM (Qwen) is not statistically significant after correction, and the baseline's advantage is concentrated on multi-hop robustness at medium and high difficulty. Difficulty level thus appears to be a more informative discriminator than global accuracy for choosing a suitable model for a given application context.

6. Conclusions and Future Work

We presented an empirical comparison of three Small Language Models (Gemma 4 `e4b`, Ministral-3 8B, Qwen 3.5 9B) and a larger LLM baseline (Gemini 3.5 Flash) as reasoning engines for a multi-hop question-answering agent on HotpotQA with Wikipedia tool access. The evaluation, conducted via an external LLM judge, corroborated by a manual validation subsample, and enriched by a complete qualitative analysis of the rationales, shows that under identical tools and prompts the models exhibit distinct accuracy, latency, and-most importantly-error profiles. Qwen leads among the SLMs and closes the gap with the LLM baseline to the point where the difference is no longer statistically detectable at this sample size; Ministral is the fastest but least reliable, owing to hallucinations and pure reasoning errors; Gemma is an intermediate, weakly dominated compromise. At this sample size most SLM-vs-SLM accuracy differences are not statistically significant, which makes the qualitative error profiles the more discriminative finding.

Several limitations should be borne in mind. The sample, although stratified, is limited to 75 questions, so the quantitative evidence is indicative rather than definitive, and the confidence intervals are correspondingly wide. The judging procedure relies on a single external LLM; while our manual-validation subsample shows high agreement, a larger human-annotated set would further quantify the judge's bias. The ReAct architecture is a strong simplification, restricted to two tools that search and read concise Wikipedia summaries; richer mechanisms (full-text or section-level access, multiple sources) could substantially change performance, especially on the questions failed by all models. Results refer to the `Q4_K_M`-quantised SLM variants; full-precision or differently-quantised weights may differ. Inference times, measured end-to-end for the local SLMs only, include system overheads; with the corrected tool layer no non-convergence outliers were observed, and the median remains the most reliable indicator of typical latency. The Gemini baseline is cloud-served and is therefore excluded from the latency comparison. Finally, despite fixed seeds, zero temperature, and locally-executed open models, marginal hardware and runtime non-determinism persists, though it does not affect the robustness of the observed rankings.

As future work we plan: (i) extending the sample and the human validation of the judgments to better quantify the judge's bias; (ii) enriching the tool set with full-page reading, section search, and target-entity identity verification; (iii) studying the impact of self-reflection and tool-augmented chain-of-thought on these same SLMs; (iv) broadening the baseline comparison to additional large models under the same agentic architecture; and (v) extending the analysis to multiple seeds and non-zero temperatures, to separate systematic from stochastic contributions, along with finer agentic metrics.

Declaration on Generative AI

During the preparation of this work, the authors used a generative AI assistant for drafting and code support. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] A. Plaat, M. van Duijn, N. van Stein, M. Preuss, P. van der Putten, K. Batenburg, Agentic large language models, a survey, Journal of Artificial Intelligence Research 84 (2025). doi:10.1613/jair.1.18675.

[2] Z. Lu, X. Li, D. Cai, R. Yi, F. Liu, W. Liu, J. Luan, X. Zhang, N. D. Lane, M. Xu, Demystifying small language models for edge deployment, in: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Vienna, Austria, 2025, pp. 14747–14764.

[3] J. Zhang, P. Wu, J. London, D. Tenney, Benchmarking and evaluating large language models in

phishing detection for small and midsize enterprises: A comprehensive analysis, IEEE Access 13 (2025) 28335–28352. doi:10.1109/ACCESS.2025.3540075.

[4] D. Annunziata, P. Qi, K. Li, F. Piccialli, Towards trustworthy IIoT: A reinforcement learning framework with federated unlearning, IEEE Internet of Things Journal (2026).

[5] Z. Yang, P. Qi, S. Zhang, Y. Bengio, W. Cohen, R. Salakhutdinov, C. D. Manning, HotpotQA: A dataset for diverse, explainable multi-hop question answering, in: E. Riloff, D. Chiang, J. Hockenmaier, J. Tsujii (Eds.), Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 2369–2380. URL: https://aclanthology.org/D18-1259/. doi:10.18653/v1/D18-1259.

[6] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, Y. Cao, ReAct: Synergizing reasoning and acting in language models, in: The Eleventh International Conference on Learning Representations (ICLR), Kigali, Rwanda, 2023.

[7] G. Mialon, R. Dessì, M. Lomeli, C. Nalmpantis, R. Pasunuru, R. Raileanu, B. Rozière, T. Schick, J. Dwivedi-Yu, A. Celikyilmaz, E. Grave, Y. LeCun, T. Scialom, Augmented language models: a survey, Transactions on Machine Learning Research (TMLR) (2023).

[8] T. Schick, J. Dwivedi-Yu, R. Dessì, R. Raileanu, M. Lomeli, E. Hambro, L. Zettlemoyer, N. Cancedda, T. Scialom, Toolformer: Language models can teach themselves to use tools, in: Advances in Neural Information Processing Systems 36 (NeurIPS 2023), New Orleans, LA, USA, 2023.

[9] S. M. A. Huda, M. Canzaniello, D. Annunziata, F. Piccialli, Graph-based federated learning for smart healthcare: A comprehensive survey, Future Generation Computer Systems (2026) 108524.

[10] F. Amato, L. Qiu, M. Tanveer, S. Cuomo, D. Annunziata, F. Giampaolo, F. Piccialli, Towards one-shot federated learning: Advances, challenges, and future directions, Neurocomputing (2025) 132088.

[11] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, I. Stoica, Judging LLM-as-a-judge with MT-Bench and chatbot arena, in: Advances in Neural Information Processing Systems 36 (NeurIPS 2023), Curran Associates, Inc., New Orleans, LA, USA, 2023, pp. 46595–46623.

[12] D. Li, B. Jiang, L. Huang, A. Beigi, C. Zhao, Z. Tan, A. Bhattacharjee, Y. Jiang, C. Chen, T. Wu, K. Shu, L. Cheng, H. Liu, From generation to judgment: Opportunities and challenges of LLM-as-a-judge, in: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Suzhou, China, 2025, pp. 2757–2791. doi:10.18653/v1/2025.emnlp-main.138.