Better Retrieval, Worse Robustness: How Multi-hop RAG Amplifies Upstream ASR Errors

Anonymous ACL submission

Abstract

Speech-based applications pass spoken queries through automatic speech recognition (ASR) before any retrieval module, so ASR errors enter the pipeline as a fixed upstream constraint. We empirically test whether two extensions to standard retrieval-augmented generation (RAG), entity-graph linking and iterative reformulation, absorb or amplify these errors. Using four English accents synthesized through neural TTS, we evaluate four RAG configurations on three multi-hop QA benchmarks (HotpotQA, 2WikiMultiHopQA and MuSiQue) against a clean-text oracle. Both extensions amplify the error: the F1 gap from clean text to the highest-WER accent is 36–67% larger under their combination than under naive dense retrieval, on all three benchmarks. The dominant failure mode is corruption of one or more query entities, accounting for 87–94% of degradation cases on 2WikiMultiHopQA. Two lightweight surface-form mitigations leave most of the gap intact, indicating that downstream retrieval structure amplifies remaining entity errors. We release code and data at https://anonymous.4open.science/r/anonymous-repo-arr2026-90C6/.

1 Introduction

Voice has become a primary communication channel in everyday applications. Whether interacting with a voice assistant or multimodal agent, the user's spoken query is passed through an automatic speech recognition (ASR) module that transcribes it into text. However, speakers differ within a language: accents vary widely, and ASR models do not transcribe all variants equally well. State-of-the-art systems achieve single-digit word error rates (WER) on clean read-speech benchmarks (Radford et al., 2023), but their accuracy degrades sharply on accented input. Koenecke et al. (2020) report that ASR systems produce nearly 2× higher WER for African American speakers than for white speakers across five commercial systems, and Markl (2022) reports similar disparities on Google and Amazon ASR for stigmatized regional varieties of British and Irish English and for second-language speakers. Figure 1 previews this in our setup: the oracle–NG F1 gap on HotpotQA grows from 0.104 under Naive RAG to 0.142 under IRCoT+HippoRAG2, a 36.5% larger drop, though the more complex method achieves higher F1 on clean text.

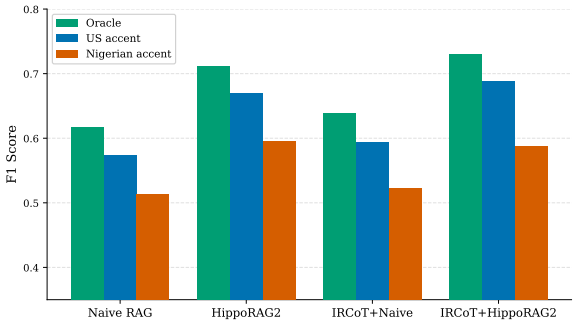

Figure 1: HotpotQA F1 under oracle (clean text), US-accented (WER 9.4%), and Nigerian-accented (WER 14.5%) speech. The oracle–NG gap grows from 0.104 for the simplest baseline (Naive RAG) to 0.142 for the most structurally complex configuration (IRCoT+HippoRAG2).

The downstream consequences of ASR errors have been studied for machine translation (Ruiz and Federico, 2014), spoken language understanding (Ruan et al., 2020), and reading comprehension (Li et al., 2018). Yet the impact of ASR errors on *retrieval-augmented generation* (RAG) has received little attention. In its standard form, RAG retrieves passages from an external corpus using the input query and conditions a generator on the retrieved evidence (Lewis et al., 2020). This single-step formulation is already sensitive to query corruption, since retrieval quality depends directly on the surface form of the query. The problem is compounded in multi-hop QA, which requires retrieving and reasoning over evidence across mul-

tiple documents (Yang et al., 2018; Ho et al., 2020; Trivedi et al., 2022). Recent multi-hop RAG methods extend the standard pipeline with entity-graph linking (Gutiérrez et al., 2025) or iterative reformulation (Trivedi et al., 2023), each adding a new point at which upstream ASR errors can propagate. It remains underexplored whether these methods are more or less robust to accented input than Naive RAG, and through which mechanisms they fail.

To analyze whether upstream ASR errors propagate through multi-hop RAG, we ask four research questions. **RQ1:** How much does accented-speech ASR error degrade multi-hop QA performance, and does it scale with WER? **RQ2:** Do structurally complex multi-hop RAG architectures absorb or amplify these errors? **RQ3:** What is the dominant failure mechanism by which ASR errors propagate through structurally complex RAG? **RQ4:** Do lightweight surface-form mitigations close the gap, and what does the remaining gap reveal about how the failure arises? To answer these, we evaluate four representative RAG methods on three multi-hop QA benchmarks (HotpotQA, 2WikiMultiHopQA, MuSiQue), using TTS-synthesized spoken queries across four English accents.

Our experiments reveal that ASR errors propagate through multi-hop RAG, that structurally complex retrieval methods amplify rather than absorb these errors, and that entity corruption is the dominant mechanism. Our main contributions are:

- We construct a spoken multi-hop QA evaluation suite spanning three benchmarks, four English accents, and four RAG methods, totaling 12,000 spoken queries[1].
- We show that structurally complex retrieval methods (entity-graph linking, iterative reformulation) amplify upstream ASR errors despite higher peak F1 on clean text.
- We identify query-entity corruption as the dominant failure mode (87–94% on 2WikiMultiHopQA, 67–82% on HotpotQA, 59–78% on MuSiQue).
- Two surface-form mitigations (N-best decoding, phonetic correction) serve as diagnostic probes. Their limited recovery indicates downstream retrieval structure amplifies remaining entity errors. We additionally validate against 500 real Nigerian utterances, supporting TTS as a controlled probe of the mechanism.

[1]We release the source code, Whisper-large-v3 transcripts, and per-cell result summaries at https://anonymous.4open.science/r/anonymous-repo-arr2026-90C6/. The full TTS-synthesized audio (∼ 360 MB) will be released upon acceptance.

2 Related Work

Bias and propagation of ASR errors. ASR disparities are well studied. Tatman (2017) quantified gender and dialect bias in YouTube auto-captions across two genders and five dialect groups. Koenecke et al. (2020) reported nearly 2× higher WER for African American speakers than white speakers across five commercial systems. The same pattern appears across regional varieties of British and Irish English and second-language speakers (Markl, 2022), and across nine global English accents (DiChristofano et al., 2022). ASR errors also propagate into downstream NLP tasks. Ruiz and Federico (2014) showed that MT systems are particularly sensitive to ASR errors. Ruan et al. (2020) demonstrated that intent classification and slot filling F1 drop sharply when entity-bearing tokens are corrupted by ASR, suggesting that downstream tasks relying on specific lexical items are particularly vulnerable. Li et al. (2018) extended this observation to reading comprehension, showing that ASR errors substantially degrade SQuAD F1. We focus on the propagation mechanism rather than the upstream disparity: how a fixed ASR error pattern propagates through multi-hop retrieval architectures, where inter-document inference creates additional failure points.

Multi-hop retrieval-augmented generation. Multi-hop QA benchmarks (Yang et al., 2018; Ho et al., 2020; Trivedi et al., 2022) require chaining evidence across multiple documents and pose distinct challenges from single-hop open-domain QA. Naive dense retrieval (Karpukhin et al., 2020) provides a strong baseline but struggles when the bridging entity connecting sub-questions is not explicitly present in the query. To address this, HippoRAG (Gutiérrez et al., 2024) extracts entity-relation triples from the corpus, builds a knowledge graph linking documents through shared entities, and retrieves via personalized PageRank seeded from query entities. HippoRAG2 (Gutiérrez et al., 2025) extends this design by additionally retrieving propositional facts and combining the fact-level dense-retrieval signal with the PageRank score. In a separate design

direction, IRCoT (Trivedi et al., 2023) interleaves retrieval with chain-of-thought reasoning, issuing a new query at each reasoning step rather than relying on a single retrieval pass. Both designs introduce new dependencies on the surface form of the query. Sciavolino et al. (2021) study this fragility in clean text, showing that dense retrieval fails on entity-centric questions, with rare entities producing the largest performance drops. ASR corruption induces a similar effect that compounds across hops in multi-hop methods.

Spoken question answering. Prior work on spoken QA has focused on single-hop reading comprehension (Li et al., 2018; Lee et al., 2018) and conversational QA (You et al., 2020), where mitigation operates within a single retrieve-and-read pass per turn. Faisal et al. (2021) introduced SD-QA, a dialectal spoken QA benchmark with real recordings across five English varieties, and report that dialect-induced ASR errors degrade single-hop QA unevenly across speaker groups. We extend this line of research to multi-hop reasoning: how the propagation mechanism changes when the downstream system is multi-hop retrieval rather than single-hop reading comprehension. More recent work has built benchmarks for evaluating large speech-language models on single-hop spoken QA (Wang et al., 2025; Chen et al., 2026), but multi-hop reasoning over a retrieved corpus is not tested, and retrieval architectures are rarely examined. To our knowledge, no prior work systematically characterizes accent robustness in multi-hop RAG or isolates the role of entity corruption in this setting.

3 Pipeline

Figure 2 shows our evaluation pipeline. The pipeline supports configurable combinations of accents, retrieval methods, and mitigation approaches, while holding the question content, retrieval corpus, and generator fixed.

3.1 Speech Synthesis and Transcription

Each sampled question is synthesized into speech using four accented English voices from Microsoft Edge TTS[2]: `en-US-JennyNeural` (US), `en-IN-NeerjaNeural` (Indian), `en-PH-RosaNeural` (Filipino), and `en-NG-EzinneNeural` (Nigerian). All four voices are female and share comparable speaking style and audio quality, which reduces – though it does not eliminate – speaker, prosody, and channel confounds. The synthesized speech is transcribed with Whisper-large-v3 (Radford et al., 2023) using greedy decoding, yielding four transcribed queries per question that share the same source content but may differ in surface form due to ASR error. In addition to these four accented conditions, we evaluate an oracle condition that bypasses the speech pipeline and feeds the original clean text directly to retrieval; the oracle provides a per-method upper bound that isolates the degradation from ASR error. We validate the TTS-synthesized accents against real accented speech in Section 5.5.

3.2 RAG Methods

We evaluate four RAG methods. All methods use `gpt-4o-mini`[3] as the generator with temperature 0 for reproducibility.

Naive RAG. Top-k dense retrieval over a passage index built with `text-embedding-3-small`[4]; retrieved passages are concatenated as context for generation. We use $k = 10$, with documents chunked at 400 tokens and 80-token overlap. This serves as the simplest baseline with no additional graph or iterative dependencies on the query.

HippoRAG2 (Gutiérrez et al., 2025). Extracts entity-relation triples and propositional facts from the corpus, builds a knowledge graph linking documents through shared entities, and retrieves by combining personalized PageRank seeded from query entities with a fact-level dense retrieval pass. We use the official implementation with damping factor 0.5 and `gpt-4o-mini` for triple and fact extraction.

IRCoT+Naive (Trivedi et al., 2023). Wraps Naive RAG in a chain-of-thought loop. At each step, the LLM emits a reasoning sentence and a follow-up search query; the new query retrieves additional context that conditions the next step. The loop terminates when the LLM emits `DONE` or after 3 steps. Retrieved chunks are deduplicated across steps and capped at 3,000 tokens.

IRCoT+HippoRAG2. The same iterative loop, with HippoRAG2 as the underlying retriever. This configuration combines graph-based linking with iterative reformulation and achieves the highest F1 on clean text in our experiments.

[2] https://learn.microsoft.com/en-us/azure/ai-services/speech-service/

[3] https://platform.openai.com/docs/models/gpt-4o-mini

[4] https://platform.openai.com/docs/guides/embeddings

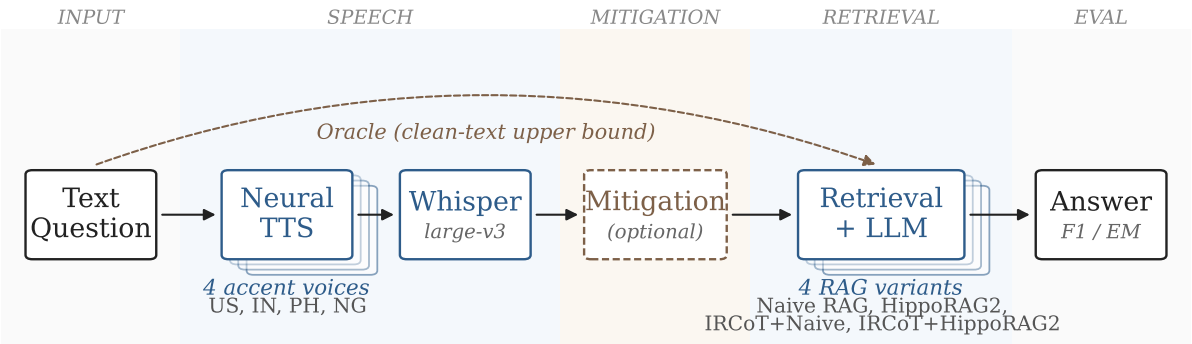

Figure 2: Overview of our spoken multi-hop QA evaluation pipeline. Each text question is synthesized in US, Indian, Filipino, and Nigerian accents through neural TTS, then transcribed by Whisper-large-v3. The transcription optionally passes through one lightweight mitigation module: N-best decoding or phonetic entity correction. It is then processed by one of the four RAG methods. The oracle condition (brown dashed arc) bypasses the speech pipeline and feeds the original text directly to retrieval, providing a clean-text upper bound.

3.3 Mitigation Strategies

We evaluate two lightweight mitigations applied to the ASR transcription before retrieval. Both are motivated by the error analysis in Section 5.3, which attributes most failures to query-entity corruption. We frame these mitigations as diagnostic probes: each targets a surface-level form of entity error, so the remaining gap characterizes what simple corrections cannot recover. Both are inference-time only, require no retraining, and compose with any retrieval method. Full algorithmic details, hyperparameters, and ablations are in Appendix A.

N-best Decoding. For each spoken query we decode Whisper at five temperatures with internal beam search, deduplicate hypotheses by text, retain the top-N (N=5) by mean log probability, and pass the union of their retrieved documents to the generator (Williams, 2008; Chia et al., 2008). The intuition is that if the greedy hypothesis corrupts a critical entity, a higher-temperature sample may recover the correct form in another decoding path.

Phonetic Entity Correction. We replace ASR-extracted entities with their phonetically nearest matches from a precomputed corpus entity index, keyed by Double Metaphone codes and reranked by edit distance (Levenshtein et al., 1966). The intuition is that ASR confusions preserve phonetic structure, so the correct surface form is often recoverable by phonetic neighborhood search even when exact string match fails.

4 Experimental Setup

4.1 Datasets

We evaluate on three widely used multi-hop QA benchmarks. **HotpotQA** (Yang et al., 2018) contains bridge and comparison questions over Wikipedia. **2WikiMultiHopQA** (Ho et al., 2020) adds compositional and inference questions with explicit reasoning chains. **MuSiQue** (Trivedi et al., 2022) spans 2- to 4-hop reasoning composed from single-hop primitives and is regarded as the hardest of the three. Following prior work (Press et al., 2023; Trivedi et al., 2023; Gutiérrez et al., 2025), we sample 1,000 questions from each validation set uniformly at random and use the associated supporting and distractor passages as the retrieval corpus for each benchmark.

To verify that our synthesized speech produces ASR error patterns comparable to real English speech, we additionally transcribe the first 500 utterances from a publicly available Nigerian-accented English corpus with the same Whisper-large-v3 system. This set is used only for ASR error pattern validation, since the utterances are general-purpose speech rather than multi-hop QA questions. Results are reported in Section 5.5.

4.2 Implementation Details

Whisper-large-v3 inference runs on a server with an `L40S` GPU, while all other components run on a local machine. We use FAISS (Douze et al., 2025) with cosine similarity for the dense index. Embedding generation (`text-embedding-3-small`) and answer generation (`gpt-4o-mini`) use the OpenAI API. For NER we use spaCy's `en_core_web_sm` model. Random sampling uses a fixed seed of 42.

4.3 Evaluation Metrics

Following standard practice on multi-hop QA, we report token-level F1 and Exact Match (EM) between the generated answer and the ground-

truth answer after SQuAD-style normalization (Rajpurkar et al., 2016). F1 and EM are averaged across questions within each dataset-accent cell. For ASR quality we report Word Error Rate (WER) between the transcription and the original question text, computed with Levenshtein-based alignment after lowercasing and whitespace tokenization. Per-accent F1 gaps are reported as $\Delta\text{F1} = \text{F1}_{\text{oracle}} - \text{F1}_{\text{accent}}$. We use absolute rather than relative differences because oracle F1 varies substantially across datasets. Statistical significance of cross-condition F1 and EM differences is assessed via paired bootstrap with 10,000 resamples.

5 Results and Analysis

We organize the section by research question. We find that (i) ASR errors degrade multi-hop QA performance, with degradation scaling with WER; (ii) structurally complex retrieval methods amplify rather than absorb these errors; (iii) query-entity corruption is the dominant failure mode; and (iv) lightweight surface-form mitigations close only a small fraction of the gap.

5.1 ASR Error Degrades Multi-hop QA and Scales with WER (RQ1)

Table 1 reports F1 across the four accent conditions and three benchmarks. Across every (dataset, method) cell, accented speech reduces F1 relative to the oracle, and Nigerian-accented input, which carries the highest WER on all three benchmarks, produces the largest drop. The oracle–NG gap ranges from 0.043 (Naive RAG on MuSiQue) to 0.195 (IRCoT+HippoRAG2 on 2WikiMultiHopQA), and is statistically significant at $p < 0.001$ across all cells. Relative to oracle F1, these gaps correspond to 15–34% degradation, with the largest relative drops on 2WikiMultiHopQA, consistent with its entity-dense bridge and compositional question format, a pattern we return to in RQ3. Exact Match results follow the same pattern: EM gaps track F1 gaps tightly (Pearson $r > 0.95$ across all 48 cells) and exhibit the same method-level amplification ordering on all three datasets (full EM numbers in Appendix C). This answers the first half of RQ1: ASR error materially degrades multi-hop QA performance.

The gap also tracks WER. Across accents, WER is lowest for US and highest for NG on all three benchmarks, and the oracle–accent F1 gaps in Table 1 follow the same ordering in nearly every cell. Across all cells, mean WER and the oracle–accent F1 gap correlate at Pearson $r = 0.88$, confirming that higher upstream WER corresponds to larger downstream degradation. WER therefore accounts for much of the cross-cell variance; what aggregate WER misses is which tokens get corrupted, a question we examine in RQ3. For a finer-grained view, Figure 3 stratifies questions in the Nigerian-accented 2WikiMultiHopQA condition by individual WER. Per-question degradation rises monotonically with WER for all four methods: questions under 5% WER fail at rates close to the oracle baseline, while questions above 20% WER fail in 25.2% (Naive RAG) to 37.8% (IRCoT+HippoRAG2) of cases. This high WER bin is not a long tail: it contains $n = 413$ questions (41% of the NG pool), the largest of the four WER bins. Taken with the cross-accent ordering, this establishes the second half of RQ1: degradation scales with WER across accents and questions.

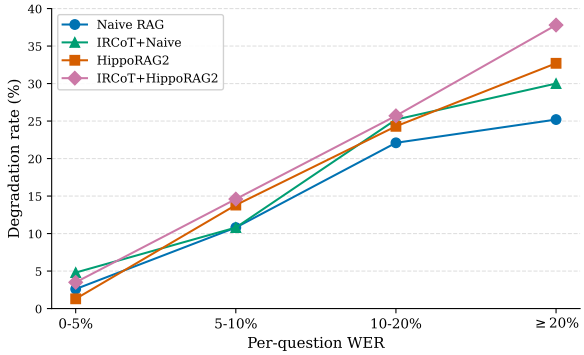

Figure 3: Per-question degradation rate on 2WikiMultiHopQA under Nigerian-accented speech, stratified by Whisper WER. The gap between Naive RAG and IRCoT+HippoRAG2 reaches 12.6% at $\geq 20\%$ WER.

5.2 Structural Complexity Amplifies Rather Than Absorbs ASR Errors (RQ2)

The larger absolute gaps in Table 1 align with the added structural extensions. We report absolute F1 differences for direct cross-method comparison within a dataset, and additionally report relative gap increases to control for the higher oracle ceiling of the more complex methods. Adding iterative reformulation alone (IRCoT+Naive vs. Naive RAG) widens the oracle–NG F1 gap by 10.6% on HotpotQA (0.104 → 0.115), 29.2% on 2WikiMultiHopQA (0.137 → 0.177), and 46.5% on MuSiQue (0.043 → 0.063). Adding entity-graph linking alone (HippoRAG2 vs. Naive RAG) widens it by 11.5%, 18.2%, and 58.1% respectively. The rela-

Dataset	Method	Oracle	US	IN	PH	NG	Gap
HotpotQA	WER	—	9.4%	11.0%	10.6%	14.5%	—
	Naive RAG	0.617	0.574	0.558	0.561	0.513	0.104
	IRCoT+Naive	0.638	0.594	0.573	0.583	0.523	0.115
	HippoRAG2	0.711	0.670	0.645	0.652	0.595	0.116
	IRCoT+HippoRAG2	0.730	0.688	0.652	0.671	0.588	**0.142**
2WikiMultiHopQA	WER	—	13.6%	14.1%	13.8%	17.1%	—
	Naive RAG	0.468	0.399	0.378	0.393	0.331	0.137
	IRCoT+Naive	0.516	0.399	0.406	0.415	0.339	0.177
	HippoRAG2	0.583	0.501	0.510	0.494	0.421	0.162
	IRCoT+HippoRAG2	0.645	0.562	0.534	0.536	0.450	**0.195**
MuSiQue	WER	—	5.2%	5.5%	5.7%	7.9%	—
	Naive RAG	0.282	0.264	0.256	0.261	0.239	0.043
	IRCoT+Naive	0.306	0.267	0.260	0.264	0.243	0.063
	HippoRAG2	0.355	0.326	0.324	0.321	0.287	0.068
	IRCoT+HippoRAG2	0.381	0.339	0.344	0.346	0.309	**0.072**

Table 1: F1 across three multi-hop QA datasets under oracle (clean text), US-accented, Indian-accented (IN), Filipino-accented (PH), and Nigerian-accented (NG) speech. WER per accent is reported per dataset. Gap (shaded) is the oracle–Nigerian F1 difference. Bold marks the largest gap per dataset; underline marks IRCoT+Naive to isolate the contribution of iterative reformulation. All oracle–Nigerian gaps are significant at $p < 0.001$.

tive ordering between these two single-extension configurations varies by dataset, but combining both yields the largest gap on every dataset (36.5%, 42.3%, and 67.4% respectively). The combined IRCoT+HippoRAG2 amplification is statistically significant on all three benchmarks ($p \leq 0.007$).

The combined configuration (IRCoT+HippoRAG2) achieves the highest oracle F1 on all three benchmarks (0.730 HotpotQA, 0.645 2WikiMultiHopQA, 0.381 MuSiQue) yet produces the largest oracle–NG gap, larger by 36.5% on HotpotQA (0.104 → 0.142), 42.3% on 2WikiMultiHopQA (0.137 → 0.195), and 67.4% on MuSiQue (0.043 → 0.072) than Naive RAG. The configuration that performs best on clean input shows the largest drop under ASR noise. Figure 3 (RQ1) previewed this at the per-question level: as WER increases, per-question failure rates grow faster for graph-based and iterative methods than for Naive RAG, showing the same pattern.

To understand which iterative steps contribute most to the amplification, we vary the IRCoT step count on HotpotQA and report the oracle–NG gap at each step (Figure 4). Step 0 corresponds to the baselines without IRCoT. The pattern is non-monotone. At step 1, the gap closes slightly on Naive RAG (0.104 → 0.096) and has little impact on HippoRAG2 (0.116 → 0.117). A single iterative round therefore does not amplify the error. For Naive RAG it slightly helps: retrieved context can surface the correct entity, which the

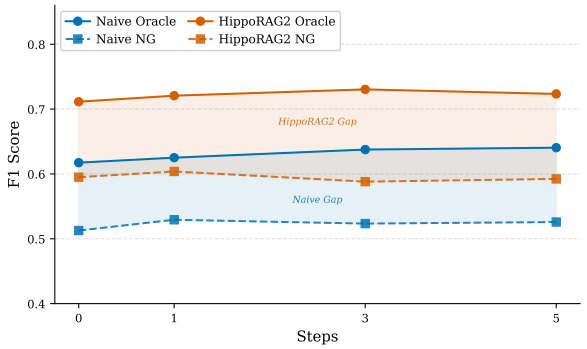

Figure 4: F1 of Naive RAG and HippoRAG2 across IRCoT steps on 1,000 HotpotQA questions. Solid lines show oracle F1, dashed lines show Nigerian-accented (NG) F1, and shaded regions show the oracle–NG gap.

LLM uses in its follow-up query. At step 3, the gap reaches 0.115 on Naive RAG and peaks at 0.142 on HippoRAG2. Beyond step 3, the two retrievers diverge. Naive RAG saturates, with the gap holding at 0.115 at step 5 and oracle and NG F1 both drifting up by 0.003. HippoRAG2's gap narrows to 0.131, but this reflects degradation rather than recovery: oracle F1 drops by 0.007 from step 3 to step 5 while NG rises by only 0.004, meaning the upper bound falls toward the corrupted condition rather than the other way around. Even after this partial closure, HippoRAG2's step-5 gap (0.131) remains 14% larger than Naive RAG's (0.115): the structural amplification persists.

This answers RQ2: structural extensions amplify rather than absorb ASR error, and the iterative

Error type	US	IN	PH	NG
Entity corruption	90%	87%	91%	94%
Severe garbling	30%	35%	37%	40%
Number/date corrupt.	4%	4%	4%	4%
Function-word noise	1%	0%	0%	0%
Other content change	7%	7%	5%	3%
# degradation cases	113	124	116	174

Table 2: Distribution of error types within degradation cases on 2WikiMultiHopQA under Naive RAG. Cases can carry multiple labels.

component contributes through cumulative reformulation steps over successive retrieval rounds.

5.3 Entity Corruption is the Dominant Failure Mechanism (RQ3)

Error categorization is automated and rule-based. For each question we compute a word-level difference between the original text and the ASR transcription using Python's `difflib.SequenceMatcher`[5]. We classify each non-equal operation by token type: a capitalized multi-word span (entity proxy), a digit token (number proxy), or a function word from a fixed list. Per-question labels (e.g., *entity corruption*, *severe garbling*) are assigned based on these per-operation counts and the question's overall WER. Labels are not mutually exclusive. Rules, thresholds, and the function-word list are in Appendix B.

Table 2 reports the distribution on 2WikiMultiHopQA under Naive RAG. Entity corruption is the dominant category for every accent (87%–94% of cases), and severe garbling appears in 30%–40%. Numerical, function-word, and miscellaneous edits are individually small ($\leq$ 7%). The Nigerian accent has the highest rates in both dominant categories (94% entity corruption, 40% severe garbling): at higher WER, only the frequency of the dominant mechanism rises, not the mechanism itself. On HotpotQA and MuSiQue the entity-corruption share is lower but remains the largest category, at 67%–82% across accents on HotpotQA and 59%–78% on MuSiQue (Appendix D). The effect is strongest on 2WikiMultiHopQA, where bridge and comparison questions chain multiple named entities by design.

Figure 5 illustrates how a single entity corruption manifests across the four RAG methods. A

[5] https://docs.python.org/3/library/difflib.html

Q: Which film has the director born earlier, **Hanuman Patal Vijay** or Young And Dangerous: The Prequel?
US: ... **Hanuman Patil Vijay** or Young and ...
NG: ... **Honourable Peter Vijay** or Young and ...

Method	Oracle	US	NG
Naive RAG	1.00	1.00	**0.00**
HippoRAG2	1.00	1.00	0.50
IRCoT+Naive	1.00	*0.67*	**0.00**
IRCoT+HippoRAG2	1.00	*0.67*	**0.00**

Figure 5: Case study on 2WikiMultiHopQA illustrating how a single entity corruption (*Hanuman Patal Vijay*) affects all four RAG methods. Cell values are F1 scores.

2WikiMultiHopQA query about the bridging entity *Hanuman Patal Vijay* is mildly corrupted under the US accent (*Hanuman Patil Vijay*) and severely corrupted under the NG accent (*Honourable Peter Vijay*). Under NG, Naive RAG fails entirely: the surrounding tokens (*film*, *director born earlier*, *Young and Dangerous*) dominate the dense retrieval signal, and the generator answers with the comparison film. HippoRAG2 partially recovers: the uncorrupted *Vijay* token still anchors the entity graph to the correct region, and the generator produces *Vijay* for partial credit. The two IRCoT variants fail completely. IRCoT+Naive offers no recovery, since the underlying retriever has no signal for the iterative loop to build on. IRCoT+HippoRAG2 is the more revealing case: the reformulation step anchors on *Honourable Peter Vijay*, abandons the remaining *Vijay* signal used by the base retriever, and returns the wrong answer. Even the mild US corruption (a single-letter change, *Patal* $\rightarrow$ *Patil*) breaks both IRCoT variants while the base retrievers stay intact.

This answers RQ3: corruption of query entities is the dominant failure mode across all three benchmarks, and structural extensions can erase remaining signal preserved by the underlying retriever during later reformulation and linking steps.

5.4 Lightweight Mitigations Reveal a Structural Gap (RQ4)

We apply both surface-form mitigations as diagnostic probes on the cell with the largest degradation (2WikiMultiHopQA under NG). Each targets a specific hypothesis about the failure. N-best decoding tests whether the gap is reducible to Whisper's sampling noise: if so, another hypothesis from the same model should recover the correct entity. Phonetic

Method	N-best		Phonetic	
	F1	Rec.	F1	Rec.
Naive RAG	0.328	-2.2%	0.337	$+4.4\%$
HippoRAG2	0.425	$+2.5\%$	0.439	$+11.1\%$
IRCoT+Naive	0.337	-1.1%	0.347	$+4.5\%$
IRCoT+HippoRAG2	0.448	-1.0%	0.464	$+7.2\%$

Table 3: Mitigation results on 2WikiMultiHopQA NG (vs. Table 1 baselines). Recovery is the fraction of the Oracle–NG gap closed by each mitigation.

entity correction tests whether the gap is reducible to surface-level entity confusion: if so, the correct surface form should be recoverable from the corpus by phonetic neighborhood search. Algorithmic details and hyperparameters are in Appendix A.

Table 3 reports F1 and recovery rate for both mitigations. N-best decoding closes essentially none of the gap: recovery ranges from -2.2% (Naive RAG) to $+2.5\%$ (HippoRAG2), with three of four methods showing slightly negative recovery. The greedy hypothesis is therefore not a noisy sample around a recoverable correct transcription: pooling hypotheses across five decoding temperatures and beam-search candidates yields the same entity corruption. Phonetic correction does better but still modestly: recovery ranges from $+4.4\%$ to $+11.1\%$, with HippoRAG2 the largest beneficiary. The graph-based method's dependence on exact entity match makes it the most sensitive to surface-form correction, yet even at this peak phonetic correction closes only about 11% of the oracle–NG gap. Neither mitigation closes more than 11% of the gap in any configuration.

This answers RQ4: lightweight surface-form mitigations close at most a small fraction of the gap. Targeting these specific error modes does not close the gap, indicating that downstream retrieval structure amplifies remaining entity errors beyond what simple surface corrections recover.

5.5 Beyond Synthetic Speech and Whisper

To probe sensitivity to two artifacts of our experimental setup, we validate the ASR error pattern against real Nigerian speech and rerun the synthesized NG audio with an alternative ASR system.

Real-speech validation. We transcribed 500 real Nigerian-accented utterances from a corpus of general-purpose Nigerian speech[6] with Whisper-large-v3. Real WER reaches 28.9% on average (median 22.0%), $1.7\times$ higher than the 17.1% on our synthesized 2WikiMultiHopQA speech, and at least one entity is mistranscribed in 21% of utterances. Synthesized speech therefore does not overstate ASR noise for real Nigerian speech, and the rate of entity corruption is comparable across synthetic and real Nigerian speech (Appendix E).

ASR system sensitivity. We re-transcribed the 2WikiMultiHopQA NG audio with SeamlessM4T-v2-large (Barrault et al., 2023) and re-ran Naive RAG and HippoRAG2. Despite a 64% higher WER (28.0% vs 17.1%), F1 is essentially unchanged (Naive RAG 0.333 vs 0.331, HippoRAG2 0.421 under both), and HippoRAG2 still outperforms Naive RAG by the same margin ($+0.088$ vs $+0.090$). The ordering between dense and graph-based retrieval therefore does not appear to be Whisper-specific (Appendix F).

Together, these two checks support the entity-corruption mechanism that drives our main amplification finding, and extending the ASR check to iterative configurations is left for future work.

6 Conclusion

We analyzed how upstream ASR errors propagate through four multi-hop RAG architectures across three benchmarks and four English accents. ASR errors in accented speech degrade multi-hop QA performance, and degradation tracks upstream WER (RQ1). Structurally complex retrieval methods amplify rather than absorb these errors (RQ2). Combining entity-graph linking and iterative reformulation widens the oracle–NG F1 gap by 36–67% relative to naive dense retrieval on all three benchmarks. Entity corruption is the dominant failure mechanism, driving 87%–94% of degradation cases on 2WikiMultiHopQA and remaining the largest category on HotpotQA and MuSiQue (RQ3). Lightweight surface-form mitigations close at most a small fraction of the gap, suggesting that the failure is not explained by random ASR noise or surface-level entity confusion alone (RQ4). The entity-corruption rate is consistent with real Nigerian-accented speech, and the ordering between dense and graph-based retrieval does not appear to be Whisper-specific. These results show that structurally complex multi-hop RAG methods remain vulnerable to upstream ASR noise, leaving robust multi-hop retrieval under spoken input an open challenge for future voice-driven systems.

[6] https://huggingface.co/datasets/benjaminogbonna/nigerian_accented_english_dataset

Limitations

Our evaluation uses synthesized speech rather than recorded human speech, with a single voice per accent. Beyond Whisper-large-v3, we also evaluate with SeamlessM4T-v2-large and confirm our finding for dense and graph-based retrieval, with iterative configurations left for future work. The 500 real Nigerian utterances in Section 5.5 show a comparable entity-corruption rate, but testing additional voices and speaker attributes, and evaluating on a corpus of real spoken multi-hop QA queries would strengthen generalizability. The evaluation is restricted to English. Code-switched and tonal languages may exhibit different ASR error patterns. The error-type analysis in Section 5.3 uses a rule-based proxy for named-entity corruption, and human validation would strengthen confidence in this attribution. We use a single LLM (`gpt-4o-mini`) to generate answers, and larger or open-source LLMs may handle corrupted input differently, particularly inside IRCoT. Promising directions include LLM-based ASR correction (Ma et al., 2023), extending correction to every IRCoT step, falling back to dense retrieval under low ASR confidence, and evaluation on real accented-speech corpora such as Common Voice (Ardila et al., 2020) and AfriSpeech-200 (Olatunji et al., 2023).

Ethical Considerations

Accent-related fairness. In real deployments, speakers with different accents face ASR error rates that can differ sharply from those of speakers whose varieties are better represented in ASR training data (Koenecke et al., 2020; Markl, 2022). Our amplification finding implies that complex multi-hop RAG methods propagate these upstream errors more than naive dense retrieval in voice-driven deployments. We flag this as a fairness implication but do not claim a fairness result. Our speech is synthesized, and the real-speech validation establishes only a comparable per-entity corruption rate. It does not establish that end-to-end RAG behavior on real speech matches the synthetic setting. Concrete fairness claims require multi-hop QA evaluation on real spoken queries, which is left for future work.

Synthetic speech and group representation. We use a single TTS voice per accent. This is sufficient as a controlled setting but does not represent within-accent variation in any speaker group. Findings should not be read as quantitative claims about any demographic population. We avoid demographic labels in our analysis and frame all accent comparisons as controlled error conditions, not as estimates about specific speaker groups.

Risk of misuse. The mechanism we identify (entity corruption breaks graph-based retrieval) is a diagnostic finding, not an attack design. The mitigations we benchmark are public, deterministic, and computationally modest. We do not anticipate dual-use concerns beyond those general to RAG and ASR research, such as biased or inaccurate outputs in voice-driven applications.

References

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the twelfth language resources and evaluation conference*, pages 4218–4222.

Loïc Barrault, Yu-An Chung, Mariano Cora Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, and 1 others. 2023. Seamlessm4t: Massively multilingual & multimodal machine translation. *arXiv preprint arXiv:2308.11596*.

Yiming Chen, Xianghu Yue, Chen Zhang, Xiaoxue Gao, Robby T Tan, and Haizhou Li. 2026. Voicebench: Benchmarking llm-based voice assistants. *Transactions of the Association for Computational Linguistics*, 14:378–398.

Tee Kiah Chia, Khe Chai Sim, Haizhou Li, and Hwee Tou Ng. 2008. A lattice-based approach to query-by-example spoken document retrieval. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 363–370.

Alex DiChristofano, Henry Shuster, Shefali Chandra, and Neal Patwari. 2022. Global performance disparities between english-language accents in automatic speech recognition. *arXiv preprint arXiv:2208.01157*.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2025. The faiss library. *IEEE Transactions on Big Data*.

Fahim Faisal, Sharlina Keshava, Md Mahfuz Ibn Alam, and Antonios Anastasopoulos. 2021. Sd-qa: Spoken dialectal question answering for the real world. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3296–3315.

Bernal J Gutiérrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. 2024. Hipporag: Neurobiologically inspired long-term memory for large language models. *Advances in neural information processing systems*, 37:59532–59569.

Bernal Jiménez Gutiérrez, Yiheng Shu, Weijian Qi, Sizhe Zhou, and Yu Su. 2025. From rag to memory: Non-parametric continual learning for large language models. *arXiv preprint arXiv:2502.14802*.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. 2020. Constructing a multi-hop qa dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6609–6625.

Vladimir Karpukhin, Barlas Oguz, 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 the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 6769–6781.

Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R Rickford, Dan Jurafsky, and Sharad Goel. 2020. Racial disparities in automated speech recognition. *Proceedings of the national academy of sciences*, 117(14):7684–7689.

Chia-Hsuan Lee, Shang-Ming Wang, Huan-Cheng Chang, and Hung-Yi Lee. 2018. Odsqa: Open-domain spoken question answering dataset. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 949–956. IEEE.

Vladimir I Levenshtein and 1 others. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710. Soviet Union.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Chia-Hsuan Li, Szu-Lin Wu, Chi-Liang Liu, and Hung-yi Lee. 2018. Spoken squad: A study of mitigating the impact of speech recognition errors on listening comprehension. *arXiv preprint arXiv:1804.00320*.

Rao Ma, Mengjie Qian, Potsawee Manakul, Mark Gales, and Kate Knill. 2023. Can generative large language models perform asr error correction? *arXiv preprint arXiv:2307.04172*.

Nina Markl. 2022. Language variation and algorithmic bias: understanding algorithmic bias in british english automatic speech recognition. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 521–534.

Tobi Olatunji, Tejumade Afonja, Aditya Yadavalli, Chris Chinenye Emezue, Sahib Singh, Bonaventure FP Dossou, Joanne Osuchukwu, Salomey Osei, Atnafu Lambebo Tonja, Naome Etori, and 1 others. 2023. Afrispeech-200: Pan-african accented speech dataset for clinical and general domain asr. *Transactions of the Association for Computational Linguistics*, 11:1669–1685.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A Smith, and Mike Lewis. 2023. Measuring and narrowing the compositionality gap in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5687–5711.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 2383–2392.

Weitong Ruan, Yaroslav Nechaev, Luoxin Chen, Chengwei Su, and Imre Kiss. 2020. Towards an asr error robust spoken language understanding system. In *Proc. Interspeech 2020*, pages 901–905.

Nicholas Ruiz and Marcello Federico. 2014. Assessing the impact of speech recognition errors on machine translation quality. In *Proceedings of the 11th Conference of the Association for Machine Translation in the Americas: MT Researchers Track*, pages 261–274.

Christopher Sciavolino, Zexuan Zhong, Jinhyuk Lee, and Danqi Chen. 2021. Simple entity-centric questions challenge dense retrievers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6138–6148.

Rachael Tatman. 2017. Gender and dialect bias in youtube's automatic captions. In *Proceedings of the first ACL workshop on ethics in natural language processing*, pages 53–59.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. Musique: Multi-hop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2023. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 10014–10037.

Bin Wang, Xunlong Zou, Geyu Lin, Shuo Sun, Zhuohan Liu, Wenyu Zhang, Zhengyuan Liu, AiTi Aw,

and Nancy Chen. 2025. Audiobench: A universal benchmark for audio large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4297–4316.

Jason D Williams. 2008. Exploiting the asr n-best by tracking multiple dialog state hypotheses. In *Interspeech*, pages 191–194.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 2369–2380.

Chenyu You, Nuo Chen, Fenglin Liu, Dongchao Yang, and Yuexian Zou. 2020. Towards data distillation for end-to-end spoken conversational question answering. *arXiv preprint arXiv:2010.08923*.

A Mitigation Details

A.1 N-best Decoding

For each spoken query we decode five hypotheses at decoding temperatures $\tau \in \{0, 0.2, 0.4, 0.6, 0.8\}$. $\tau = 0$ is identical to the greedy decode used without any mitigations. The four sampled hypotheses provide slightly different surface forms drawn from regions of higher acoustic uncertainty. Each hypothesis is independently passed to the retriever, yielding up to five sets of retrieved documents per query. We pass the union of these document sets to the generator as context, deduplicated by document ID and capped at the per-method context budget. For IRCoT-based methods, N-best is applied only to the initial ASR transcription, subsequent IRCoT-generated queries pass through unchanged. The generator and all other pipeline components are unchanged.

A.2 Phonetic Entity Correction

Offline corpus indexing. The corpus entity index combines two sources: (i) all document titles in the retrieval corpus, and (ii) named entities extracted from passage bodies using spaCy NER[7]. Each entity string is normalized and encoded. The index is a mapping from phonetic-code tuples to canonical entity strings, held in memory.

Query-time correction. Given an ASR transcription, the procedure proceeds as follows:

1. **NER extraction.** Run spaCy NER on the transcription and collect mentioned entities.

2. **Phonetic filtering.** Encode each candidate with the same Double Metaphone procedure. Retrieve corpus entities whose phonetic-code Jaccard similarity with the candidate exceeds 0.4, and retain the top 50.

3. **Edit-distance reranking.** Score each retained corpus entity against the candidate using `rapidfuzz.fuzz.ratio`, a normalized Levenshtein similarity on a 0–100 scale.

4. **Threshold gating.** If the top match exceeds 75, accept it. Otherwise, fall back to a full-corpus search and accept matches only above 85. The stricter fallback threshold protects

[7] Model: `en_core_web_sm`; entity types retained: `PERSON`, `ORG`, `GPE`, `LOC`, `FAC`, `WORK_OF_ART`, `EVENT`, `PRODUCT`, `NORP`.

against spurious string matches when phonetic similarity is no longer available as a prior.

5. **Replacement.** Replace the original candidate span in the transcription with the accepted canonical form, using word-boundary regex matching to avoid corrupting adjacent tokens.

Hyperparameter selection. The thresholds (Jaccard 0.4, edit-distance 75, fallback edit-distance 85) and the top-$K = 50$ candidate pool were chosen by inspection of representative ASR error cases prior to the main evaluation. They are kept fixed across all datasets and accents. No tuning is performed on the $1{,}000$ evaluation samples.

B Error Categorization Rules

Pipeline. For each (question, ASR transcription) pair we compute a word-level difference using Python's `difflib.SequenceMatcher` after whitespace tokenization, no lowercasing, to preserve case for entity detection.

Per-operation labels. A non-equal difference operation (`replace`, `delete`, `insert`) is labelled based on the original word:

- **entity corruption** if the word matches the capitalized-phrase regex `\b[A-Z][a-z]+(?:\s+[A-Z][a-z]+)*\b`;
- **number corruption** if the word matches `\b\d+(?:st|nd|rd|th)?\b`;
- **function-word change** if the word is in our function-word list (Table 4);
- **content-word change** otherwise.

Aggregate labels. A question receives severe garbling if its WER exceeds 0.20 and it has at least one content-word change. It receives function-word noise if it has function-word changes and no other label has occurred. It receives other content change when changes occurred but no other aggregate label was triggered. Labels are not mutually exclusive: *entity corruption* and *severe garbling* frequently co-occur on Nigerian speech.

C Exact Match Results

Table 5 reports Exact Match (EM) for every condition in Table 1. EM is a stricter metric than F1, requiring exact string match after normalization. EM gaps are correspondingly smaller in absolute terms but track F1 gaps tightly, with Pearson $r > 0.95$ across all cells. The method-level amplification ordering observed for F1 is preserved on every dataset, and the largest EM gap reaches 0.189 on 2WikiMultiHopQA under IRCoT+HippoRAG2, mirroring the structural amplification pattern reported for F1.

Category	Words
Articles (3)	a, an, the
Prepositions (13)	of, in, on, at, to, for, with, by, from, as, into, about, between
Forms of *be* (7)	is, was, are, were, be, been, being
Other auxiliaries (6)	have, has, had, do, does, did
Modals (8)	will, would, shall, should, may, might, can, could
Conjunctions (7)	and, or, but, not, no, if, than
Pronouns / det. (10)	that, which, who, whom, this, these, those, it, its, whose
Wh-words (5)	what, where, when, how, why

Table 4: Function-word list used for ASR error categorization. Matching is case-insensitive after stripping non-alphanumeric characters.

D Error-Type Distribution on HotpotQA and MuSiQue

Table 2 focuses on error analysis on 2WikiMultiHopQA, where each question references multiple named entities. This makes entity corruption the most visible failure mode. We apply the same rule-based labeling to HotpotQA and MuSiQue degradation cases under Naive RAG (Tables 6 and 7). Entity corruption remains the largest single failure category on every accent across both datasets, confirming that our finding generalizes beyond 2WikiMultiHopQA.

Two patterns are worth noting. Severe garbling on HotpotQA NG reaches 44%, co-occurring with entity corruption rather than competing with it. NG questions are partially destroyed at the sentence level rather than losing just one entity. "Other content change" on MuSiQue US reaches 28%, the only cell where a non-entity category meaningfully competes with entity corruption. This likely reflects MuSiQue's small number of degradation cases ($n = 46$) and low overall WER (5.2%).

E Real-World Speech Validation

Dataset. We sample the first 500 utterances from a publicly available corpus on HuggingFace[8]. The

[8] https://huggingface.co/datasets/benjaminogbonna/nigerian_accented_english_dataset

Dataset	Method	Oracle	US	IN	PH	NG	Gap
HotpotQA	WER	—	9.4%	11.0%	10.6%	14.5%	—
	Naive RAG	0.488	0.456	0.439	0.446	0.403	0.085
	IRCoT+Naive	0.508	0.476	0.453	0.469	0.417	0.091
	HippoRAG2	0.568	0.534	0.514	0.519	0.473	0.095
	IRCoT+HippoRAG2	0.588	0.546	0.519	0.534	0.463	**0.125**
2WikiMultiHopQA	WER	—	13.6%	14.1%	13.8%	17.1%	—
	Naive RAG	0.419	0.346	0.330	0.343	0.286	0.133
	IRCoT+Naive	0.475	0.354	0.359	0.366	0.293	0.182
	HippoRAG2	0.510	0.425	0.432	0.420	0.347	0.163
	IRCoT+HippoRAG2	0.569	0.486	0.451	0.459	0.380	**0.189**
MuSiQue	WER	—	5.2%	5.5%	5.7%	7.9%	—
	Naive RAG	0.180	0.160	0.151	0.156	0.138	0.042
	IRCoT+Naive	0.209	0.172	0.166	0.174	0.158	0.051
	HippoRAG2	0.242	0.218	0.212	0.212	0.188	0.054
	IRCoT+HippoRAG2	0.270	0.239	0.240	0.246	0.215	**0.055**

Table 5: Exact Match (EM) across three multi-hop QA datasets under oracle and four English accents. All settings follow the convention in Table 1.

Error type	US	IN	PH	NG
Entity corruption	67%	70%	75%	82%
Severe garbling	14%	24%	24%	44%
Number/date corrupt.	5%	5%	5%	4%
Function-word noise	4%	9%	8%	7%
Other content change	21%	13%	11%	7%
# degradation cases	81	103	88	136

Table 6: Distribution of error types within degradation cases on HotpotQA under Naive RAG. Cases can carry multiple labels.

Error type	US	IN	PH	NG
Entity corruption	59%	78%	66%	68%
Severe garbling	4%	7%	14%	21%
Number/date corrupt.	0%	0%	0%	1%
Function-word noise	9%	4%	2%	1%
Other content change	28%	13%	20%	23%
# degradation cases	46	54	50	71

Table 7: Distribution of error types within degradation cases on MuSiQue under Naive RAG. Cases can carry multiple labels.

corpus contains short Nigerian-accented English utterances from multiple speakers and is released for research use. Because the utterances are general-purpose speech rather than multi-hop QA queries, we use the set only for ASR error-pattern validation.

Transcription and WER. We transcribe each utterance with the same Whisper-large-v3 configuration used in the main pipeline. Real Nigerian speech is substantially noisier than our synthesized Nigerian condition: mean WER on real speech reaches 28.9% (median 22.0%), compared to 17.1% on 2WikiMultiHopQA, 14.5% on HotpotQA, and 7.9% on MuSiQue under synthesized speech. Our synthesized condition therefore underestimates the actual ASR difficulty by a factor of $1.7\times$ to $3.7\times$ across the three benchmarks. The WER distribution on real speech also has a heavy tail: a noticeable fraction of utterances exceed 50% WER, reflecting variation in recording channel and speaker that TTS cannot reproduce.

Entity corruption rate. For each utterance we extract named entities from the reference text using spaCy NER and check whether each entity is preserved verbatim in the transcription. At least one entity is mistranscribed in 21% of real utterances. The rate on our synthesized Nigerian 2WikiMultiHopQA condition is comparable, differing from real speech by a factor of $1.08\times$. This supports the use of TTS as a controlled probe for the entity-

Method	WER (W / S)	F1 (W)	F1 (S)
Naive RAG	17.1 / 28.0	0.331	0.333
HippoRAG2	17.1 / 28.0	0.421	0.421

Table 8: End-to-end F1 with Whisper-large-v3 (W) and SeamlessM4T-v2-large (S) on 2WikiMultiHopQA NG.

corruption mechanism behind the amplification reported in Section 5.2.

F ASR Sensitivity Analysis

Setup. We replace Whisper-large-v3 with SeamlessM4T-v2-large (Barrault et al., 2023), using greedy decoding. All other components of the pipeline remain unchanged: TTS voices, question samples, retrieval corpus, RAG methods, and generator. We rerun the Nigerian-accent condition with Naive RAG and HippoRAG2 on 2WikiMultiHopQA, the configuration that exhibited the largest oracle–NG gap in our main results.

Results. SeamlessM4T yields a mean WER of 28.0% on the same 2WikiMultiHopQA NG audio, 64% higher than Whisper-large-v3's 17.1%. Despite this larger upstream error, end-to-end F1 changes by less than 0.002 on both methods (Table 8): Naive RAG moves from 0.331 to 0.333, and HippoRAG2 stays at 0.421 under both ASR systems. HippoRAG2 still outperforms Naive RAG by the same margin ($+0.090$ under Whisper, $+0.088$ under SeamlessM4T). The F1 stability under a substantial WER change indicates that the dense and graph-based retrieval ordering is not driven by the specific ASR system. The failure mode is sensitive to entity-level surface-form changes in the corrupted query, which both ASR systems produce despite their different overall error rates.