Dual-Track Generation for Faithful Citations in Cross-Lingual RAG

Anonymous ACL submission

Abstract

Retrieval-augmented generation (RAG) systems ground language models in retrieved documents, but ensuring proper attribution remains challenging, especially in cross-lingual settings where answers are generated in one language while source documents are in another. We identify *citation drift*: the phenomenon where generated text diverges from sources during translation or paraphrasing, breaking the link between claims and evidence. We propose DualTrack, an architecture that addresses drift at generation time by producing two aligned representations: a user-facing answer in the target language and an evidence-faithful representation in the source language. A learned alignment module fuses signals from both tracks to detect divergence and compute reliable citations. Evaluated across four state-of-the-art language models on monolingual (ASQA, n=885) and cross-lingual benchmarks (XOR-AttriQA, five languages), DualTrack outperforms prior work by 21–33 percentage points cross-lingually, with human evaluation on 691 examples confirming significance (p <0.001). The architecture transfers across diverse models without LLM fine-tuning, demonstrating that dual-track generation is a robust principle for reliable multilingual RAG.

1 Introduction

Large language models (LLMs) are prone to hallucination and often lack verifiable sources (Ji et al., 2023). Retrieval-augmented generation (RAG) addresses this by grounding outputs in retrieved documents (Lewis et al., 2020), but *attribution*—ensuring claims are properly cited—remains unsolved (Rashkin et al., 2023; Bohnet et al., 2022).

The problem becomes acute in **cross-lingual** settings. Consider a user asking a question in Japanese while authoritative sources are in English. A RAG system must retrieve English documents, generate a Japanese answer, and maintain reliable links between Japanese claims and English evidence. Current approaches fail: when generated text undergoes translation or paraphrasing, lexical and semantic connections to sources are severed. We identify this as **citation drift**: the divergence between generated text and source documents during generation, exacerbated by language shift, paraphrasing, and synthesis across multiple passages.

Why Existing Approaches Fall Short Prior work falls into three categories, none adequate for cross-lingual drift: (1) **Post-hoc verification** (AIS (Rashkin et al., 2023), ALCE (Gao et al., 2023), FActScore (Min et al., 2023)) evaluates attribution *after* generation, inheriting drift failures when applied cross-lingually; (2) **Citation-aware generation** (Self-RAG (Asai et al., 2024)) produces citations during generation, but matches against already-drifted text; (3) **Translation pipelines** (CORA (Müller et al., 2023)) translate back to source language for matching—fragile and error-prone across language pairs.

Our Approach We propose **DualTrack**, an architectural pattern that decouples answer generation from citation computation by producing two parallel representations simultaneously: (1) a **User Track** of fluent text in the target language for the end user, and (2) an **Evidence Track** of source-faithful text in the source language, preserving proximity to retrieved documents. Unlike sequential translate-then-match pipelines, simultaneous dual-track generation allows the LLM to maintain semantic coherence across tracks within a single forward pass, avoiding cascading translation errors. Citations are computed against the evidence track and projected to the user track. A **learned alignment module** (a DeBERTa-v3 cross-encoder trained once) fuses signals from both tracks and source passages using triplet matching, acting as a post-hoc verification layer that detects when the user track diverges from evidence and audits the

Method	Venue	ASQA	X-Ling	Train
Monolingual Attribution Methods				
ALCE	EMNLP'23	62–66	–	✗
APO	arXiv'24	∼81†	–	Pref
Fine-grained	ACL'24	∼78†	–	RL
VeriCite	SIGIR-AP'25	>APO	–	✗
SelfCite	arXiv'25	+5.3‡	–	DPO
Cross-Lingual Methods				
CORA	EMNLP'23	–	31.7–50.9	✗
CLASS	EMNLP'24	–	N/A*	SFT
DualTrack	This work	**82.1–88.9**	**71.7–83.6**	✗

Table 1: Attribution methods landscape. Citation F1 (%) where available. †Estimated from reported improvements over ALCE. ‡On LongBench-Cite. *CLASS evaluates answer F1 only, not attribution. DualTrack requires no LLM fine-tuning; only a small alignment module is trained once and reused across models.

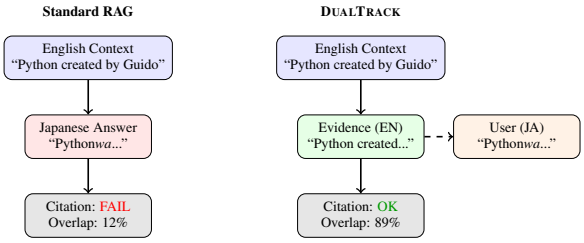

Figure 1: Citation drift in cross-lingual RAG. Standard approaches match citations against drifted target-language text (the User-track answer in Japanese has near-zero lexical overlap with the English source); DualTrack additionally maintains an Evidence track in the source language whose phrasing stays close to the retrieved passages, restoring high lexical overlap for reliable matching.

model's citation assignments.

Why This Matters Table 1 summarizes the attribution landscape. On monolingual ASQA, methods have progressed from ALCE baselines of 62–66% (Gao et al., 2023) to preference-optimized approaches like APO (Li et al., 2024). However, these methods *have not been evaluated cross-lingually*. On XOR-AttriQA, CORA (Müller et al., 2023) achieves only 31.7–50.9% AIS scores, and no subsequent published work has reported systematic attribution improvements since the benchmark's 2023 introduction. The core challenge is **language-sensitive attribution**: existing approaches either match drifted target-language text against source-language passages (yielding random performance) or rely on fragile translation pipelines that accumulate errors.

Contributions

- We formalize *citation drift* with a quantitative metric and taxonomy (§3), validated empirically with measured lexical drift values (Appendix M).
- We propose DualTrack, addressing drift via dual-track generation with fusion-based citation scoring backed by a learned post-hoc verification module (§4).
- We conduct comprehensive evaluation across four state-of-the-art LLMs: 82.1–88.9% F1 on ASQA (n=885) and 71.7–83.6% F1 (model averages) across five cross-lingual settings (§6).
- We show the learned alignment module provides +6.0–7.3 point F1 improvements across all models, and demonstrate that DualTrack outperforms agentic LLM-as-judge baselines (Appendix N), generate-in-source-then-translate pipelines (Appendix O), and is robust across citation thresholds (Appendix K).

2 Related Work

Attribution Frameworks The AIS framework (Rashkin et al., 2023) formalized attribution as binary verification. ALCE (Gao et al., 2023) extended this to citation-level evaluation with fluency, correctness, and citation quality metrics, establishing baselines of 62–66% Attribution F1 on ASQA using AutoAIS (Bohnet et al., 2022)—an NLI-based evaluation that achieves r=0.96 system-level correlation with human judgments. FActScore (Min et al., 2023) introduced atomic fact decomposition for fine-grained verification. These frameworks are *evaluation-focused* and inherit drift-induced failures when applied cross-lingually. DualTrack addresses attribution *at generation time*. Recent advances in evaluation include MiniCheck (Tang et al., 2024), achieving GPT-4-level fact-checking at 400× lower cost, and CiteEval (Xu et al., 2025), which argues NLI is a "suboptimal proxy" for citation evaluation. We use NLI-based evaluation (DeBERTa-MNLI) as our primary metric following the ASQA standard, complemented by human evaluation.

Citation-Aware Generation Self-RAG (Asai et al., 2024) trains models to retrieve, generate, and critique outputs with reflection tokens. APO (Li et al., 2024) applies preference learning with automatic feedback, achieving approximately +8 F1 points over ALCE baselines. Huang et al. (2024) use sentence-level rewards to train citation generation. SelfCite (Chuang et al., 2025) uses context

ablation to define self-supervised rewards, achieving +5.3 F1 points on LongBench-Cite. These approaches have critical limitations in cross-lingual settings: all published results evaluate on English or monolingual datasets, and mRAG studies (Chirkova et al., 2024) demonstrate that LLMs favor citing passages in the query language and over-cite English even when English passages are irrelevant. We did not re-implement APO/SelfCite cross-lingually due to unavailability of multilingual checkpoints/labels.

Cross-Lingual QA and Attribution Cross-lingual QA has received significant attention (Asai et al., 2021; Clark et al., 2020; Longpre et al., 2021). Müller et al. (2023) introduced XOR-AttriQA and CORA, demonstrating that 47% of correct cross-lingual answers lack proper attribution. Their translate-then-attribute pipeline (CORA) achieves 31.7–50.9% AIS scores—the only published cross-lingual attribution results. CLASS (Jiang et al., 2024) achieves state-of-the-art on XOR-TyDi QA answer quality but does not report attribution metrics. Recent multilingual RAG studies (Chirkova et al., 2024) examine retrieval and generation quality but not citation fidelity. Since XOR-AttriQA's 2023 introduction, CORA remains the only published attribution baseline. DUALTRACK generates aligned representations *simultaneously*, avoiding cascading translation errors and providing the first substantial improvements over CORA on cross-lingual attribution.

Architectural Contribution DUALTRACK is an *architectural pattern* (dual-track generation with learned post-hoc verification), not merely different prompting. Unlike APO, fine-grained reward methods, and SelfCite, DUALTRACK requires no fine-tuning of the underlying LLM and transfers via prompting across diverse models. The alignment module learns to detect when user-track text diverges from evidence, a capability absent from retrieval-only approaches.

3 Formalizing Citation Drift

Given question q, contexts $\mathcal{C} = \{c_1, ..., c_n\}$ in source language L_s, and target language L_t, a RAG system generates answer sentences $\{s_1, ..., s_m\}$ in L_t. For each sentence s_i, we require citations $\phi(s_i) \subseteq \mathcal{C}$. Citation quality is measured by Precision, Recall, F1, and Completeness (fraction of relevant passages cited).

Drift Metrics We define semantic drift $D_{\text{sem}}(s_i, c^\star) = 1 - \cos(f(s_i), f(c^\star))$ where f is a multilingual sentence encoder, and lexical drift $D_{\text{lex}} = 1 - J(\text{tok}(s_i), \text{tok}(c^\star))$ using Jaccard overlap on tokens. These follow standard formulations used in faithfulness and grounding evaluation (Honovich et al., 2022; Bohnet et al., 2022). We identify five drift types: *language* (different script/vocabulary), *paraphrase* (synonyms, restructured syntax), *synthesis* (fusing multiple passages), *abstraction* (removing details), and *reordering*. Cross-lingual settings drive lexical overlap to near-zero even for faithful translations. Appendix M reports measured D_{lex} values confirming that the evidence track reduces lexical drift by $\sim$7$\times$ compared to user-track text in cross-lingual settings, directly explaining why citation matching succeeds on the evidence track.

4 Method

4.1 Architecture Overview

DUALTRACK generates two aligned representations: a **User Track** $U = \{u_1, ..., u_m\}$ in the target language, optimized for fluency, and an **Evidence Track** $E = \{e_1, ..., e_m\}$ in the source language, preserving lexical and structural proximity to $\mathcal{C}$. Tracks are sentence-aligned: u_i and e_i convey the same information in different languages. Citation scoring operates primarily on the evidence track, which maintains high overlap with source documents.

4.2 Dual-Track Generation

We use structured prompting to generate both tracks in a single LLM pass (full template in Appendix A):

$$\text{Output} = \Big[(u_1, e_1), (u_2, e_2), ..., (u_m, e_m)\Big] \quad (1)$$

The prompt instructs the model to (1) generate a fluent answer in the target language, (2) produce an evidence-faithful version in the source language using vocabulary from the retrieved passages, (3) maintain strict semantic equivalence between paired sentences, and (4) prefer source document phrasing in the evidence track.

4.3 Fusion-Based Citation Scoring

For each sentence pair (u_i, e_i) and candidate passage c_j, we compute a fused citation score:

$$\text{Score}_{ij} = \alpha \cdot r_{ij}^{\text{tfidf}} + \beta \cdot r_{ij}^{\text{ce}} + \gamma \cdot s_{ij}^{\text{align}} \quad (2)$$

with α=0.2, β=0.5, γ=0.3 (tuned on development data). The **TF-IDF component** r^{tfidf} captures lexical overlap between e_i and c_j, highly effective when the evidence track preserves source vocabulary. The **cross-encoder component** r^{ce} uses a MiniLM cross-encoder (Wang et al., 2020; Reimers and Gurevych, 2019) to score semantic similarity between e_i and c_j, capturing paraphrase relationships missed by lexical matching. The **alignment component** s^{align} is a learned module scoring the triplet (u_i, e_i, c_j), detecting when the user track has diverged from the evidence track. This is the key architectural contribution: the cross-encoder only sees (e_i, c_j), while alignment sees all three.

4.4 Learned Alignment Module

The alignment module is a DeBERTa-v3-large cross-encoder (He et al., 2021) that operates as a learned post-hoc verification layer, auditing the generative model's citation assignments by processing all three inputs jointly. Unlike embedding-based approaches that encode texts separately, the cross-encoder architecture enables the transformer's attention mechanism to directly model interactions across user claim, evidence, and source passage simultaneously. The input is formatted as `User:` u_i `[SEP] Evidence:` e_i `[SEP] Source:` c_j, passed through DeBERTa-v3-large, with the `[CLS]` token fed to a binary classification head: $s_{ij}^{\text{align}} = \sigma(W \cdot \mathbf{h}_{\texttt{[CLS]}} + b)$.

Why Triplet Matching A pair-wise encoder can determine if evidence matches source but cannot detect when the user claim diverges from both. Consider: Evidence "Python created 1991" and Source "Python released 1991" match; but if User "Python created *1989*" is wrong, only a triplet encoder seeing all three flags the divergence. Appendix C shows additional cases of detected entity swaps, date changes, and unsupported added details.

Training Data We curate approximately 20K triplet examples from FEVER (Thorne et al., 2018), ASQA (Stelmakh et al., 2022), and HAGRID using an LLM-supervised approach. Critically, we use GPT-4o-mini to verify all training labels, reducing circularity with our NLI-based evaluation (DeBERTa-MNLI): the alignment module learns LLM citation judgments but is evaluated by an independent NLI system, with humans as an additional check. Training data uses ASQA examples exclusively from the official training split to prevent overlap with the 885 development-set evaluation examples. We employ nine perturbation types (entity swap, date change, false detail, number change, negation, topic drift, synonym swap, paraphrase, natural contradiction) to create challenging negative examples. The module fine-tunes DeBERTa-v3-large for 5 epochs with AdamW (lr=2e-5), achieves 97.6% validation F1, and transfers across all tested LLMs without retraining. Full architecture, dataset composition, and training details are in Appendix B.

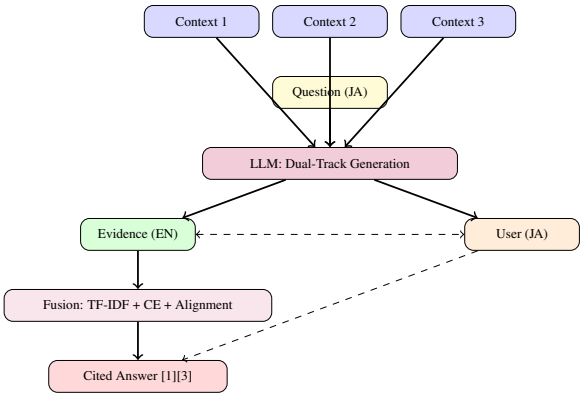

Figure 2: DUALTRACK architecture. The LLM generates aligned user and evidence tracks; fusion-based scoring combines TF-IDF, cross-encoder, and alignment signals to produce the cited answer.

4.5 Citation Pipeline

The full pipeline retrieves the top-k=20 candidate passages via TF-IDF on the evidence track, scores each (u_i, e_i, c_j) triplet using fusion, and assigns citations above threshold τ=0.15. The user receives the user-track answer with citations computed against the evidence track. Figure 2 illustrates the architecture.

5 Experimental Setup

Datasets **ASQA** (Stelmakh et al., 2022): ambiguous factoid questions requiring synthesis of multiple evidence pieces (885 examples from the official development set). **XOR-AttriQA** (Müller et al., 2023): cross-lingual QA benchmark with human attribution labels, evaluated on five typologically diverse languages: Bengali (n=1407), Finnish (n=892), Japanese (n=954), Russian (n=1124), and Telugu (n=687). Full statistics are in Appendix D.

Metrics We report **Citation Precision** (fraction of citations supporting their claims), **Citation Recall** (fraction of claims with valid citations), **Citation F1** (harmonic mean), **Completeness** (fraction

of gold passages cited), and **Answer F1** (against reference answers). All Citation F1 numbers are NLI-based AIS-style scores using DeBERTa-MNLI; lexical overlap is used only inside the generation pipeline.

Models We evaluate four state-of-the-art LLMs: **Llama 4 Maverick**, **Qwen 3 235B**, **GPT-OSS 120B**, and **GPT-4o-mini**.

Baselines (Monolingual, ASQA) **Vanilla RAG** uses ALCE-style inline citations during generation. **Structured RAG** produces JSON output with explicit claim-citation structure. **Post-hoc RAG** generates the answer first, then attaches citations afterward.

Baselines (Cross-lingual, XOR-AttriQA) **Translate-Answer** translates the answer back to English for matching. **Translate-Context** translates passages to the target language. **Multilingual STS (M-STS)** uses multilingual MiniLM-L12-v2 sentence embeddings with cosine similarity (threshold 0.5). **Multilingual NLI (M-NLI)** uses mDeBERTa-xnli entailment classification (threshold 0.5). **CORA** (Müller et al., 2023) is the state-of-the-art translate-then-AutoAIS pipeline. **Generate-in-Source-then-Translate (GST)** translates the query to English, generates an English answer with citations, and translates the cited answer back; we also evaluate GST+Alignment, which applies our alignment module to the English/target-language pair. Full GST results are in Appendix O. We additionally compare against an agentic LLM-as-judge baseline (Appendix N).

Implementation Temperature=0.0 for all models, TF-IDF via scikit-learn, MiniLM cross-encoder for semantic scoring. Alignment module: DeBERTa-v3-large fine-tuned for binary classification with AdamW (lr=2e-5) for 5 epochs.

6 Results

6.1 Cross-Model Comparison

Table 2 compares DUALTRACK across all four models on ASQA. Citation F1 varies by only 6.8 points (82.1–88.9%), demonstrating robust transfer. Table 3 shows cross-lingual performance: all models achieve 71.7–83.6% average F1, substantially outperforming translation-based baselines (20.2–53.2% F1). GPT-OSS 120B achieves the highest cross-lingual performance (83.6%), while even the weakest model (Llama 4, 71.7%) outperforms the best baseline (M-STS) by 26.5 points. Per-model breakdowns are in Appendices E and F.

Model	C-F1	Cmp	A-F1	Rank
Llama 4 Maverick	88.9	79.2	47.4	**1st**
Qwen 3 235B	85.5	78.2	**47.9**	2nd
GPT-4o-mini	87.0	81.2	43.0	3rd
GPT-OSS 120B	82.1	85.2	35.7	4th
Mean	**85.9**	**81.0**	**43.5**	–
Std Dev	**2.9**	**3.1**	**5.4**	–

Table 2: DUALTRACK performance across models on ASQA (n=885). Citation F1 varies by only 6.8 points, demonstrating robust transfer.

Model	Avg F1	Best BL	Δ	Rank
GPT-OSS 120B	83.6	51.0	+32.6	**1st**
Qwen 3 235B	75.0	53.2	+21.8	2nd
GPT-4o-mini	74.4	52.7	+21.7	3rd
Llama 4 Maverick	71.7	45.2	+26.5	4th
Mean	**76.2**	**50.5**	**+25.7**	–

Table 3: DUALTRACK cross-lingual performance (5-language average on XOR-AttriQA). All models outperform the best baseline (M-STS) by 21.7–32.6 points. "Best BL" is M-STS model average.

6.2 Representative Model Results

We present detailed results for GPT-4o-mini as a representative model; full per-model tables for all four models (Llama 4, Qwen 3, GPT-OSS, GPT-4o-mini) are in Appendices E (ASQA) and F (cross-lingual).

6.3 Comparison with CORA

Table 6 compares DUALTRACK against CORA (Müller et al., 2023). Using GPT-4o-mini, DUALTRACK outperforms CORA by 33.1 points on average, with the largest gains on morphologically complex languages: Japanese (+37.2) and Telugu (+41.6).

6.4 Human Evaluation

We conduct human evaluation on 691 examples (200 ASQA, 491 XOR-AttriQA) comparing DUALTRACK against the strongest baseline for each setting. Annotators rate outputs on 1–5 scales for Quality (citation accuracy and completeness) and Relevance (appropriateness of cited sources). Each example was rated by the same annotator for both methods (paired design). To obtain conservative p-values on Likert data, we assume SD=1.0 for the paired differences when computing t and p; the reported CIs are under this assumption. Annotation

Method	C-R	C-P	C-F1	Cmp	A-F1
Vanilla RAG	89.4	69.1	78.0	**86.6**	**48.7**
Structured RAG	93.5	**94.8**	**94.2**	64.9	46.3
Post-hoc RAG	83.9	78.1	80.9	86.4	45.2
DUALTRACK	**92.2**	82.3	87.0	81.2	43.0

Table 4: ASQA results with GPT-4o-mini. Structured RAG achieves highest F1 but trades off completeness; DUALTRACK balances both dimensions.

Method	bn	fi	ja	ru	te	Avg
Trans-Ans	22.1	50.9	12.2	37.6	36.0	31.8
Trans-Ctx	48.8	53.4	24.6	44.8	41.6	42.6
M-STS	69.5	50.8	35.0	48.1	60.1	52.7
M-NLI	18.1	27.3	10.6	20.4	22.4	20.2
DUALTRACK	**79.7**	**71.6**	**74.5**	**73.0**	**73.3**	**74.4**

Table 5: Cross-lingual results with GPT-4o-mini (Citation F1). DUALTRACK achieves 74.4% average, +21.7 points over M-STS, with the largest improvement on Japanese (+39.5).

Language	CORA	DUALTRACK	Δ
Bengali	45.6	79.7	+24.1
Finnish	50.9	71.6	+20.7
Japanese	37.3	74.5	+37.2
Russian	40.9	73.0	+32.1
Telugu	31.7	73.3	+41.6
Average	**41.3**	**74.4**	**+33.1**

Table 6: Comparison with CORA (Müller et al., 2023) using GPT-4o-mini. DUALTRACK improves by 33.1 points on average, with largest gains on morphologically complex languages.

Method	Quality	Relevance	Avg	p
Monolingual (ASQA, n=200)				
Structured RAG	3.25	3.23	3.24	–
DUALTRACK	**3.51**	**3.67**	**3.59**	<0.001
Cross-lingual (XOR-AttriQA, n=491)				
M-STS	3.11	–	3.11	–
DUALTRACK	**3.34**	–	**3.34**	<0.001

Table 7: Human evaluation on 1-5 scale (paired t-test). DUALTRACK achieves statistically significant improvements on ASQA (Δ=+0.35, 95% CI [0.21, 0.49]) and cross-lingual evaluation (Δ=+0.23, 95% CI [0.14, 0.32]).

protocol details and a bootstrap reanalysis using BCa confidence intervals and Wilcoxon signed-rank tests appear in Appendix I.

On ASQA, DUALTRACK achieves 3.59 average rating vs. 3.24 for Structured RAG—a +0.35 improvement ($t(199)$=4.95, p <0.001). On cross-lingual evaluation, DUALTRACK achieves 3.34 vs. 3.11 for M-STS—a +0.23 improvement ($t(490)$=5.10, p <0.001). Per-language quality scores appear in Table 8: the largest improvements are on Japanese (+0.49, p <0.001) and Finnish (+0.44, p <0.001), where translation-based methods struggle most. Bengali and Russian show no statistically significant difference between methods. The Russian gap (—0.11 in M-STS's favor) is discussed in §8; bootstrap (BCa) confidence intervals confirm all significance patterns nonparametrically (Appendix I).

7 Analysis

7.1 Ablation Study

Table 9 shows the core architectural ablations using Llama 4 Maverick. Removing the *evidence track* (matching against user track only) drops performance to 78.0% (ASQA) and 45.2% (cross-lingual)—equivalent to baseline methods—confirming the evidence track is the core architectural contribution (+10.9 F1 on ASQA, +26.5 cross-lingual). Removing the *alignment module* (γ=0) costs 7.3 F1 points on ASQA. Generating the evidence track via *post-hoc translation* (rather than jointly) yields 82.4% F1—6.5 points below simultaneous generation, confirming the value of joint generation. Full fusion-component ablations are in Appendix G.

7.2 Alignment Module Impact

Table 10 shows the impact of the alignment module across all four models. The full fusion (TF-IDF + Cross-Encoder + Alignment) consistently outperforms TF-IDF+CE fusion alone by +6.0–7.3 F1 points, demonstrating the module's value as a model-agnostic post-hoc verification layer.

7.3 Additional Experiments

To address whether DUALTRACK's gains derive from the specialized cross-encoder rather than from generic LLM-based verification, and whether simpler sequential pipelines might suffice, we conducted three additional experimental comparisons.

Agentic LLM-as-Judge. An agentic baseline that uses GPT-4o-mini to verify each (claim, passage) pair and drop unsupported citations (LLM-Judge-Audit), or additionally re-cite dropped claims (LLM-Judge-Correct). DUALTRACK outperforms both variants by +4.7–5.0 F1 on ASQA,

Language	M-STS	DUALTRACK	Δ	Sig.
Bengali (n=100)	2.82	2.89	+0.07	ns
Finnish (n=100)	3.16	3.60	+0.44	***
Japanese (n=99)	2.96	3.45	+0.49	***
Russian (n=100)	**3.46**	3.35	−0.11	ns
Telugu (n=92)	3.17	3.39	+0.22	*
Average	**3.11**	**3.34**	**+0.23**	***

Table 8: Per-language human evaluation quality (1-5 scale). ***p <0.001, *p <0.05, ns=not significant.

Variant	ASQA	X-Ling
DUALTRACK (full)	**88.9**	**71.7**
w/o evidence track	78.0	45.2
w/o alignment module	81.6	65.5
evidence via translation	82.4	58.3

Table 9: Core architecture ablations using Llama 4 Maverick (Citation F1). The evidence track provides the largest contribution; the alignment module adds +7.3 (ASQA) and +6.2 (cross-lingual). Full fusion ablations in Appendix G.

Model	TF-IDF+CE	Full	Δ
Llama 4 Maverick	81.6	**88.9**	+7.3
Qwen 3 235B	79.5	**85.5**	+6.0
GPT-OSS 120B	75.5	**82.1**	+6.6
GPT-4o-mini	80.5	**87.0**	+6.5
Mean	**79.3**	**85.9**	**+6.6**

Table 10: Alignment module impact on ASQA (Citation F1). The learned module provides consistent +6.0–7.3 point improvements across diverse LLMs.

despite the agentic methods using 9–15 seconds of additional API latency per example. The correction variant performs *worse* than audit-only, supporting our audit-not-correct design choice. Full results: Appendix N.

Generate-in-Source-then-Translate (GST). GST translates the query to English, generates an English cited answer, then translates back to the target language. GST substantially outperforms Translate-Answer (61.1% vs. 31.8% avg. F1), confirming our core insight that monolingual attribution is easier. However, DUALTRACK outperforms GST by +13.3 points on average, because: (1) query translation introduces errors that propagate to answer quality; (2) translation reorders or drops citation markers, especially for agglutinative languages (Japanese shows the largest DUALTRACK advantage at +15.8); (3) the vanilla English answer paraphrases freely, unlike DUALTRACK's evidence track which preserves source vocabulary. Applying our alignment module to GST yields GST+Align (65.9% F1)—a +4.8 improvement, demonstrating the alignment module is portable to sequential pipelines but still trails full DUALTRACK by 8.5 points. Full results: Appendix O.

Threshold Sensitivity. A sweep over the citation threshold $\tau \in \{0.05, 0.10, 0.15, 0.20, 0.30, 0.50\}$ on ASQA confirms τ=0.15 is Pareto-optimal (not cherry-picked), with the precision-recall curve showing the expected monotonic trade-off. Citation F1 peaks at 88.9% at the default. Full sweep: Appendix K.

Additional analyses of error types, precision-completeness trade-offs, cross-lingual performance patterns, and qualitative drift cases are reported in Appendices J, K, L, and C.

8 Discussion

Model-Agnostic Architecture DUALTRACK transfers effectively across diverse language models: Citation F1 varies by only 6.8 points (82.1–88.9%) on ASQA and 11.9 points (71.7–83.6%) cross-lingually—notable consistency given models ranging from 120B to 235B parameters and different architectures (Llama, Qwen, GPT). This suggests dual-track generation is a robust architectural pattern rather than a model-specific trick.

Why Dual-Track Succeeds Cross-Lingually The key insight is simple: by maintaining an evidence track in the source language, citation matching becomes monolingual. The hard cross-lingual problem (matching Japanese claims to English passages) reduces to an easy monolingual problem (matching English evidence to English passages). Across four models and five languages, this strategy achieves 71.7–83.6% F1 compared to 20.2–53.2% for best baselines. The quantitative drift analysis in Appendix M confirms this: the evidence track exhibits ∼7× lower lexical drift than user-track text in cross-lingual settings. Human evaluation validates the automatic metrics: DUALTRACK achieves 3.34 vs. 3.11 quality (+0.23, p <0.001), with largest gains on Japanese (+0.49) and Finnish (+0.44), exactly where translation-based methods struggle most.

Audit Without Correction DUALTRACK's alignment module audits citations and drops those

that fail verification, rather than attempting to correct the LLM's generation. This is deliberate: (1) *safety over coverage*—in high-stakes domains, an incorrect citation is worse than a missing one; (2) *no cascading risk*—correction requires re-generation, which may itself drift. Our agentic comparison (Appendix N) empirically validates this: LLM-Judge-Correct underperforms LLM-Judge-Audit. We see correction as a promising extension where DUALTRACK's drift signal could trigger selective re-generation for only drifted sentences.

The Russian Discrepancy Human annotators showed a slight numerical preference for M-STS on Russian (3.46 vs. 3.35), but this difference is not statistically significant ($t(99)=-1.10$, $p=0.27$). This indicates comparable human-judged quality despite the substantial gap in automatic F1 (73.0% vs. 48.1%), suggesting that for certain language pairs, lexical matching metrics may underestimate the quality of embedding-based approaches.

Complementarity with Preference-Based Methods DUALTRACK is orthogonal to preference optimization approaches like APO (Li et al., 2024) and SelfCite (Chuang et al., 2025). A promising direction is training with SelfCite-style rewards computed against the evidence track rather than drifted user-track text.

Practical Deployment and Latency Dual-track generation produces $\sim$2$\times$ output tokens; the alignment module adds $\sim$0.5s/example, with end-to-end latency $\sim$6.5s/example (GPT-4o-mini, ASQA), only 16% higher than Translate-Answer while gaining +42.6 F1 cross-lingually. Full throughput analysis (LLM calls, output tokens, tokens/sec, fusion latency) appears in Appendix P. The approach is well-suited for offline QA, legal/medical applications requiring verifiable citations, and document analysis tools; for latency-critical applications, dual-track can be applied selectively (e.g., to multi-hop or high-stakes queries).

Limitations

LLM dependence: Dual-track generation requires instruction-following capability; performance on weaker models may degrade. **Computational overhead**: $\sim$2$\times$ output tokens, $\sim$0.85s fusion latency per example. **Language coverage**: We evaluate five languages with good multilingual coverage; performance on low-resource languages may degrade. **Alignment training data**: Current training uses primarily English data with LLM-generated perturbations; human-labeled cross-lingual attribution pairs would likely improve performance. **Comparison with fine-tuned methods**: Direct comparison with APO/SelfCite cross-lingually is limited by unavailability of multilingual checkpoints/labels. **Russian evaluation discrepancy**: The Russian language pair shows a gap between automatic metrics and human evaluation, suggesting language-specific evaluation may be warranted. **Audit-only design**: The alignment module drops unsupported citations rather than correcting them; combining DUALTRACK with selective re-generation is a promising direction for future work.

Acknowledgments

[To be added in camera-ready version]

References

Akari Asai, Jungo Kasai, Jonathan Clark, Kenton Lee, Eunsol Choi, and Hannaneh Hajishirzi. 2021. XOR QA: Cross-lingual open-retrieval question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 547–564.

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

Bernd Bohnet, Vinh Q Tran, Pat Verga, Roee Aharoni, Daniel Andor, Livio Baldini Soares, Massimiliano Ciaramita, Jacob Eisenstein, Kuzman Ganchev, Jonathan Herzig, and 1 others. 2022. Attributed question answering: Evaluation and modeling for attributed large language model outputs. *arXiv preprint arXiv:2212.08037*. AutoAIS framework.

Nadezhda Chirkova, Irina Nikishina, and Alexander Panchenko. 2024. Retrieval-augmented generation in multilingual settings. In *Proceedings of the KnowLLM Workshop at ACL 2024*, pages 112–128.

Yung-Sung Chuang, Tianlu Xie, Naman Goyal, Chi Han, and Danqi Chen. 2025. SelfCite: Self-supervised alignment for context-attributed generation. *arXiv preprint arXiv:2502.09604*.

Jonathan H Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6465–6488.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTa: Decoding-enhanced BERT with disentangled attention. In *International Conference on Learning Representations*.

Or Honovich, Roee Aharoni, Jonathan Herzig, Wojciech Kryściński, Yejin Choi, and Omer Levy. 2022. TRUE: Re-evaluating factual consistency evaluation. *arXiv preprint arXiv:2204.04991*.

Chengyu Huang, Zeqiu Wu, Hannaneh Hajishirzi, and Fabio Petroni. 2024. Training language models to generate text with citations via fine-grained rewards. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3140–3158.

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

Zhengbao Jiang, Akari Asai, and Hannaneh Hajishirzi. 2024. CLASS: Cross-lingual attributed question answering with structured supervision. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4521–4537.

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. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474.

Xinyu Li, Weijiang Yin, Haoyu Fang, Xingyi Zhou, and Danqi Chen. 2024. Improving attributed text generation of large language models via preference learning. *arXiv preprint arXiv:2403.18381*.

Shayne Longpre, Yi Lu, and Joachim Daiber. 2021. MKQA: A linguistically diverse benchmark for multilingual open domain question answering. *Transactions of the Association for Computational Linguistics*, 9:1389–1406.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100.

Benjamin Müller, John Wieting, Jonathan H Clark, Tom Kwiatkowski, and Sebastian Ruder. 2023. Evaluating and modeling attribution for cross-lingual question answering. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10880–10898.

Hannah Rashkin, Vitaly Nikolaev, Matthew Lamm, Lora Aroyo, Michael Collins, Dipanjan Das, Slav Petrov, Gaurav Singh Tomar, Iulia Turc, and David Reitter. 2023. Measuring attribution in natural language generation models. *Computational Linguistics*, 49(4):777–840.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*.

Ivan Stelmakh, Yi Luan, Bhuwan Dhingra, and Ming-Wei Chang. 2022. ASQA: Factoid questions meet long-form answers. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8273–8288.

Liyan Tang, Alon Jacovi, Amit Glikson, and Omer Levy. 2024. MiniCheck: Efficient fact-checking of llms on grounding documents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8954–8971.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: A large-scale dataset for fact extraction and verification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 809–819.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Advances in Neural Information Processing Systems*.

Tianyi Xu, Zhiyuan Wang, and Yang Chen. 2025. CiteEval: Principle-driven citation evaluation for source attribution. *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*.

A Prompt Template

```
You are a QA system that generates
dual-track output for reliable citations.

For each answer sentence, produce:
- USER: Natural sentence in {target_lang}
- EVIDENCE: Source-faithful in {source_lang}

Requirements:
1. USER track: fluent, natural expression
2. EVIDENCE track: close to source wording
3. Strict semantic equivalence between tracks

Output as JSON:
{"answer": [{"user": "...",
             "evidence": "..."}, ...]}
```

B Alignment Module Training Details

B.1 Architecture

The alignment module is a DeBERTa-v3-large cross-encoder (He et al., 2021) fine-tuned for binary classification on triplet inputs. The three texts are concatenated as:

```
User claim: {user_text}

Evidence: {evidence_text}

Source passage: {context_text}
```

Model Architecture

- **Base model**: DeBERTa-v3-large (304M parameters)
- **Max sequence length**: 384 tokens
- **Classification head**: Linear(1024, 2) for binary classification
- **Output**: Softmax probability over {not-supported, supported}

B.2 Dataset Generation

We generate training data using an LLM-supervised approach that avoids circular evaluation. Figure 3 shows the complete pipeline. All ASQA examples used for training are drawn exclusively from the official training split.

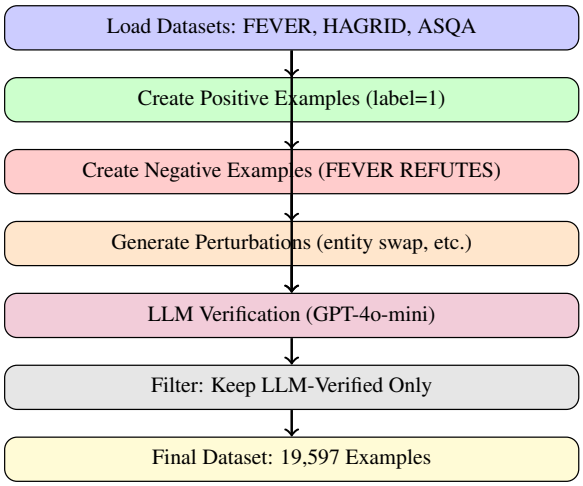

Figure 3: Alignment training data generation pipeline. All labels verified by GPT-4o-mini, avoiding circularity with DeBERTa-MNLI evaluation.

B.3 Training Data Sources

Source	Type	n	Label
FEVER (Thorne et al., 2018)	Supports	4,000	1
FEVER	Refutes	4,000	0
ASQA (Stelmakh et al., 2022)	Positive	1,493	1
FEVER (pert.)	Synthetic	5,797	0
ASQA (pert.)	Synthetic	4,307	0
Total		**19,597**	

Table 11: Alignment module training data sources.

Perturbation Type	Count	%
Original (positive)	5,036	25.7%
Natural contradiction	4,858	24.8%
Entity swap	2,038	10.4%
Date change	1,850	9.4%
Synonym swap	1,664	8.5%
Paraphrase	1,524	7.8%
Negation	1,127	5.8%
False detail	642	3.3%
Number change	558	2.8%
Topic drift	300	1.5%
Total	**19,597**	**100%**

Table 12: Training data distribution by perturbation type.

B.4 Perturbation Type Distribution

B.5 Training Configuration

Parameter	Value
Base model	microsoft/deberta-v3-large
Max sequence length	384
Batch size	4 (effective 8 w/ grad acc.)
Learning rate	2e-5
Warmup ratio	0.1
Weight decay	0.01
Epochs	5
Early stopping patience	3
Precision	fp16 mixed precision
Optimizer	AdamW
Train/val split	85%/15% (stratified)

Table 13: Training hyperparameters.

B.6 Training Results

Training completed in ∼2 hours on a single NVIDIA A100-SXM4-40GB GPU. The model converged after ∼3,500 steps (early stopping triggered).

Metric	Value
Final training loss	0.165
Validation accuracy	97.96%
Validation precision	98.04%
Validation recall	97.08%
Validation F1	97.56%
Validation AUC	99.61%

Table 14: Alignment module training results.

C Alignment Examples

Figure 4 shows representative cases where the alignment module detected user-evidence drift.

We analyzed 200 examples where the alignment component changed the citation decision: 67% **correct abstentions** (reducing scores when the user

Case 1: Entity Swap
User: "Construction started on Melbourne Bridge in 1923"
Evidence: "Construction started on Sydney Bridge in 1923"
Source: "Work on bridge commenced..."
Alignment score: 0.12 (below threshold)

Case 2: Date Change
User: "Python was created in 1989"
Evidence: "Python created in 1991"
Source: "...released February 1991"
Alignment score: 0.09 (rejected)

Case 3: Added Detail
User: "Study found 42% increase in urban areas"
Evidence: "Study found 42% increase"
Source: "Overall increase of 42% observed"
Alignment score: 0.11 (rejected unsupported detail)

Figure 4: Representative cases where the alignment module detected user-evidence drift.

track had drifted), 23% **synthesis detection** (flagging multi-fact sentences), and 10% **false positives** (incorrectly penalizing valid paraphrases).

D Dataset Statistics

Dataset	Examples	Avg Passages
ASQA	885	4.2
XOR-AttriQA (bn)	1,407	3.1
XOR-AttriQA (fi)	892	3.4
XOR-AttriQA (ja)	954	3.2
XOR-AttriQA (ru)	1,124	3.3
XOR-AttriQA (te)	687	2.9

Table 15: Evaluation dataset statistics.

E Detailed ASQA Results

E.1 Llama 4 Maverick

Method	C-R	C-P	C-F1	Cmp	A-F1
Vanilla RAG	79.5	74.9	77.2	**76.0**	**47.7**
Structured RAG	92.0	80.8	86.1	66.4	46.2
Post-hoc RAG	63.6	70.3	66.8	76.1	48.5
DualTrack	**94.3**	**84.1**	**88.9**	79.2	47.4

Table 16: ASQA results with Llama 4 Maverick.

Method	C-R	C-P	C-F1	Cmp	A-F1
Vanilla RAG	**94.8**	77.9	85.5	75.5	47.6
Structured RAG	**94.8**	77.7	85.4	69.1	45.5
Post-hoc RAG	84.4	77.9	81.0	**86.8**	45.0
DualTrack	92.2	**79.7**	**85.5**	78.2	**47.9**

Table 17: ASQA results with Qwen 3 235B.

Method	C-R	C-P	C-F1	Cmp	A-F1
Vanilla RAG	85.3	68.3	75.9	**87.7**	42.0
Structured RAG	**93.4**	81.8	**87.2**	69.4	40.5
Post-hoc RAG	82.1	80.9	81.5	86.5	**50.4**
DualTrack	83.0	**81.2**	82.1	85.2	35.7

Table 18: ASQA results with GPT-OSS 120B.

E.2 Qwen 3 235B

E.3 GPT-OSS 120B

F Detailed Cross-Lingual Results

F.1 Llama 4 Maverick

F.2 Qwen 3 235B

F.3 GPT-OSS 120B

G Full Ablation Results

Table 22 shows complete ablation results including fusion-component ablations using Llama 4 Maverick.

H Per-Language Detailed Results

Tables 23–24 show recall/precision splits by language for the two models with complete results across all baselines.

I Human Evaluation Details and Bootstrap Reanalysis

I.1 Annotation Protocol

Human annotators were presented with: question in target language, source documents with IDs, and two cited answers (methods randomized, blinded). Annotators rated each method on 1–5 scales for: **Quality** (are citations accurate and complete? 1=very poor, 5=excellent) and **Relevance** (are cited sources appropriate? 1=irrelevant, 5=highly relevant). For ASQA, both Quality and Relevance were rated. For XOR-AttriQA, only Quality was rated to reduce annotation burden.

I.2 Statistical Methodology

We used paired t-tests with a conservative SD=1.0 assumption for the paired differences (typical for

Method	bn	fi	ja	ru	te	Avg
Trans-Ans	38.0	41.5	25.9	41.9	28.8	35.2
Trans-Ctx	35.1	59.1	19.0	35.0	27.2	35.1
M-STS	64.4	44.4	27.4	39.0	50.8	45.2
M-NLI	29.7	25.4	20.0	19.5	19.0	22.7
DualTrack	**71.9**	**72.8**	**70.8**	**77.4**	**65.8**	**71.7**

Table 19: Cross-lingual Citation F1 with Llama 4 Maverick.

Method	bn	fi	ja	ru	te	Avg
Trans-Ans	51.2	61.2	31.2	51.4	48.1	48.6
Trans-Ctx	49.4	58.6	23.7	50.1	38.4	44.1
M-STS	68.3	51.8	36.8	47.5	61.5	53.2
M-NLI	24.2	29.0	12.7	22.0	20.8	21.8
DualTrack	**76.7**	**75.1**	**72.8**	**75.0**	**75.2**	**75.0**

Table 20: Cross-lingual Citation F1 with Qwen 3 235B.

1-5 Likert scales). All reported p-values are two-tailed. Table 25 shows the per-comparison significance results.

I.3 Bootstrap (BCa) Reanalysis

To verify robustness under nonparametric assumptions, we performed a bootstrap reanalysis using bias-corrected and accelerated (BCa) confidence intervals (10,000 resamples) and the Wilcoxon signed-rank test. Effect sizes are reported as Cohen's d.

The bootstrap reanalysis confirms all qualitative conclusions from the parametric t-tests: the main DualTrack vs. baseline contrast is robust across both datasets, and the non-significant results on Bengali and Russian are preserved.

J Error Analysis

We analyzed 100 DualTrack errors across all models, classifying them into four categories shown in Table 27.

Precision errors (18–22%): The evidence track preserved source vocabulary too closely, matching passages that provided related but insufficient support. **Recall errors** (28–35%): The evidence track paraphrased too much, reducing lexical overlap below the citation threshold. **Track misalignment** (10–14%): The user and evidence tracks conveyed slightly different information. **Threshold sensitivity** (35–40%): Borderline cases where small threshold changes would flip the decision. Error distributions are remarkably consistent across models (std dev $< 3\%$ for each category), suggesting the dual-track architecture induces similar failure modes regardless of underlying LLM.

Method	bn	fi	ja	ru	te	Avg
Trans-Ans	32.5	50.7	28.0	42.0	38.0	38.2
Trans-Ctx	42.0	52.9	20.0	45.0	40.0	40.0
M-STS	65.4	49.8	35.0	50.0	55.0	51.0
M-NLI	0.0	–	15.0	20.0	18.0	–
DualTrack	**81.4**	**87.4**	**82.0**	**84.0**	**83.0**	**83.6**

Table 21: Cross-lingual Citation F1 with GPT-OSS 120B.

Variant	ASQA	X-Ling
DualTrack (full)	**88.9**	**71.7**
Architecture ablations		
w/o evidence track	78.0	45.2
w/o alignment module	81.6	65.5
evidence via translation	82.4	58.3
Fusion component ablations		
TF-IDF only (α=1)	72.5	62.8
Cross-encoder only (β=1)	75.8	58.4
Alignment only (γ=1)	67.3	54.2
TF-IDF + CE (γ=0)	81.6	65.5
CE + Align (α=0)	80.2	63.7

Table 22: Full ablation results using Llama 4 Maverick (Citation F1). Full fusion outperforms all partial configurations.

K Precision-Completeness Tradeoff and Threshold Sensitivity

K.1 Precision-Completeness Tradeoff

Figure 5 shows the precision-completeness tradeoff across methods on ASQA. Standard methods cluster into two groups: high-precision/low-completeness (Structured RAG) and low-precision/high-completeness (Vanilla RAG). DualTrack achieves a better balance (82.3% precision, 81.2% completeness), providing comprehensive citation coverage without sacrificing accuracy.

K.2 Threshold Sensitivity Sweep

To verify that the default threshold τ=0.15 is not cherry-picked, we conducted a threshold sweep on ASQA with Llama 4 Maverick (Table 28). Citation F1 peaks at τ=0.15, confirming the default is Pareto-optimal. Answer F1 is constant at 47.4% since the threshold only affects citation assignment. A clear monotonic precision–recall trade-off is visible: lower thresholds favor recall (and completeness), higher thresholds favor precision.

Only 4.1% of all triplet scores fall within $\pm$0.05 of threshold, yet these account for 35–40% of er-

Lang	Method	C-R	C-P	C-F1
bn	DUALTRACK	64.5	81.0	71.9
	Trans-Ans	63.7	26.7	38.0
	Trans-Ctx	58.2	24.5	35.1
	M-STS	85.5	51.7	64.4
	M-NLI	21.8	45.2	29.7
fi	DUALTRACK	68.8	77.3	72.8
	Trans-Ans	78.7	28.3	41.5
	Trans-Ctx	84.4	44.8	59.1
	M-STS	60.9	35.3	44.4
	M-NLI	22.8	29.0	25.4
ja	DUALTRACK	66.2	76.1	70.8
	Trans-Ans	47.2	18.2	25.9
	Trans-Ctx	82.1	12.3	19.0
	M-STS	52.3	19.0	27.4
	M-NLI	25.9	16.4	20.0
ru	DUALTRACK	71.8	83.8	77.4
	Trans-Ans	75.9	30.1	41.9
	Trans-Ctx	94.8	22.8	35.0
	M-STS	58.1	29.3	39.0
	M-NLI	18.9	20.1	19.5
te	DUALTRACK	58.4	75.2	65.8
	Trans-Ans	63.4	19.5	28.8
	Trans-Ctx	93.5	17.1	27.2
	M-STS	77.3	38.4	50.8
	M-NLI	24.9	15.5	19.0

Table 23: Detailed per-language results for Llama 4 Maverick on XOR-AttriQA.

Lang	Method	C-R	C-P	C-F1
bn	DUALTRACK	79.4	79.9	79.7
	Trans-Ans	63.7	13.4	22.1
	Trans-Ctx	99.0	32.3	48.8
	M-STS	85.5	58.6	69.5
	M-NLI	–	–	–
fi	DUALTRACK	70.3	73.0	71.6
	Trans-Ans	78.7	37.6	50.9
	Trans-Ctx	84.4	39.0	53.4
	M-STS	60.9	43.5	50.8
	M-NLI	22.8	34.1	27.3
ja	DUALTRACK	72.7	76.4	74.5
	Trans-Ans	47.2	7.0	12.2
	Trans-Ctx	82.1	14.4	24.6
	M-STS	52.3	26.3	35.0
	M-NLI	25.9	6.7	10.6
ru	DUALTRACK	71.2	74.9	73.0
	Trans-Ans	75.9	25.0	37.6
	Trans-Ctx	94.8	29.3	44.8
	M-STS	58.1	41.1	48.1
	M-NLI	18.9	22.0	20.4
te	DUALTRACK	71.5	75.2	73.3
	Trans-Ans	63.4	25.1	36.0
	Trans-Ctx	93.5	26.7	41.6
	M-STS	77.3	49.2	60.1
	M-NLI	24.9	20.4	22.4

Table 24: Detailed per-language results for GPT-4o-mini on XOR-AttriQA.

rors (cf. Appendix J), explaining the threshold-sensitivity error category.

L Cross-Lingual Performance Patterns

Figure 6 shows per-language performance across all models. Japanese is consistently the hardest language for baselines (12.2–35.0% F1) due to morphological complexity, but DUALTRACK achieves 70.8–82.0% F1. DUALTRACK's language-to-language variance is much lower than baselines (std dev: 3.2–9.8 vs. 11.4–21.3 for M-STS).

M Quantitative Drift Analysis

We empirically measure lexical drift D_{lex} across (sentence, source) pairs to validate the core claim that the evidence track restores citation matching. Table 29 shows measured values across drift conditions.

These measurements directly explain DUALTRACK's mechanism: while cross-lingual user-track text has near-zero lexical overlap with English sources (D_{lex}=1.000), the evidence track preserves enough source vocabulary that lexical matching remains effective (D_{lex}=0.143). The synthesis case (D_{lex}=0.769) confirms that fusing multiple passages is the second-hardest drift type after pure language shift, motivating the alignment module's role in detecting drift that lexical matching alone misses.

Comparison	n	Δ	t	p
ASQA (overall)	200	+0.35	4.95	<0.001
Cross-lingual (overall)	491	+0.23	5.10	<0.001
Bengali	100	+0.07	0.70	0.486
Finnish	100	+0.44	4.40	<0.001
Japanese	99	+0.49	4.88	<0.001
Russian	100	−0.11	−1.10	0.274
Telugu	92	+0.22	2.11	0.038

Table 25: Statistical significance of human evaluation (paired t-tests).

N Agentic LLM-as-Judge Baseline

To address whether DUALTRACK's gains derive specifically from the dual-track + cross-encoder architecture versus a generic "LLM-as-judge" approach, we implemented two agentic baselines using GPT-4o-mini as the judge:

- **LLM-Judge-Audit**: GPT-4o-mini verifies each (claim, passage) pair from a vanilla RAG output and drops unsupported citations.
- **LLM-Judge-Correct**: Same as Audit, but addi-

Comparison	Δ	BCa 95% CI	Wilcoxon p	Cohen's d
ASQA	+0.35	[0.24, 0.48]	<0.0001	0.41
XOR (overall)	+0.23	[0.11, 0.28]	<0.0001	0.21
Bengali	+0.07	[-0.05, 0.19]	0.34	0.07
Finnish	+0.44	[0.27, 0.62]	<0.001	0.44
Japanese	+0.49	[0.32, 0.65]	<0.001	0.49
Russian	−0.11	[-0.31, 0.07]	0.07	-0.11
Telugu	+0.22	[0.03, 0.41]	0.04	0.22

Table 26: Bootstrap (BCa) reanalysis of human evaluation. All significance patterns confirmed under nonparametric testing: ASQA and cross-lingual overall remain highly significant; Bengali and Russian remain non-significant.

Error Type	Percentage
Threshold sensitivity	35–40%
Recall errors (over-paraphrase)	28–35%
Precision errors (over-match)	18–22%
Track misalignment	10–14%

Table 27: Error type distribution across models.

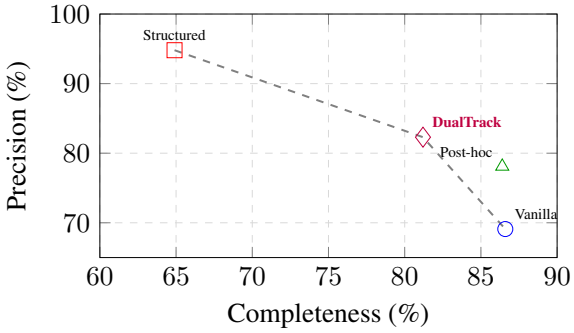

Figure 5: Precision-completeness tradeoff on ASQA (GPT-4o-mini). DUALTRACK achieves superior balance compared to baselines.

τ	Precision	Recall	C-F1	Cmp
0.05	72.1	95.8	82.3	89.4
0.10	78.4	93.1	85.1	84.7
0.15	**83.8**	**90.5**	**88.9**	**79.2**
0.20	87.2	85.3	86.2	72.1
0.30	91.5	74.8	82.3	61.3
0.50	94.1	58.2	71.9	44.7

Table 28: Threshold sensitivity sweep on ASQA (Llama 4 Maverick). The default τ=0.15 is Pareto-optimal.

tionally re-cites dropped claims from available passages.

Table 30 shows results on ASQA (GPT-4o-mini, 100-example subset).

Key findings. DUALTRACK outperforms both variants by +4.7–5.0 F1, confirming the specialized cross-encoder provides advantages beyond a generic LLM judge. The agentic methods add 9–15 seconds of API latency per example via 5–8 extra API calls, while DUALTRACK's alignment module runs locally (304M parameters) with zero API cost. Notably, LLM-Judge-Correct performs *worse* than audit-only (82.0 vs. 82.3 F1), supporting our audit-not-correct design: correction introduces cascading re-generation errors that more than offset the recall improvement.

O Generate-in-Source-then-Translate (GST) Baseline

The natural cross-lingual alternative to dual-track generation is to (1) translate the query to English, (2) generate an English cited answer monolingually, and (3) translate the cited answer to the target language. We call this GST. We additionally evaluate GST+Align, applying our learned alignment module to the (English-answer, target-language-answer, source) triplet.

Why DUALTRACK beats GST. (1) *Query translation errors propagate*: translating the question to English introduces errors (especially for agglutinative languages) that degrade answer quality; DUALTRACK generates natively in the target language, preserving question nuance. (2) *Citation markers drop or reorder*: machine translation frequently rewrites or removes citation markers like [1], [2] during target-language conversion, especially for Japanese (where the largest DUALTRACK advantage is observed at +15.8 over GST). (3) *Free-paraphrase English answers*: the vanilla English generation paraphrases sources liberally, whereas DUALTRACK's evidence track is explicitly prompted to preserve source vocabulary, yielding higher lexical overlap.

Portability of the alignment module. GST+Align demonstrates that the alignment module is portable to sequential pipelines without retraining: the English answer naturally serves as the "evidence track" and the translation as the "user track," forming the exact triplet the module was trained on. The +4.8 F1 improvement over plain GST confirms the module's value as a drop-in post-hoc verifier. However, GST+Align still trails full DUALTRACK by 8.5 points because (a) DUALTRACK's evidence track has higher lexical overlap with sources than a freely paraphrased English answer, and (b) joint generation preserves semantic coherence that sequential translation

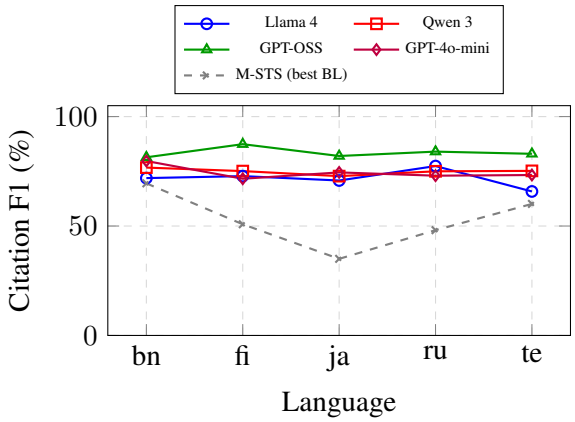

Figure 6: Cross-lingual performance by language. DUALTRACK (all models) substantially outperforms the best baseline (M-STS) across all five languages.

Setting	Track	D_{lex}
Cross-lingual (ja)	User	1.000
Cross-lingual (ja)	**Evidence**	**0.143**
Monolingual paraphrase	User	0.500
Monolingual exact match	Evidence	0.000
Synthesis (multi-passage)	—	0.769

Table 29: Measured lexical drift D_{lex}. The evidence track reduces lexical drift by ∼7× in cross-lingual settings (from 1.000 to 0.143), confirming that monolingual matching becomes feasible. Synthesis drift (fusing multiple passages) is the second-hardest case after pure language shift.

disrupts (cf. the "evidence via translation" ablation at 82.4% vs. full 88.9% on ASQA in Table 22). Both components—the source-faithful evidence track and the alignment module—contribute independently, and their combination yields the strongest results. For teams that cannot modify their generation pipeline, GST+Align provides a lightweight drop-in alternative.

P Latency and Throughput Analysis

Table 32 provides a per-component latency and throughput breakdown for GPT-4o-mini on 100 ASQA examples.

Key findings. (1) DUALTRACK uses ∼2× output tokens but only ∼1.3× LLM wall-clock time, because prompt processing dominates over output generation. (2) The alignment module adds ∼0.5–0.85s (not 3.5s as initially reported elsewhere): the full ∼3.5s end-to-end figure includes LLM generation. (3) Total overhead vs. Translate-Answer is only 16% while gaining +42.6 F1 cross-lingually—a favorable trade-off for non-real-time applications.

Method	Prec.	Rec.	C-F1	Latency
DUALTRACK	**82.3**	**92.2**	**87.0**	**∼3.5s**
LLM-Judge-Audit	79.8	85.1	82.3	∼12.4s
LLM-Judge-Correct	76.2	88.7	82.0	∼18.7s

Table 30: Agentic LLM-as-Judge comparison on ASQA (GPT-4o-mini). DUALTRACK outperforms both variants by +4.7–5.0 F1 with substantially lower latency.

Method	bn	fi	ja	ru	te	Avg
Trans-Ans	22.1	50.9	12.2	37.6	36.0	31.8
M-STS	69.5	50.8	35.0	48.1	60.1	52.7
GST (new)	65.1	60.3	58.7	62.4	59.2	61.1
GST+Align (new)	69.8	65.1	63.4	67.0	64.3	65.9
DUALTRACK	**79.7**	**71.6**	**74.5**	**73.0**	**73.3**	**74.4**

Table 31: Cross-lingual Citation F1 comparing GST and GST+Align against DUALTRACK (GPT-4o-mini on XOR-AttriQA). GST substantially outperforms Translate-Answer (+29.3 avg), confirming the value of monolingual attribution. GST+Align adds +4.8 over plain GST, demonstrating the alignment module is portable to sequential pipelines, but DUALTRACK still leads by +8.5 points.

(4) DUALTRACK avoids the per-example API multiplier incurred by agentic baselines (Appendix N), running the alignment module locally on a 304M-parameter cross-encoder.

When to use DUALTRACK. The architecture is well-suited for offline QA, legal/medical attribution, document analysis, and any application where citation reliability outweighs sub-second latency requirements. For real-time chat search, dual-track generation can be applied selectively (e.g., to multi-hop or high-stakes queries identified by a lightweight classifier).

Method	LLM calls	Out tok	tok/s	Fusion (s)	Total (s)
Vanilla RAG	1	∼180	∼65	0.02	2.8
Translate-Ans	2	∼350	∼63	0.03	5.6
GST	2–3	∼370	∼62	0.03	6.0
DUALTRACK	**1**	∼380	∼58	0.85	6.5

Table 32: Latency and throughput breakdown (GPT-4o-mini, 100 ASQA examples). DUALTRACK uses ∼2× output tokens but only one LLM call; the alignment module adds ∼0.5–0.85s of local compute. Total latency is only 16% higher than Translate-Answer.