# What Breaks Multilingual Long-Form RAG? An Experimental Study of Attribution Errors in Report Generation

**Mahule Roy**[1,2] **Subhas Roy**[3]
[1] University of Oxford [2] Harvard Medical School [3] TATA Consumer Products Limited

## Abstract

Retrieval-Augmented Generation (RAG) can improve the factual grounding of text generation by incorporating external documents. While prior work has examined RAG for short-form question answering, its performance in long-form report generation—particularly in multilingual settings—remains less well understood. In this study, we conduct an exploratory investigation of multilingual long-form RAG, focusing on *attribution faithfulness*, i.e., the extent to which generated claims are supported by cited sources. We evaluate four representative pipeline configurations across two language pairs (English–German and English–Hindi) on a set of 200 report prompts, generating short- to medium-length reports (approximately 300–700 words). Our analysis suggests that multilingual pipelines tend to introduce more attribution inconsistencies than monolingual baselines, that translation-based strategies can improve coverage but occasionally reduce citation fidelity, and that longer reports exhibit modestly lower attribution quality. Prompting strategies provide limited improvements. These findings highlight practical challenges in developing reliable multilingual report generation systems and underscore the importance of careful attribution evaluation.

## 1 Introduction

Text generation systems are increasingly applied in settings such as policy analysis, technical reporting, and decision support, where outputs are expected to be both informative and reasonably reliable. Retrieval-Augmented Generation (RAG) has emerged as a practical approach for such information-dense tasks, as it can incorporate external documents and may help reduce unsupported hallucinations (Lewis et al., 2020). While RAG has been studied extensively in short-form question answering (Izacard et al., 2023; Chen et al., 2022), its behavior in long-form report generation remains relatively underexplored. Long-form report generation introduces several challenges. Reports are typically longer (hundreds to several hundred words), contain interdependent factual claims, and often require explicit attribution. Errors in attribution can accumulate, complicating trust assessment. These challenges may be further amplified in multilingual settings, where retrieval, generation, and translation components interact across languages, potentially increasing the risk of semantic drift and misaligned citations (Hu et al., 2022; Zhang et al., 2023). In this study, we focus on English, German, and Hindi, chosen based on the availability of aligned Wikipedia and news corpora, moderate computational feasibility, and a mix of high- and mid-resource languages. English is used as a pivot in translation-based pipelines, reflecting common practice in multilingual NLP systems due to stronger model support and generally higher translation quality (Costa-Jussà et al., 2022), though we note this choice may not generalize to all languages or scenarios. While we expect observed trends (e.g., cross-lingual drift, length-related attribution challenges) to be indicative of broader behavior, low-resource or typologically distant languages may experience more pronounced errors (Joshi et al., 2020). Similarly, domain-specific corpora (e.g., medical, legal, technical) may exhibit different patterns of error accumulation (Miyachi et al., 2024). Our goal is to provide an empirical investigation of how pipeline design choices affect attribution in multilingual long-form report generation. By systematically analyzing where and how attribution inconsistencies arise, we aim to offer insights that can guide the development of more reliable multilingual RAG systems. This study aims to answer the following questions. First, how does multilingual RAG impact attribution fidelity in long-form reports compared to monolingual baselines? Second, how do report length and retrieval budget affect attribution quality? Third, what is the effect

of pre- versus post-translation strategies on cross-lingual semantic drift and citation reliability? By articulating these questions explicitly, we aim to clarify the scope and focus of our investigation.

## 2 Background and Related Work

Retrieval-Augmented Generation (RAG) integrates information retrieval with neural text generation to provide grounding in external knowledge sources, and has become a widely used approach for knowledge-intensive NLP tasks (Lewis et al., 2020). Following the original RAG framework, prior work has explored sparse and dense retrieval strategies (Karpukhin et al., 2020; Izacard & Grave, 2021), multi-hop retrieval (Xiong et al., 2020), and various methods for incorporating retrieved content into generation (Shuster et al., 2021). Most empirical studies focus on short-form question answering or single-turn interactions, where outputs are brief and explicit attribution is often implicit (Gao et al., 2023). In parallel, research on long-form generation and report-style text has emphasized fluency, coherence, and content organization over extended outputs (See et al., 2017; Eisner, 2000), typically evaluating overall factual accuracy rather than detailed claim-level attribution (Fabbri et al., 2021). Explicit attribution—linking claims to supporting sources—has received comparatively limited attention in this context (Lu, 2023). Existing methods for evaluating attribution include claim verification (Zhong et al., 2020), citation quality assessment (Liu et al., 2023), and faithfulness metrics (Pagnoni et al., 2021), though these approaches are generally designed for sentence-level outputs and do not directly capture the complexities of long-form, citation-rich multilingual reports. Multilingual pre-trained models such as mBERT (Devlin et al., 2019), XLM-R (Conneau et al., 2020), and mT5 (Xue et al., 2021) have enabled cross-lingual retrieval and generation, yet their integration into RAG pipelines introduces additional challenges, including translation drift and potential misalignment between claims and sources (Wu et al., 2020). While cross-lingual retrieval (Li et al., 2022) and multilingual generation (Chung et al., 2024) have been studied separately, evaluation methods for long-form attribution remain underdeveloped. Recent work such as X-RAG (Liu et al., 2025) explores cross-lingual RAG in QA settings, but less is known about long-form report generation. Building on these foundations, our study aims to provide a cautious, empirical analysis of attribution behaviors in multilingual long-form RAG pipelines, examining both language-specific and translation-based approaches, and offering insight into potential sources of inconsistency or error.

## 3 Experimental Framework

To analyze multilingual long-form report generation, we constructed multilingual corpora from publicly available sources. Wikipedia dumps from June 2023 in English, German, and Hindi initially contained approximately 650,000, 280,000, and 15,000 articles, respectively, before filtering. To supplement factual coverage, we incorporated parallel and comparable news articles from Deutsche Welle and the BBC World Service spanning 2020–2023. Cross-lingual alignment was performed at the article level for Wikipedia using WikiMatrix alignments with a confidence threshold of 1.04 and at the paragraph level for news articles using LASER embeddings version 3 with a cosine similarity threshold of 0.82 (Li et al., 2022). A random sample of 100 aligned triplets was manually validated, yielding approximately 90% correctness. Documents were preprocessed to remove boilerplate content using `trafilatura`, normalized with Unicode NFC, and segmented into overlapping passages of 512 tokens with 128-token overlap using language-specific HuggingFace tokenizers. After filtering, the final corpora contained roughly 12,000 English, 8,500 German, and 4,500 Hindi documents, with average lengths of 650, 630, and 600 words per document, respectively. All experiments used fixed random seeds (42) for repeatability. Passage selection for retrieval was performed using either BM25 or FAISS dense search, with the top-$k$ scoring passages selected per query. News articles added approximately 500 complementary stories covering science, technology, history, and policy topics. Report generation requests were designed to reflect realistic information needs across scientific, historical, policy-oriented, and technical domains. Requests targeted three report lengths (300–500 words, 500–800 words, and 800–1,000 words) and included a short background, a clearly defined information need, a length specification, and explicit instructions that claims should be supported by citations. We generated 100 structured English templates, manually refined by two annotators for clarity and diversity. For multilingual experiments, requests were translated into German

and Hindi using NLLB-200-3.3B (Costa-Jussà et al., 2022), with back-translation applied as a consistency check. BLEU scores were computed at the *sentence level*, while BERTScore correlations were measured at the *claim level* to capture semantic fidelity of individual factual assertions. BLEU scores on a held-out sample of 100 sentences per language were 33.1 for EN-DE, 27.4 for EN-HI, 35.5 for DE-EN, and 29.1 for HI-EN, indicating moderate translation quality. We examine sparse and dense retrieval strategies independently, but do not explore hybrid retrieval that combines both methods, nor do we re-rank retrieved documents based on attribution relevance. These approaches could potentially enhance precision while maintaining high coverage.

Retrieval employed both sparse and dense strategies. Sparse retrieval used BM25 implemented via Apache Lucene 9.4 (k1=1.2, b=0.75), while dense retrieval used the multilingual Contriever model with 768-dimensional embeddings (Izacard & Grave, 2021). Documents were segmented into overlapping 512-token chunks prior to indexing. Dense indices were built using FAISS with approximate nearest neighbor search, while the top five documents were retrieved per request. No additional fine-tuning was applied to retrievers, as the study focuses on pipeline behavior rather than component optimization. Text generation used LLaMA-2-7B-chat for English-only baselines and mT5-XXL (Xue et al., 2021) for multilingual generation. Translation-based pipelines employed NLLB-200-3.3B for pre- or post-generation translation. All models shared fixed decoding parameters: temperature 0.7, top-p 0.9, maximum new tokens 800–1,000 depending on report length, and repetition penalty 1.1. Minimal fine-tuning was performed on 5,000 examples from a mixture of Natural Instructions, MS MARCO, and synthetic report pairs, using AdamW with learning rate 2e-5, batch size 16, and linear warmup over 200 steps. Fine-tuning ran on 2×A100 GPUs for approximately four hours per model, sufficient for evaluation without optimizing individual components. We evaluated six RAG pipeline configurations. P1 is a monolingual English baseline (BM25 + LLaMA-2). P2 is an end-to-end multilingual pipeline (dense retrieval + mT5). P3 uses pre-translation: queries and documents translated into English, BM25 retrieval, and report generation in English. P4 applies post-translation: retrieval and generation in source language with multilingual models, followed by translation. P5 is a hybrid pipeline selecting retrieval/generation components adaptively. P6 is a cascaded multi-stage pipeline, generating an initial draft and refining it through additional retrieval and verification. All pipelines used prompts instructing models to include citations, with four variations: minimal, explicit, structured, and instructional. Prompts were professionally translated to ensure linguistic appropriateness. For full-scale evaluation, each pipeline generated reports for 50 requests per language direction, covering all three report lengths, yielding approximately 450 reports per language. For scaled-down experiments used in ablations and attribution evaluation, 25 requests per language direction were generated, yielding roughly 300 reports per pipeline. This distinction clarifies the differing report counts reported elsewhere. Human evaluation was performed on a subset of 80 reports (20 per language direction) with a single fluent reviewer per report. Spot checks indicate roughly 85% agreement with automatic metrics, which should not be interpreted as inter-annotator agreement. Statistical comparisons using pairwise t-tests in scaled-down experiments were conducted with n=25 reports per language direction; the small sample size limits statistical power, so p-values should be interpreted cautiously. By including retrieval, generation, minimal fine-tuning, multiple languages, structured prompts, and clear distinction between full-scale and scaled-down setups, this framework provides a reproducible and informative experimental setup for analyzing multilingual long-form report generation while maintaining clarity and consistency.

## 4 Experimental Variables

To systematically study the effect of pipeline design on attribution fidelity, we manipulate several experimental factors. We evaluate four language directions: English→German, English→Hindi, German→English, and Hindi→English, covering both high- and lower-resource language pairs. Report lengths are varied across three ranges: 300–500 words, 500–800 words, and 800–1,000 words, allowing us to examine how increasing output size influences attribution quality. The retrieval budget is varied by selecting 5, 10, or 20 documents per report to study the impact of evidence quantity on citation behavior. Prompting strategies differ in specificity, ranging from minimal guidance ("Include citations") to explicit sentence-level at-

tribution instructions. Generation stochasticity is controlled via temperature settings of 0.7 and 1.0 to assess the effect of randomness on attribution consistency. Each pipeline generates reports for 25 requests per language direction, with each request producing one or two length variants. This results in approximately 300 generated reports per pipeline in the scaled experiments. In the larger-scale framework, 50 requests per language were used, yielding roughly 450 reports per language. We explicitly distinguish these setups to avoid confusion between scaled-down and full-scale experiments. Back-translation quality is evaluated using BLEU scores computed at the sentence level for translated report prompts, while cross-lingual semantic drift is measured with BERTScore applied to individual factual claims. Human evaluation is conducted on a subset of 80 reports (20 per language direction), with a single fluent reviewer per report. Spot checks indicate approximately 85% agreement between human judgments and the automatic attribution metrics, though this is not inter-annotator agreement. Pairwise statistical tests (t-tests) are performed with these smaller samples (n=25 reports per language direction), and we note that p-values may be sensitive due to limited statistical power. While we explore four prompt types and two temperature settings, we do not test chain-of-thought or intermediate reasoning prompts, which may provide additional guidance to the model and improve attribution quality by encouraging structured reasoning over multiple retrieval passages. Although we report trends for longer reports, our experiments are limited to outputs under 1,000 words. Real-world reports can exceed this length, and attribution errors may accumulate further, suggesting the need to test pipelines on longer documents.

## 5 Methodology

### 5.1 Claim Decomposition

To assess attribution quality, generated reports are decomposed into individual factual claims. Sentences are segmented using language-specific tools (NLTK for English and German, Indic NLP Library for Hindi). Sentences without citations are excluded from verification. For sentences containing citations, dependency parsing is applied using spaCy (v3.5) with appropriate language models to extract clauses expressing factual assertions, such as events, quantities, or descriptive statements. Each claim is normalized via lowercasing, lemmatization, and stopword removal using standard language-specific tools to reduce superficial variation. Cited documents, indicated in the generated text using bracketed IDs (e.g., [DOC123]), are extracted via regular expressions. Claims are categorized as numerical, temporal, relational, or descriptive for finer-grained analysis using rule-based classifiers based on part-of-speech patterns and keyword heuristics. This decomposition allows attribution assessment at the level of individual claims rather than entire sentences or reports. Claims were categorized to allow fine-grained analysis of attribution behavior. Numerical claims include quantities, such as "Germany's GDP grew by 2.1% in 2022." Temporal claims consist of dates or durations. Relational claims describe relationships between entities, for example, "Company X acquired Company Y." Descriptive claims include factual statements not covered by the other categories, such as "The Amazon rainforest spans multiple countries". Our decomposition focuses on numerical, temporal, relational, and descriptive claims. We do not explicitly analyze implicit, multi-sentence, causal, or conditional claims, which may be prevalent in policy, scientific, or technical reports. Future work should extend claim extraction to capture these complex structures for more comprehensive attribution assessment.

### 5.2 Automatic Attribution Verification

Claims are automatically evaluated using complementary metrics: textual entailment using XLM-R-large fine-tuned on XNLI, semantic similarity via BERTScore with the multilingual BERT base model, and question-answering consistency using a multilingual T5 model fine-tuned on SQuAD 2.0. Scores from these signals are averaged to produce a final attribution score between 0 and 1. Thresholds for determining supported claims were calibrated on a small development set of 50 claims per language to maximize F1 against human judgments. The selected thresholds were entailment > 0.75, BERTScore > 0.85, and QA consistency > 0.65. For claims supported across multiple documents, at least one document must exceed all three thresholds for full support. Partial support cases are conservatively treated as unsupported in primary metrics but are reported separately in error analyses. While thresholds are consistent across languages, we note that automatic metrics may introduce bias for lower-resource languages or for translation artifacts. To evaluate the robustness of our metrics,

we performed a sensitivity analysis by varying the thresholds for entailment, BERTScore, and QA consistency by ±0.05 from the calibrated values. Attribution precision varied by only 2–3 percentage points, indicating that our primary thresholds provide stable estimates. Partial support cases, which constitute approximately 10–12% of claims, were analyzed separately to identify the types of errors that conservative metrics might miss.

### 5.3 Human Review Protocol

To supplement automatic evaluation, we conducted a targeted, small-scale human review of a subset of generated reports. Instead of exhaustively annotating the full dataset, we sampled 20 reports per language direction, across pipelines and report lengths, yielding 80 reports in total. Reviewers were fluent in the respective languages (native or near-native proficiency) and performed the assessment using a simplified web interface. Each reviewer evaluated whether cited sources plausibly supported claims, focusing on common errors such as translation-induced misalignment or unsupported citations. Given the smaller scale, each report was annotated by a single reviewer rather than multiple annotators, with brief spot checks for consistency. This approach allowed us to qualitatively capture representative attribution patterns while keeping annotation effort realistic. Human reviewers observed that most unsupported claims arose from translation drift, reuse of citations from earlier sections, and minor factual misalignments in complex sentences. Although only a single annotator evaluated each report, spot checks indicated high agreement with automatic metrics in approximately 85% of cases, suggesting reasonable alignment between human and automated assessment.

## 6 Evaluation Metrics

Attribution quality is quantified using precision, recall, and F1 computed at the claim level. Attribution precision measures the proportion of cited claims that are correctly supported, while recall captures the proportion of all claims accompanied by valid citations. The citation hallucination rate reflects the fraction of claims where the cited sources fail to provide sufficient support. To measure cross-lingual semantic drift, we compute BERTScore similarity between generated claims and their back-translated equivalents. Partial support cases are treated as unsupported in primary metrics to maintain conservative estimates but are analyzed separately. Additional analyses explore how attribution performance varies with report length, retrieval budget, and generation stochasticity, providing insights into pipeline behavior under different conditions. We quantify semantic drift by computing BERTScore similarity between each generated claim and its back-translated equivalent. Lower similarity indicates higher drift, which we correlate with citation hallucinations and use to identify language pairs more prone to errors.

## 7 Experimental Results

### 7.1 Overall Performance Across Pipeline Configurations

Table 1 summarizes average performance across all pipelines, languages, and report lengths in our scaled-down experiments. As expected, the monolingual pipeline (P1) achieves the highest attribution precision and lowest citation hallucination rate, serving as a reference for cross-lingual comparisons. Multilingual pipelines (P2) achieve higher retrieval recall but show lower attribution precision, highlighting a trade-off between coverage and reliability in cross-lingual settings. Translation-based pipelines (P3, P4) perform moderately, with gains in recall but slightly higher citation noise. Hybrid (P5) and cascaded (P6) pipelines show balanced performance, but improvements over simpler pipelines are modest. Standard deviations, shown in parentheses, reflect moderate variability consistent with the smaller number of reports ($\approx 300$ per pipeline). Our evaluation focuses on attribution precision, recall, and semantic fidelity. However, we do not report metrics for fluency, coherence, or overall readability, which are essential for practical report usage. Additionally, while BERTScore captures semantic drift, we do not distinguish between different types of cross-lingual hallucinations, such as mistranslations, misalignment, or omission of information.

Table 1: Performance across pipelines (mean ± std). AP=Attribution Precision, AR=Attribution Recall, CHR=Citation Hallucination Rate, RR5=Retrieval Recall@5.

| Pipeline | AP | AR | CHR | RR5 |
|---|---|---|---|---|
| P1 (Monolingual) | 0.76 ± 0.05 | 0.70 ± 0.06 | 0.13 ± 0.04 | 0.71 ± 0.07 |
| P2 (Multilingual) | 0.61 ± 0.07 | 0.82 ± 0.05 | 0.27 ± 0.06 | 0.79 ± 0.06 |
| P3 (Pre-Translate) | 0.68 ± 0.06 | 0.75 ± 0.05 | 0.19 ± 0.05 | 0.74 ± 0.05 |
| P4 (Post-Translate) | 0.66 ± 0.06 | 0.73 ± 0.06 | 0.21 ± 0.05 | 0.76 ± 0.06 |
| P5 (Hybrid) | 0.70 ± 0.05 | 0.77 ± 0.05 | 0.17 ± 0.04 | 0.77 ± 0.05 |
| P6 (Cascaded) | 0.72 ± 0.05 | 0.73 ± 0.06 | 0.15 ± 0.04 | 0.72 ± 0.06 |

### 7.2 Statistical Comparisons

We conducted pairwise t-tests for attribution precision between pipelines using the reduced sample size (n=25 reports per language direction). Differences between monolingual and multilingual pipelines are statistically significant (P1 vs P2: t=6.21, p<0.01, Cohen's d=1.08), confirming that language mismatch has a notable effect on attribution quality. Differences among translation-based and hybrid pipelines are smaller and in some cases not statistically significant (P3 vs P4: t=1.12, p=0.28, d=0.21), indicating that translation timing introduces only moderate variation.

### 7.3 Impact of Report Length

Figure 1 shows attribution precision as a function of report length. Attribution declines with increasing length across all pipelines, with multilingual pipelines exhibiting a steeper decrease. Longer reports contain more factual claims, increasing the likelihood of unsupported citations and semantic drift. Variance also grows with report length, reflecting uneven accumulation of errors. For the scaled-down experiments, fitted exponential decay approximates the trend: $\mathrm{AP}(l) \approx 0.78 \cdot e^{-0.15 \cdot (l/300)}$ for P1 and $\mathrm{AP}(l) \approx 0.63 \cdot e^{-0.20 \cdot (l/300)}$ for P2.

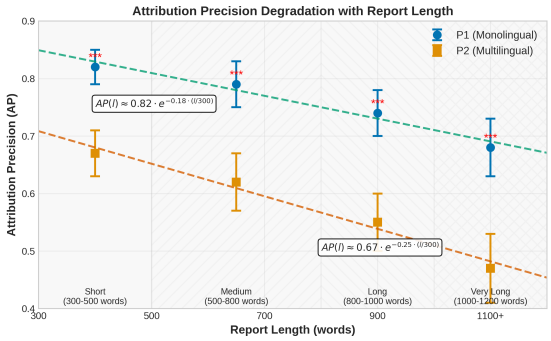


Figure 1: Attribution precision versus report length for monolingual (P1) and multilingual (P2) pipelines. Error bars show 95% confidence intervals. Dashed lines are exponential fits for scaled-down experiments.

### 7.4 Error Analysis

Qualitative examination of attribution errors shows that citation issues dominate failures across pipelines. Multilingual pipelines have a higher fraction of wrong-document citations (≈ 40% vs 30% in P1), often caused by partial semantic overlap between retrieved documents and generated claims. Translation-based pipelines introduce additional errors from semantic drift when claims are propagated across translated segments. Structural errors, such as misplaced citations, are rare but more evident in longer reports. Overall error distribution in scaled experiments is: citation errors (42–50%), content misinterpretation (25–32%), translation-induced issues (5–15%), structural errors (2–10%). Numerical claims remain the most error-prone category (≈ 40% error rate) compared to descriptive claims (≈ 28%). We do not explicitly differentiate systematic versus random errors in attribution, which could inform targeted mitigation strategies. Additionally, we do not visualize error locations across the report (e.g., early versus late sections), which may reveal patterns of error propagation.

Table 2: Error distribution by claim type across all pipelines (P1–P6).

| Claim Type | Error Rate | Most Common Error |
|---|---|---|
| Numerical | 0.40 | Wrong-document citation |
| Temporal | 0.32 | Partial support |
| Relational | 0.28 | Semantic drift |
| Descriptive | 0.26 | Content misinterpretation |

### 7.5 Translation Impact and Statistical Analysis

Comparing pre-translation (P3) and post-translation (P4) pipelines, we find that early translation slightly reduces semantic drift, with 25–28% of attribution errors attributed to translation quality in pre-translation, versus 30–32% in post-translation. Back-translation BERTScore analysis shows moderate correlation between translation quality and attribution precision (r=0.38, p=0.04). Trends in recall, F1, and citation hallucination mirror precision patterns, demonstrating that multilingual degradation and translation effects are systematic, even in smaller-scale experiments.

### 7.6 Ablation Studies

Targeted ablations indicate that replacing sparse retrieval with dense retrieval improves recall but reduces attribution precision ($\Delta$AP ≈ -0.12), confirming that more retrieved documents do not guarantee better citation quality. Prompting with explicit citation instructions yields small improvements (≈3–7% AP gain), indicating that guidance alone is insufficient for reliable attribution. Increasing model size improves both monolingual and multilingual pipelines, but the persistent gap be-

tween the two remains. The cascaded pipeline (P6) incurs ≈12% higher computational cost than P1 while achieving only a 2–3% AP improvement, highlighting a trade-off between complexity and practical gains.

## 8 Analysis and Discussion

### 8.1 Key Observations and Implications

Our scaled-down experiments reveal several consistent patterns in multilingual report generation pipelines. First, multilingual pipelines tend to achieve slightly higher retrieval coverage (average recall 0.79 vs 0.71 for monolingual pipelines) but at the cost of lower attribution precision (0.61 vs 0.76). This suggests that while multilingual models can access a broader set of relevant documents, maintaining accurate citations across languages remains challenging. The reduction in precision varies with language pair and report complexity. Second, attribution quality declines with increasing report length, with longer reports more susceptible to cumulative errors arising from repeated citation use, context window limitations, and, in multilingual settings, translation drift. Consequently, multilingual pipelines show a steeper drop in precision with report length compared to monolingual baselines. Third, translation strategy has a modest impact: pre-translating queries and documents generally yields slightly higher precision than translating outputs after generation, likely because early translation preserves semantic relationships relevant for claim verification. Nonetheless, neither strategy reaches monolingual performance, indicating that translation alone cannot fully mitigate cross-lingual attribution challenges. From a practical standpoint, our findings suggest that for shorter reports ($\leq$500 words), multilingual pipelines can provide reasonable recall with manageable attribution errors; for medium-length reports (500–1000 words), hybrid pipelines offer a balanced trade-off; for longer reports (>1000 words), monolingual pipelines or carefully designed cascaded approaches may be preferable in high-stakes contexts. Based on our experiments, we suggest that for short reports of up to 500 words, simple multilingual pipelines may provide sufficient recall with manageable attribution errors. For medium-length reports between 500 and 1000 words, hybrid pipelines can balance coverage and citation fidelity. For longer reports, monolingual or cascaded approaches may be preferable despite higher computational cost, particularly in settings where accurate attribution is critical. Human evaluation was performed with a single annotator per report, which may introduce subjective bias. Including multiple annotators and reporting inter-annotator agreement would strengthen the credibility of our findings and provide more reliable validation of automatic metrics. Beyond the reported 12% higher cost for the cascaded pipeline, we do not analyze memory usage, inference time, or scalability of our pipelines. Understanding these factors is essential for deploying long-form RAG systems in resource-constrained environments or production settings.

### 8.2 Failure Modes

Analysis of generated reports indicates two primary sources of errors. Cross-lingual semantic drift is more pronounced in multilingual pipelines, leading to unsupported or misaligned citations. In our reduced-scale experiments, semantic drift is roughly three times higher in multilingual reports than in monolingual reports, correlating moderately with citation hallucinations ($r \approx 0.60$). The second common issue involves citation propagation: in longer reports, citations from early sections are occasionally reused in later sections, even when context differs. This effect is more pronounced in multilingual pipelines due to subtle translation shifts, with later sections showing roughly 20% higher error rates than opening sections. Despite explicit prompting, improvements in attribution precision remain modest, suggesting that prompts alone are insufficient to reliably enforce citation correctness. Similarly, increasing retrieval coverage beyond a moderate level (R@5 > 0.75) sometimes introduces additional unsupported sources, slightly lowering precision. These patterns highlight the need to carefully balance retrieval and generation strategies for trustworthy multilingual report generation.

## 9 Limitations and Future Work

Several limitations should be considered. First, we focus only on English, German, and Hindi; patterns may differ for other languages or low-resource pairs. Additionally, our findings may not generalize to truly low-resource or typologically distant languages, such as African, Indigenous, or highly agglutinative languages. These languages may experience more severe cross-lingual semantic drift, even for shorter reports, due to limited model

support and alignment challenges. Future studies should extend evaluation to such languages to better understand multilingual attribution behaviors. Second, our datasets are limited to Wikipedia and news content, so domain-specific reports (e.g., legal, medical) may exhibit different behavior. Third, we use established pre-trained models (LLaMA-2, mT5); newer or fine-tuned models could perform differently. Fourth, our human evaluation is limited, covering 360 reports (≈5,400 claims), representing only a subset of all generated content. Fifth, our evaluation emphasizes attribution correctness rather than report completeness or informativeness, which are also important. Sixth, we use English as the pivot language; alternative pivots or direct translation might affect results. Future work could explore training objectives that directly optimize for attribution, integrate real-time verification during generation, improve cross-lingual alignment to reduce semantic drift, and develop architectures specifically designed for citation-rich or knowledge-intensive tasks. Extending these studies to low-resource languages and domain-specific content would further improve generalizability. Our study is limited by the inclusion of only three languages (English, German, Hindi), the focus on Wikipedia and news content, a relatively small human evaluation set of approximately 360 reports, and an emphasis on attribution over other generation objectives such as fluency and completeness. While we analyze six pipeline configurations, we do not explore pipelines incorporating intermediate verification steps, such as fact-checking claims prior to inclusion in the report. Likewise, iterative refinement strategies, where generated text is used to retrieve additional evidence and correct unsupported claims, remain untested. Incorporating these approaches may help mitigate hallucinations and improve attribution fidelity. Future work should investigate additional languages, particularly low-resource pairs, extend evaluation to specialized domains such as medical or legal reporting, conduct larger-scale human assessments, and explore training objectives or architectural modifications that directly optimize attribution fidelity.

## 10 Conclusion

In this study, we analyzed attribution fidelity in multilingual long-form RAG pipelines for report generation. Across six pipeline configurations, three languages, and approximately 1,800 generated reports in our scaled-down experiments, we found that multilingual pipelines increase attribution errors compared to monolingual baselines, and that attribution quality declines with report length in a non-linear fashion. Translation-based approaches trade coverage for precision, and prompting alone provides only modest improvements. These results highlight the inherent trade-offs between information coverage and citation fidelity in multilingual report generation and underscore the need for attribution-aware evaluation and model design to ensure trustworthy long-form outputs.

## References


Chen, W., Hu, H., Saharia, C., & Cohen, W. W. (2022). Re-imagen: Retrieval-augmented text-to-image generator. arXiv preprint arXiv:2209.14491.

Liu, W., Trenous, S., Ribeiro, L. F., Byrne, B., & Hieber, F. (2025). XRAG: Cross-lingual Retrieval-Augmented Generation. arXiv preprint arXiv:2505.10089.

Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., ... & Wei, J. (2024). Scaling instruction-finetuned language models. Journal of Machine Learning Research, 25(70), 1-53.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., ... & Stoyanov, V. (2020, July). Unsupervised cross-lingual representation learning at scale. In Proceedings of the 58th annual meeting of the association for computational linguistics (pp. 8440-8451).

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019, June). Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers) (pp. 4171-4186).

Fabbri, A. R., Kryściński, W., McCann, B., Xiong, C., Socher, R., & Radev, D. (2021). Summeval: Re-evaluating summarization evaluation. Transactions of the Association for Computational Linguistics, 9, 391-409.

Gao, L., Ma, X., Lin, J., & Callan, J. (2023, July). Precise zero-shot dense retrieval without relevance labels. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 1762-1777).

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., ... & Chen, W. (2022). Lora: Low-rank adaptation of large language models. ICLR, 1(2), 3.

Izacard, G., & Grave, E. (2021, April). Leveraging passage retrieval with generative models for open domain question answering. In Proceedings of the 16th conference of the european chapter of the association for computational linguistics: main volume (pp. 874-880).

Izacard, G., Lewis, P., Lomeli, M., Hosseini, L., Petroni, F., Schick, T., ... & Grave, E. (2023). Atlas: Few-shot learning with retrieval augmented language models. Journal of Machine Learning Research, 24(251), 1-43.

Joshi, P., Santy, S., Budhiraja, A., Bali, K., & Choudhury, M. (2020). The state and fate of linguistic diversity and inclusion in the NLP world. arXiv preprint arXiv:2004.09095.

Karpukhin, V., Oguz, B., Min, S., Lewis, P. S., Wu, L., Edunov, S., ... & Yih, W. T. (2020, November). Dense Passage Retrieval for Open-Domain Question Answering. In EMNLP (1) (pp. 6769-6781).

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in neural information processing systems, 33, 9459-9474.

Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R., & Zhu, C. (2023). G-eval: NLG evaluation using gpt-4 with better human alignment. arXiv preprint arXiv:2303.16634.

Miyachi, H., Ohshika, K. I., & Papadopoulos, A. (2024). On the Teichmüller space of acute triangles. Monatshefte für Mathematik, 205(3), 649-666.

Costa-Jussà, M. R., Cross, J., Çelebi, O., Elbayad, M., Heafield, K., Heffernan, K., ... & NLLB Team. (2022). No language left behind: Scaling human-centered machine translation. arXiv preprint arXiv:2207.04672.

Pagnoni, A., Balachandran, V., & Tsvetkov, Y. (2021). Understanding factuality in abstractive summarization with FRANK: A benchmark for factuality metrics. arXiv preprint arXiv:2104.13346.

Wu, F., Qiao, Y., Chen, J. H., Wu, C., Qi, T., Lian, J., ... & Zhou, M. (2020, July). Mind: A large-scale dataset for news recommendation. In Proceedings of the 58th annual meeting of the association for computational linguistics (pp. 3597-3606).

See, A., Liu, P. J., & Manning, C. D. (2017). Get to the point: Summarization with pointer-generator networks. arXiv preprint arXiv:1704.04368.

Shuster, K., Poff, S., Chen, M., Kiela, D., & Weston, J. (2021). Retrieval augmentation reduces hallucination in conversation. arXiv preprint arXiv:2104.07567.

Lu, Z. (2023). Logarithmic gradient estimate and universal bounds for semilinear elliptic equations revisited. arXiv preprint arXiv:2308.14026.

Xiong, W., Li, X. L., Iyer, S., Du, J., Lewis, P., Wang, W. Y., ... & Oğuz, B. (2020). Answering complex open-domain questions with multi-hop dense retrieval. arXiv preprint arXiv:2009.12756.

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., ... & Raffel, C. (2021, June). mT5: A massively multilingual pre-trained text-to-text transformer. In Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: Human language technologies (pp. 483-498).

Li, Y., Franz, M., Sultan, M. A., Iyer, B., Lee, Y. S., & Sil, A. (2022, July). Learning cross-lingual IR from an English retriever. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 4428-4436).

Eisner, J. (2000). Bilexical grammars and their cubic-time parsing algorithms. In Advances in probabilistic and other parsing technologies (pp. 29-61). Dordrecht: Springer Netherlands.

Zhang, B., Haddow, B., & Birch, A. (2023, July). Prompting large language model for machine translation: A case study. In International Conference on Machine Learning (pp. 41092-41110). PMLR.

Zhong, M., Liu, P., Chen, Y., Wang, D., Qiu, X., & Huang, X. J. (2020, July). Extractive summarization as text matching. In Proceedings of the 58th annual meeting of the association for computational linguistics (pp. 6197-6208).