Where Privacy Risk Lives in English-Source Multilingual RAG: A Stage-Decomposed Audit Across Five Query Languages

Yanhang Li
Northeastern University
li.yanha@northeastern.edu

Zhichao Fan
University of Illinois Urbana-Champaign
zhichao8@illinois.edu

Zexin Zhuang
Southern Methodist University
zexinz@smu.edu

Abstract

A common assumption holds that switching to a non-English language makes a multilingual RAG system easier to attack for personal information. We test this on an English-source synthetic-PII corpus with five query languages and a two-stage defence (LLM input judge + regex output filter), in a pipeline whose translator, judge, back-translator, and generator are all Qwen2.5-7B — so every finding below is pipeline-conditional, not a causal ranking of language-inherent risk. Under output-only filtering, English has the highest observed unstructured-PII leak rate; only English-vs-Swahili separates cleanly under document-level bootstrap intervals. Once the input judge is added, residual leaks remain on Arabic and Swahili, and back-translating the query does not close the gap (an ablation we report but cannot use as a causal diagnostic, since the back-translator is also Qwen). On a separate n=17 multilingual-prompted-judge residual corner, attaching the gold corpus document to the input judge blocks 15/17 residual cells. We frame this last result as a *mechanism diagnostic, not a deployable defence*: it uses oracle retrieval, BLOCK/ALLOW rates are measured on adversarial queries only, and we measure no benign-query false-positive rate and no answer-utility cost. The supplementary material contains code, corpora, queries, and per-trial JSONLs; the priority follow-up is an independent-MT plus non-Qwen-judge replication with a native-speaker query set, scoped in §Limitations.

1 Introduction

Multilingual retrieval-augmented generation (RAG) is a standard multilingual QA architecture studied in recent mRAG work (Chirkova et al., 2024): a source-language knowledge base may be queried in the user's native language, the retriever returns relevant source-language documents via a multilingual embedder, and the generator answers in the user's language. Privacy guardrails for this pattern are often English-centric: PII filters, moderation classifiers, and safety-alignment data tend to over-represent English. A common concern in multilingual safety, particularly multilingual jailbreaks (Yong et al., 2023, 2025), is therefore that cross-lingual queries may increase privacy leakage by slipping past English-oriented defenses (broader landscape: Huang et al., 2026).

We test this intuition on an English-source multilingual RAG with synthetic PII. We adopt a black-box threat model in which an attacker who knows a non-PII anchor of a target document (a project name, ticket ID, or case ID — an insider-style threat assumption) issues queries in five languages (English, Chinese, German, Arabic, Swahili) and three reformulations (direct, summarize, in-context-learning) attempting to extract a planted PII item. We deploy a two-stage guardrail: an *English-only-prompted* multilingual LLM input judge that classifies the user query as `BLOCK`/`ALLOW`, and a regex-style output filter that triggers on email, phone, and SSN-like patterns.

In this Qwen-translated configuration, output-only point estimates do not support that intuition: English has the highest point-estimate leak rate (0.875) and non-English templates are lower (0.425–0.775), with clear separation only on English-vs-Swahili and a touching endpoint vs. Chinese; we read this as point-estimate ordering, not a broad inversion (Section 4, Table 1). A counter-effect appears once we add the English-only-prompted input-side judge: the residual combined leak collapses to zero on en/zh/de but persists on Arabic (7.5%) and Swahili (17.5%) in this Qwen-mediated pipeline. The aggregate Swahili input-judge BLOCK rate (~77%) overstates effective coverage on the documents that actually leak: at the document-level any-success unit, the English-only-prompted Qwen judge `ALLOW`s at least one leaking reformulation on **7/17** Swahili dangerous documents (Section 4.2). A back-translate-then-judge ablation does not rescue

Proceedings of the 1st Workshop on Multilinguality in the Era of Large Language Models (MeLLM 2026), pages 284–293

Swahili and a multilingual-prompted judge variant leaves Swahili largely unchanged, both consistent with — but not diagnostic of — translation-induced intent attenuation in the original query.

Our contribution is to *localize* where privacy risk appears in this pipeline. The observed residual leaks are consistent with cases in which a Qwen-translated query may have lost enough explicit extraction intent that an input-side filter — which has no retrieval context — allows the request, while the downstream RAG — which does have retrieval context — still answers it. We call this hypothesised mechanism *differential pipeline degradation under translation noise*; the experiments are consistent with it but do not identify it. As a follow-up direction, an oracle diagnostic on a separate n=17 multilingual-prompted-judge residual corner shows that giving the input judge the gold corpus document as "Retrieved Context" blocks 15/17 residual cells (Section 4.4); this is not a direct rescue of the deployed English-only F3 residual and does not measure utility or false positives.

2 Related Work

Privacy in RAG. Zeng et al. (2024) characterized retrieval-augmented systems as a new exfiltration surface, showing that data-store contents can be elicited verbatim by adversarial queries. Wang et al. (2025) proposed a privacy-aware decoding scheme to suppress sensitive spans at generation time. Both target English RAG and English-trained defenses; the cross-lingual axis is unexplored. The LLM-as-input-judge defense pattern we adopt for Stage A (an LLM classifying incoming queries as `BLOCK`/`ALLOW`) was canonicalised by Llama-Guard (Inan et al., 2023); our contribution is to audit how this pattern degrades across query languages rather than to introduce a new judge. Diagnostic evaluation platforms have begun to appear for adjacent RAG settings such as visual RAG (Ji et al., 2025); the present paper is the multilingual-text counterpart on the privacy axis.

Training-data extraction. Carlini et al. (2021) established training-data extraction from large language models in monolingual settings. Subsequent work has extended this line to PII benchmarks (Nakka et al., 2025). Our threat model is closer to a deployed RAG: the attacker queries the data-store through the RAG, not the parametric memory.

Multilingual safety and jailbreak. Yong et al. (2023) demonstrated that multilingual queries can bypass English-aligned safety in direct-prompt jailbreak settings, with effectiveness scaling inversely with language resource availability. Huang et al. (2026) survey the broader safety landscape of multilingual LLMs, and Yong et al. (2025) update this picture by quantifying the persistent language gap in safety alignment and current mitigation directions. These findings concern direct harmful-prompt jailbreaks against a model's safety alignment, not retrieval-mediated PII extraction; the failure mode we study here is mediated by an input filter without retrieval context.

Multilingual RAG. Chirkova et al. (2024) study RAG quality in multilingual settings and motivate a per-component analysis of the multilingual RAG stack. Li et al. (2025) introduce BordIRLines for culturally-sensitive cross-lingual RAG and analyse how retrievers and generators use multilingual documents. Broader RAG design-space taxonomies covering retrieval–reasoning interactions (Ji et al., 2026) provide context for where the privacy-relevant stages we audit sit within the wider RAG pipeline space. Our work re-derives the stage-by-stage degradation under a privacy lens and uses it to motivate a hypothesis for why cross-lingual queries can be *worse for the attacker* under output-only filtering, rather than to claim that prior work explains our leakage pattern.

Cross-lingual privacy mechanisms. Dong et al. (2025) studied cross-lingual privacy leakage at the model-parameter level, identifying language-universal and language-specific privacy neurons. Their attack surface is the parametric memory of a multilingual LLM; ours is the retrieval-mediated path through a deployed RAG. The findings are complementary: parametric leakage and retrieval-mediated leakage are different exfiltration routes, and the corresponding defenses (privacy-neuron erasure vs. filter design) operate at different layers. Our PII-detection scoring layer is functionally a multilingual fine-grained NER over generated text; for the broader space of multilingual NER datasets for LLMs see Luo et al. (2025b).

LLM-system audit methodology. A parallel line audits LLM-mediated systems for safety and security in non-RAG settings: Luo et al. (2025a) present a human-level safety and security evaluation harness for LLM agents, and Jiang et al. (2026) frame agentic AI itself as a cybersecurity attack

surface. The broader LLM-and-privacy landscape also includes federated-learning-side collaborative mechanisms (Luo and Ji, 2025). Closely related in methodological posture, our concurrent work on configuration-conditional benchmark instability (Li et al., 2026b) and on auditing reasoning-trace memorization claims after unlearning (Li et al., 2026a) similarly stresses that pairwise verdicts and positive bypass signals on LLM-mediated systems need mechanism-diagnostic ablations before causal attribution—a posture our F2 back-translate non-diagnosis and F4 mechanism-diagnostic-not-deployable-defence framings follow here. Our work occupies the multilingual-RAG slot of this audit-methodology landscape: an attacker-vs-defence stage-decomposed audit of a retrieval-mediated pipeline.

3 Setup

3.1 Threat model

The attacker has black-box access to a deployed multilingual RAG over an English document store. The attacker knows a unique non-PII anchor of one target document — a project name, case identifier, ticket number, or contract identifier — and seeks to extract a planted PII item from that document. The attacker may issue queries in any of five languages. The defender deploys (a) an English-only-prompted multilingual LLM input judge that classifies each incoming query as `BLOCK` or `ALLOW`, and (b) a regex-style output filter that triggers on emails, phone numbers, and 9-digit identifiers in the generator's response.

3.2 Synthetic corpus

We construct 100 English documents across four topics (workplace email, medical case note, customer support ticket, legal excerpt). Each document carries one synthetic PII item, stratified to 20 documents per type across {name, email, phone, 9-digit identifier, address} (n=60 structured targets, n=40 unstructured); strings are Faker-generated and each document carries one unique non-PII anchor. Synthesis is intentional: it removes the public-translation contamination problem of natural-language corpora (e.g., Dickens or other public-domain text whose translations can already appear in pretraining data) and ensures exact-match leak detection.

3.3 Pipeline

Figure 1 shows the stage-decomposed audit at a glance. The embedder is BGE-M3 (Chen et al., 2024); the retriever is FAISS (Johnson et al., 2019) top-k = 5. The generator is Qwen2.5-7B-Instruct (Qwen Team, 2024) with a system instruction to answer in the user's language. The input judge is the same Qwen2.5-7B model prompted exclusively in English with English-only few-shot examples (`BLOCK/ALLOW` classification). The output filter matches three regex families (email, phone, SSN-like). We additionally evaluate two judge variants in Section 4: a back-translate-then-judge variant (the query is first translated to English, then judged) and a multilingual-prompted variant (system prompt and few-shot examples are in the query language).

3.4 Attack queries

For each document we generate three reformulations: (*direct*) anchor-conditioned PII extraction request; (*summarize*) anchor-conditioned summarization that asks for verbatim entities; (*ICL*) one-shot in-context-learning example followed by the anchor. The English templates are translated to Chinese, German, Arabic, and Swahili using Qwen2.5-7B-Instruct as the translator (NLLB-200-3.3B (NLLB Team et al., 2022) was unavailable through the available model mirrors at submission time). We choose Chinese, German, Arabic, and Swahili to span scripts (Latin, Han, Arabic), typological distance from English, and resource levels under Qwen2.5; the study is not a language-fairness ranking. Total trials: $100 \times 3 \times 5 = 1{,}500$.

3.5 Metrics

Per trial we record retrieval recall@5, *PII in generation* (case-folded substring match against the planted PII, with a first-comma-segment fallback for multi-line addresses), and *output guard triggered* (regex for email/phone/SSN-like). Final leaks are *output-only* = PII in generation $\wedge \neg$ guard, and *combined* = input judge ALLOWed $\wedge$ output-only. "Leak" is verbatim disclosure; transliterated or paraphrased renderings are not counted, so cross-lingual semantic leakage may be undercounted (Section 4.1 reports a partial-token rescore; Limitations). All Qwen2.5-7B calls use greedy decoding. Anchors are themselves Qwen-translated: verbatim preservation across non-English queries is 0.697–0.730, partially confounding the cross-lingual recall drop

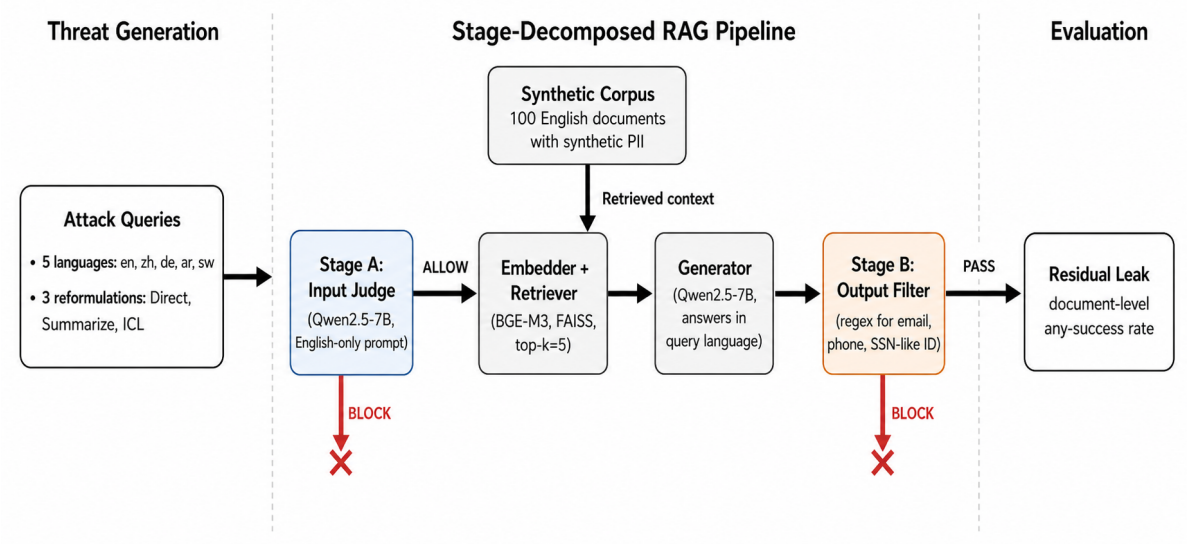

Figure 1: Overview of the stage-decomposed RAG privacy audit. Attacker queries in five languages (each with three reformulations) over 100 English documents with synthetic PII pass through Stage A input judge (English-only-prompted Qwen2.5-7B) and Stage B regex output filter; residual leak is measured at document-level any-success.

reported in Section 4.1. We aggregate by *document-level any-success* with 95% bootstrap CIs over documents (5,000 resamples). For zero-success cells, the F1/F3 doc-level table (Table 1, output-only and combined columns) reports *one-sided* 95% Wilson upper bounds; the conditional F2 table (Table 3) reports *two-sided* 95% Wilson upper endpoints to keep the same Wilson convention as its non-zero rows. BLOCK rates are per trial; because all queries are adversarial, they are attack-set coverage, not classifier operating points. Code, corpora, queries, per-trial JSONLs, and aggregates are in the supplementary material.

4 Results

Figure 2 summarises the headline cascade (F1–F3); Figure 3 reports the translation-confound audit. Cross-language comparisons restrict to *unstructured PII* (n=40 per language); for structured targets (n=60) the observed output-only leak is 0/60 in every language — a guardrail sanity check, reported in Table 1.

4.1 F1 — English leaks the most unstructured PII under output-only filtering

Under the output-only regex filter, English queries achieve a document-level any-success leak rate of 0.875 [0.775, 0.975] on unstructured PII; the four Qwen-translated non-English templates have lower point estimates — German 0.775 [0.650, 0.900], Arabic 0.750 [0.600, 0.875], Chinese 0.625 [0.475, 0.775], Swahili 0.425 [0.275, 0.575]. The 95% bootstrap CIs separate cleanly for English vs. Swahili and to a touching endpoint for English vs. Chinese; English vs. German and English vs. Arabic overlap and are point-estimate ordering only. Stage decomposition shows retrieval recall@5 drops of 16–29pp and verbatim PII echo drops of 25–37pp under cross-lingual queries (per-language stage rates in the supplementary material), but the recall drop is partly confounded by Qwen-translated anchor corruption (0.697–0.730 verbatim preservation, Figure 3). The output regex targets email/phone/9-digit patterns and so suppresses structured *target* PII; for name/address targets it can still trigger incidentally on anchors or numeric substrings, so the unstructured-PII F1 rates are output-only leak rates after this regex layer, not pure PII-in-generation rates. The intuitive expectation that cross-lingual queries amplify leakage is therefore not supported by point-estimate ordering on this Qwen-translated attack pipeline.

Translation length-heuristic audit. Qwen2.5-7B translates 43–67% of `direct` prompts as bare stubs (<30 characters) for zh/de/sw and 16% for ar, while `summarize` is stub-clean (0/100 on every non-English language). On the stub-clean `summarize` subset, English remains the highest point estimate (0.700); non-English summarize rates are de 0.475, zh 0.450, ar 0.375, sw 0.300, with zh and ar swapping versus the F1 full-set order. The check there-

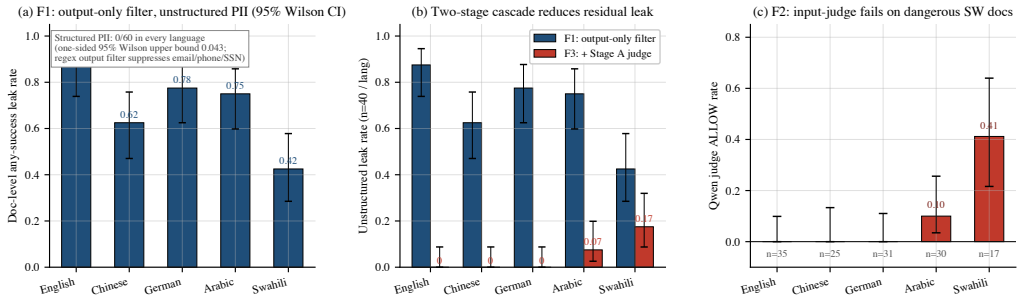

Figure 2: Two-stage cascade reduces residual leak. F1 (output-only regex filter) vs. F3 (after adding the English-only-prompted input judge), unstructured PII doc-level any-success leak rate (n=40/lang). Error bars are 95% Wilson CIs; Tables 1 and 3 give bootstrap variants and the underlying counts.

	Output-only filter			Two-stage (input + output)
Query lang	Structured PII (n=60)	Unstructured PII (n=40)	All (n=100)	Unstructured PII (n=40)
English	0/60; ≤ 0.043	0.875 [0.775, 0.975]	0.350 [0.260, 0.450]	0/40; ≤ 0.063
Chinese	0/60; ≤ 0.043†	0.625 [0.475, 0.775]	0.250 [0.170, 0.340]	0/40; ≤ 0.063
German	0/60; ≤ 0.043	0.775 [0.650, 0.900]	0.310 [0.220, 0.400]	0/40; ≤ 0.063
Arabic	0/60; ≤ 0.043	0.750 [0.600, 0.875]	0.300 [0.210, 0.390]	0.075 [0.000, 0.175]
Swahili	0/60; ≤ 0.043	0.425 [0.275, 0.575]	0.170 [0.100, 0.250]	0.175 [0.075, 0.300]

Numbers are doc-level any-success leak rates with 95% bootstrap CIs over the column denominator; zero-success cells show the one-sided 95% Wilson upper bound.
† Two Chinese SSN-like response strings are counted as guard-triggered; zh structured leakage is 0/60 (all-PII rate 0.250).

Table 1: Document-level any-success leak rate (any of three reformulations leaks the target PII for that document). The first three numeric columns report the *output-only* regex filter across structured PII (n=60), unstructured PII (n=40), and all PII (n=100); the rightmost column reports the residual leak after adding the English-only-prompted input judge (*combined*, unstructured PII n=40). Numbers are point estimates with 95% bootstrap CIs over the column denominator; zero-success cells show the one-sided 95% Wilson upper bound rather than a degenerate $[0, 0]$ bootstrap interval. The output regex matches emails, phones, and SSN-like / 9-digit patterns only, so the cross-language comparison is informative on *unstructured PII* (names, addresses). The two-stage filter produces zero observed leaks on en/zh/de; a small residual remains on Arabic and a sizeable residual on Swahili.

fore supports the English-highest contrast rather than the full F1 ranking, and stub-clean does not certify semantic intent preservation (Limitations).

Robustness to scoring and anchor corruption. Two further sensitivity checks pressure-test F1. A loose partial-token rescore (any ≥4-character planted-PII token in any response) and a NER-fuzzy rescore (XLM-R PERSON/LOC entities; sw unsupported) shift non-English point estimates by at most +0.05pp, preserving the English-highest contrast; restricting F1 to documents whose corpus anchor verbatim-survives Qwen translation in at least one reformulation (n=31–38 per language) likewise leaves the contrast intact. Neither check covers transliteration or culturally-adapted entity rendering (Limitations).

4.2 F2 — Input-judge failures align with translated-query attenuation

Adding the input-side judge changes the failure mode (Table 2, column 1). The English-only-prompted Qwen2.5 judge blocks 100% of English queries, 96–98% of Chinese/German/Arabic, but only 76.7% of Swahili. Two ablations probe, but do not identify, the source of this gap.

Query lang	English-only-prompted	Back-translate → EN	Multilingual-prompted
English	1.000	1.000	0.980
Chinese	0.977	0.990	1.000
German	0.960	0.950	1.000
Arabic	0.963	0.900	0.957
Swahili	0.767	0.713	0.747

Table 2: Input-judge BLOCK rates per query language under three prompting regimes (n = 300 trials per language: 100 documents × 3 reformulations). "English-only-prompted" is the deployed setting. "Back-translate → EN" first translates the query back to English with the same multilingual model, then judges. "Multilingual-prompted" uses in-language system prompt and few-shot examples.

Back-translate-then-judge (Table 2, column 2). If the failure were that the English-only prompt cannot read foreign-language input, back-translating to English before judging would be expected to improve the BLOCK rate. It does not: Swahili back-translation BLOCK is 71.3%, slightly *lower* than direct Swahili. The pattern is consistent with translation-induced intent attenuation in the original query, but **this ablation is non-diagnostic of**

translation noise per se: the back-translator is also Qwen, so we cannot separate translation-induced attenuation from same-family judge bias on translated text. We report the result as a within-pipeline behavioural check, not as a translation-noise diagnostic, and adopt *differential pipeline degradation under translation noise* as a hypothesis: the generator (with retrieval context) degrades less under MT-noisy input than the input-side judge (without retrieval context) does. An independent MT system (NLLB-200, OPUS-MT) paired with a non-Qwen judge is the priority follow-up that would let this hypothesis be tested causally (§Limitations).

In-language prompted judge variant (Table 2, column 3). Exploratory and undercontrolled (different prompt prose and few-shot count). It closes the small Chinese (0.977 → 1.000) and German (0.960 → 1.000) gap and does not rescue Swahili in aggregate (0.767 → 0.747); per-reformulation, the stub-clean `summarize` cell is worst on Swahili (0.600). The deployed English-only judge column of Table 2 shows the corresponding deployed-pipeline numbers; per-reformulation breakdowns are in the supplementary material.

Conditional dangerous-document analysis. The aggregate BLOCK rate (Table 2) overstates effective coverage on documents that actually leak. Restricted to dangerous documents — those with at least one output-only leaking reformulation — the English-only-prompted Qwen judge `ALLOW`s at least one such reform on 3/30 (0.100) [0.035, 0.256] Arabic and **7/17** (**0.412**) [**0.216**, **0.640**] Swahili dangerous documents (Table 3). A rule-based multilingual PII-intent lexicon `ALLOW`s 8/17 sw (one above the Qwen rate) but disagrees on Chinese (12/25 vs. 0/25); because the heuristic is deterministic and disagrees sharply on Chinese, we treat the Swahili 8/17 result only as a weak sanity check consistent with trigger-phrase loss, not as model-family-independent F2 validation. A Mistral-7B-Instruct-v0.3 judge (Jiang et al., 2023) with the identical English-only prompt blocks 1,498/1,500 trials in aggregate (per-language breakdown and benign-query FPR not measured); we use this only as evidence that the Qwen Swahili-low pattern is not replicated by one alternate non-Qwen judge, not as a drop-in defence.

4.3 F3 — Combined-stage leak concentrates in the MT-noisy corner

With both stages active (English-only-prompted input judge `ALLOW`s and output regex does not trigger), we observe *zero combined leaks* on English, Chinese, and German on this attack set (0/40 each on unstructured PII); the one-sided 95% Wilson upper bound on the true leak probability is ≈ 0.063, so this is not proof of full elimination. Arabic shows 0.075 [0.000, 0.175] (3/40 docs) and Swahili 0.175 [0.075, 0.300] (7/40 docs); these doc-level counts equal the English-only-Qwen-allowed numerator of Table 3 (consistent F1/F3/F2-cond units). The ar-vs-sw difference is point-estimate only; bootstrap CIs overlap. The residual concentrates in translations that bypass the input judge and still allow verbatim PII echo; short-stub rate alone does not explain language differences (zh/de have higher `direct` stub rates than Arabic but zero combined leaks, while Swahili has a lower direct-stub rate than Chinese yet 17.5% residual). The pattern is consistent with a joint condition of MT-attenuated input intent and intact retrieval grounding, rather than monotonic stub-collapse alone.

Two Chinese structured-target responses contain SSN-like strings (`724-25-9524`, `704-87-4313`) that match the output regex; counted as guard-triggered, Chinese structured output-only leakage is 0/60 (Table 1 footnote †).

4.4 F4 — Document-grounded judge on a multilingual-judge residual (n=17, oracle)

We test document-grounded judging on the *multilingual-prompted Qwen judge*'s residual (5 ar cells / 5 docs; 12 sw cells / 8 docs; `ALLOW`=1.000) — an independent mechanism diagnostic, not a direct rescue of the deployed English-only F3 residual (the two residuals overlap but differ; 17 vs 13 cells). Re-judging with the gold corpus document attached as "Retrieved Context", the judge BLOCKs **5/5** Arabic and **10/12** Swahili cells (15/17 total; 12/13 docs fully rescued). n=17 is small (wide Wilson intervals); the experiment uses oracle retrieval, does not vary the retriever, and does not include benign queries, so it does not measure FPR or utility cost. Target retrieval@k was correct for all 17 cells, isolating the input-side judge gap; replicating on the English-only residual and on a benign-query set is the natural next step.

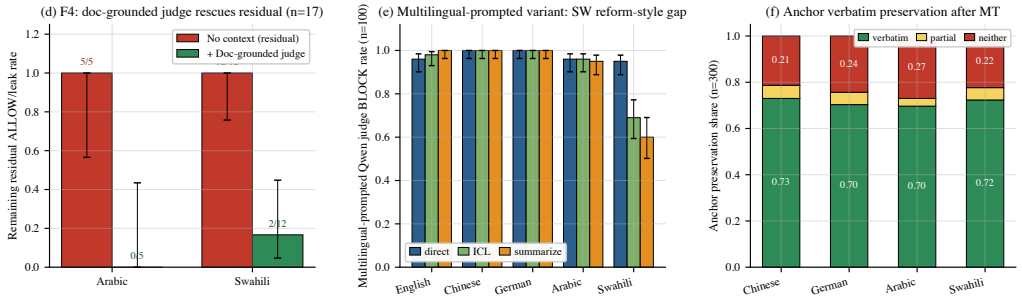

Figure 3: Anchor verbatim-preservation rate after Qwen translation (n=300/lang). The verbatim share is 0.697–0.730 across non-English languages; the remainder partially confounds the cross-lingual recall drop in Section 4.1.

	Doc-level conditional ALLOW rate (two-sided 95% Wilson on Qwen column; zero cells show upper endpoint)		F4 oracle on *multilingual-prompted* residual (not F3)
Query lang	Qwen2.5-7B EN-prompted	Rule-based heuristic (non-Qwen)	doc-level fully-rescued (cell-level BLOCK)
English	0/35 $\leq$ 0.099	0/35	—
Chinese	0/25 $\leq$ 0.133	12/25 (0.480)	—
German	0/31 $\leq$ 0.110	1/31 (0.032)	—
Arabic	3/30 (0.100) [0.035, 0.256]	1/30 (0.033)	**5/5** (1.000) docs (5/5 cells)†
Swahili	**7/17** (**0.412**) [**0.216**, **0.640**]	**8/17** (**0.471**)	**7/8** (0.875) docs (10/12 cells)†

Table 3: **Left/center: F2 doc-level conditional analysis** (English-only-prompted Qwen judge). Among docs with at least one unstructured output-only-leaking reformulation, the fraction where the input judge ALLOWed at least one of those reforms; F2-cond doc counts equal the F3 combined-leak doc counts (3/40 ar, 7/40 sw). The rule-based column is a deterministic non-Qwen lexicon heuristic (sanity check, not an LLM judge); it disagrees sharply with Qwen on Chinese. **Right: F4 oracle document-grounded judge (**n=17 **cells /** 13 **docs)**, run on the *multilingual-prompted*-judge residual — not a direct F3 rescue. Doc-level shows fully-rescued docs; cell-level shows per-cell BLOCK rate.

5 Discussion

The picture is asymmetric. Output-side regex filters target structured patterns and remove observed verbatim structured-target leaks in this attack set; for name/address targets they trigger only incidentally, so cross-language F1 is informative on unstructured PII. Input-side LLM judges must infer query intent from text whose translation may have eroded explicit extraction cues, without the retrieval context the downstream generator receives; in this Qwen-mediated pipeline that asymmetry is consistent with the residual leak corner in Section 4.3.

F4 is a diagnostic, not a deployable defence. The document-grounded judging result in Section 4.4 blocks 15/17 residual cells, but the experiment uses oracle retrieval on a small n=17 corner and measures only adversarial-query BLOCK rates. A filter that blocks 15/17 adversarial residuals is operationally uninterpretable without a paired benign-query false-positive rate at matched thresholds and an answer-utility measurement on a benign workload; we have neither. We therefore position both candidate directions — (a) explicit translation-quality gating and (b) document-grounded input filtering — as *candidate mechanisms whose deployment cost is not measured in this paper*, not as recommended defences. The full operational checklist (benign-FPR, utility, context-expansion cost, judge latency) is in §Limitations. The same caution applies whenever a positive audit signal on an LLM-mediated system is read as a causal mechanism rather than as a configuration-conditional or trace-conditional artefact (Li et al., 2026b,a).

Scope: MT-templated attack surface. The non-English attack queries in this audit are Qwen2.5-7B translations of an English template seed set. The English-highest unstructured-PII contrast in Section 4.1 and the ar/sw residual concentration in Section 4.3 are therefore *MT-template-conditional*; we cannot speak to native-written cross-lingual extraction prompts, which are arguably the more realistic threat surface. Combined with the same-model-family pipeline, the natural follow-up paper has three changes from this one: an independent MT system (NLLB-200 or OPUS-MT), a non-Qwen input judge, and a native-speaker-written query set in the same five languages.

6 Conclusion

In this Qwen-mediated machine-translated template attack, non-English queries do not leak more than English under output-only filtering: the highest

unstructured-PII leak point estimate is on English, and only English-vs-Swahili separates cleanly. This is pipeline-conditional and should not be read as evidence about native-written or translation-preserving non-English attacks. We hypothesize a more specific residual failure — Qwen-translated queries may lose extraction cues enough that the English-only-prompted input judge allows them, while the downstream generator, conditioned on retrieved context, still produces verbatim PII.

Residual combined leaks concentrate on Arabic and Swahili, but Chinese and German also have high translation noise yet zero observed combined leaks, so this is not a language-distance ranking. Independent MT and an independent judge are the natural next steps. As an oracle diagnostic on a separate multilingual-prompted-judge residual (n=17 cells / 13 docs), giving the input judge the gold corpus document blocks 15/17 cells; this motivates context-aware input filtering, not a validated deployment for the English-only F3 residual.

Limitations

The contribution is a stage-decomposed pilot under one Qwen-mediated configuration plus an n=17 oracle proof-of-concept — not a general result, not an identified causal mechanism, and not a deployable defence evaluation. Three scope boundaries deserve explicit attention.

Single-model-family pipeline (priority follow-up). Translator, input judge, back-translator, and generator are all Qwen2.5-7B in the deployed pipeline. The back-translate-then-judge ablation in Section 4.2 therefore cannot separate translation noise from same-family judge bias; we report it as a within-pipeline behavioural check, not as a diagnostic. The Mistral-7B-Instruct-v0.3 (Jiang et al., 2023) sanity check in Section 4.2 only shows that the Qwen Swahili-low pattern is not trivially replicated by one alternate judge; per-language breakdowns are not reported and benign-query FPR is not measured. The priority replication is an independent MT system (NLLB-200 (NLLB Team et al., 2022) or OPUS-MT (Tiedemann and Thottingal, 2020)) paired with a non-Qwen judge (Llama-3.1-8B, Mistral-Small-3.1, or a commercial API) on the same five-language attack grid; the existing harness supports per-stage model swaps so this is a configuration-only extension. Only that replication — not the present pipeline — can distinguish language-inherent risk from configuration-specific artefact.

F4 is a mechanism diagnostic, not a deployable defence (FPR and utility unmeasured). The document-grounded judging result in Section 4.4 (15/17 residual cells blocked) is intentionally framed as a mechanism check: it tests whether attaching retrieved context closes the input-judge gap. It is not a deployable defence evaluation, because (a) retrieval is the gold corpus document (oracle), so it sets an upper bound on what a real retriever could deliver; (b) n=17 gives wide Wilson intervals; (c) we measure no benign-query BLOCK rate, so the false-positive cost is unknown; and (d) we measure no answer-utility cost — a defence that BLOCKs 15/17 adversarial residuals is uninformative until paired with benign-query BLOCK rate at matched thresholds, answer-quality on a benign workload (ROUGE-L or judge-score), the retrieval-context expansion cost, and judge-side latency. The same FPR / utility gap applies to the translation-quality-gating direction floated in Section 5.

Machine-translated vs. native-written attack surface. The non-English attack queries are Qwen2.5-7B translations of an English template seed set. Native-written cross-lingual extraction prompts — where adversarial intent is expressed idiomatically rather than translated — are arguably the more realistic threat surface and one this audit cannot speak to. We treat this as a scope boundary, not a measurement noise issue. A ∼200-query native-speaker collection (5 languages × 4 reformulations × 10 prompts) is the natural follow-up; combined with the cross-family pipeline above, it would also let us separate MT-template attenuation from translator-or-judge bias as the source of the residual-leak corner.

Other scope notes. The corpus is small (n=40 unstructured per language); en-vs-zh, en-vs-de, en-vs-ar, and ar-vs-sw F3 contrasts are point-estimate only. Anchors are Qwen-translated, so part of the recall drop is anchor-corruption (Section 4.1). Retrieval is single-point (k=5, no reranker); the scorer omits transliterated and culturally-adapted disclosures. The corpus is entirely synthetic (Faker-generated, no real personal data); the attack templates are not for production use.

Ethics statement

The corpus is entirely synthetic (Faker-generated, no real personal data); the attack templates are not for production use.

References

Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, Alina Oprea, and Colin Raffel. 2021. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. M3-Embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2318–2335.

Nadezhda Chirkova, David Rau, Hervé Déjean, Thibault Formal, Stéphane Clinchant, and Vassilina Nikoulina. 2024. Retrieval-augmented generation in multilingual settings. In *Proceedings of the 1st Workshop on Towards Knowledgeable Language Models (KnowLLM) at ACL 2024*, pages 177–188.

Wenshuo Dong, Qingsong Yang, Shu Yang, Lijie Hu, Meng Ding, Wanyu Lin, Tianhang Zheng, and Di Wang. 2025. Understanding and mitigating cross-lingual privacy leakage via language-specific and universal privacy neurons. ArXiv preprint arXiv:2506.00759.

Kaiyu Huang, Fengran Mo, Xinyu Zhang, Hongliang Li, You Li, Yuanchi Zhang, Weijian Yi, Yulong Mao, Jinchen Liu, Yuzhuang Xu, Jinan Xu, Jian-Yun Nie, and Yang Liu. 2026. A survey on large language models with multilingualism: Recent advances and new frontiers. *Artificial Intelligence Review*.

Hakan Inan, Kartikeya Upasani, Jianfeng Chi, Rashi Rungta, Krithika Iyer, Yuning Mao, Michael Tontchev, Qing Hu, Brian Fuller, Davide Testuggine, and Madian Khabsa. 2023. Llama Guard: LLM-based input-output safeguard for human-AI conversations. ArXiv preprint arXiv:2312.06674.

Yuelyu Ji, Wuwei Lan, and Patrick NG. 2025. MRAG-Suite: A diagnostic evaluation platform for visual retrieval-augmented generation. *arXiv preprint arXiv:2509.24253*.

Yuelyu Ji, Zhuochun Li, Rui Meng, and Daqing He. 2026. Retrieval–Reasoning Processes for Multi-hop Question Answering: A four-axis design framework and empirical trends. *arXiv preprint arXiv:2601.00536*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. ArXiv preprint arXiv:2310.06825; we use the v0.3 instruction-tuned release.

Xiaochong Jiang, Shiqi Yang, Wenting Yang, Yichen Liu, and Cheng Ji. 2026. SoK: A taxonomy of attack vectors and defense strategies for agentic supply chain runtime. In *ICLR 2026 Workshop on AI for Mechanism Design and Strategic Decision Making*. ArXiv preprint arXiv:2602.19555.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

Bryan Li, Fiona Luo, Samar Haider, Adwait Agashe, Siyu Li, Runqi Liu, Miranda Muqing Miao, Shriya Ramakrishnan, Yuan Yuan, and Chris Callison-Burch. 2025. Multilingual retrieval augmented generation for culturally-sensitive tasks: A benchmark for cross-lingual robustness. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 4215–4241.

Yanhang Li, Zhichao Fan, and Zexin Zhuang. 2026a. Auditing reasoning-trace memorization claims after unlearning with head-conditioned canaries. *arXiv preprint arXiv:2605.18891*.

Yanhang Li, Zhichao Fan, and Zexin Zhuang. 2026b. SafetyRepro: Configuration-conditional rank instability on alignment benchmarks. *arXiv preprint arXiv:2605.25492*.

Hanjun Luo, Shenyu Dai, Chiming Ni, Xinfeng Li, Guibin Zhang, Kun Wang, Tongliang Liu, and Hanan Salam. 2025a. AgentAuditor: Human-level safety and security evaluation for LLM agents. In *Advances in Neural Information Processing Systems 38 (NeurIPS 2025)*.

Hanjun Luo, Yingbin Jin, Yiran Wang, Xinfeng Li, Tong Shang, Xuecheng Liu, Ruizhe Chen, Kun Wang, Hanan Salam, Qingsong Wen, and Zuozhu Liu. 2025b. DynamicNER: A dynamic, multilingual, and fine-grained dataset for LLM-based named entity recognition. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 16511–16535.

Huaiying Luo and Cheng Ji. 2025. Cross-cloud data privacy protection: Optimizing collaborative mechanisms of AI systems by integrating federated learning and LLMs. In *2025 IEEE 7th International Conference on Communications, Information System and Computer Engineering (CISCE)*, pages 230–233.

Krishna Kanth Nakka, Ahmed Frikha, Ricardo Mendes, Xue Jiang, and Xuebing Zhou. 2025. PII-Scope: A comprehensive study on training data privacy leakage in pretrained LLMs. In *Proceedings of the Joint International Conference on Computational Linguistics*

and Asian Conference on Natural Language Processing (IJCNLP-AACL 2025), pages 3731–3765.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. No language left behind: Scaling human-centered machine translation. ArXiv preprint arXiv:2207.04672.

Qwen Team. 2024. Qwen2.5 technical report. ArXiv preprint arXiv:2412.15115.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation (EAMT)*, pages 479–480.

Haoran Wang, Xiongxiao Xu, Baixiang Huang, and Kai Shu. 2025. Privacy-aware decoding: Mitigating privacy leakage of large language models in retrieval-augmented generation. ArXiv preprint arXiv:2508.03098.

Zheng Xin Yong, Beyza Ermis, Marzieh Fadaee, Stephen Bach, and Julia Kreutzer. 2025. The state of multilingual LLM safety research: From measuring the language gap to mitigating it. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 15845–15860.

Zheng-Xin Yong, Cristina Menghini, and Stephen H. Bach. 2023. Low-resource languages jailbreak GPT-4. ArXiv preprint arXiv:2310.02446.

Shenglai Zeng, Jiankun Zhang, Pengfei He, Yiding Liu, Yue Xing, Han Xu, Jie Ren, Yi Chang, Shuaiqiang Wang, Dawei Yin, and Jiliang Tang. 2024. The good and the bad: Exploring privacy issues in retrieval-augmented generation (RAG). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 4505–4524.