

Generator-Guided Amount Recovery for Voice-Based Financial Record-Keeping in Mooré–French Code-Switched Speech

Maimouna Ouattara
University of Luxembourg

El-Hacen Diallo
University of Luxembourg

Fred Philippy
University of Luxembourg

Abdoul Kader Kaboré
University of Luxembourg

Jacques Klein
University of Luxembourg

Tegawendé F. Bissyandé
University of Luxembourg

{maimouna.ouattara, el-hacen.diallo, fred.philippy,
abdoulkader.kabore, jacques.klein, tegawende.bissyande}@uni.lu

Abstract

Voice interfaces promise to bring digital financial services to the roughly 800 million adults worldwide who cannot read, yet informal markets demand more than broad speech-recognition coverage. In Burkina Faso, traders record credit in Mooré and in French–Mooré code-switched speech, where the value of a spoken amount depends on the active counting convention, a historical base-five *wakir* system in Mooré against base ten in French, so that a correctly transcribed numeral can still encode the wrong sum, and standard cascades and frontier audio Large Language Models (LLMs) fail even when the convention is supplied in context. We present a voice-to-ledger pipeline that recovers the customer, the transaction direction, and the monetary amount from naturalistic market speech. Its core is Generator-Guided Analysis-by-Synthesis (GG-Abs), a deterministic amount-recovery module that treats each spoken amount as a latent integer. GG-Abs proposes candidates with a bidirectional Mooré numeral generator, renders each phonetically, keeps only those that re-synthesise to the noisy transcript, and abstains otherwise, so every accepted amount carries an explicit reconstruction; a conservative LLM judge supplies the customer and direction. We also release a voice-ledger corpus of Mooré–French market recordings spanning four credit intents. On a sealed real-speech holdout, the pipeline attains **76.0%** amount accuracy on Mooré and **85.0%** on code-switched speech, with direction accuracy above 90% in both regimes, improving Mooré amount recovery over Automatic Speech Recognition (ASR)→LLM and audio-native LLM baselines by 16 to 47 points.¹

1 Introduction

Recent advances in multilingual AI now reach hundreds of under-represented languages, spanning

¹Code is available at <https://github.com/mouna80/latell2026-moore-french-cs>.

translation models such as No Language Left Behind (NLLB) (Costa-Jussà et al., 2022) and speech recognition systems such as Massively Multilingual Speech (MMS) (Pratap et al., 2024) and Omnilingual ASR (Keren et al., 2025). Yet broader coverage alone does not guarantee *inclusive* access for the millions of people who cannot read. UNESCO estimates that 739 million adults still lack basic literacy skills, including 225 million in sub-Saharan Africa (UNESCO, 2025). This challenge is particularly acute in countries like Burkina Faso, where the adult literacy rate stood at about 42% in 2023 (World Bank, 2023). Voice-first services promise to lift this barrier, replacing written input with speech and opening digital access in the languages people actually speak (Medhi Thies, 2015; da Silva et al., 2024).

However, these barriers are, in practice, contextual and cultural. Everyday communication is shaped by social and cultural conventions that demand particular attention. One of the most common is *code-switching*, the mixing of two languages within a single utterance. In Burkinabè markets, for instance, traders fluidly alternate between French and Mooré (Myers-Scotton, 1997; Amuzu and Singler, 2014). A historical artefact of the CFA franc ² makes this code-switching especially treacherous. An amount spoken in French denotes its exact base-ten value in CFA. The same amount spoken in Mooré, however, follows a base-five system anchored on the *wakir*, a five-franc unit inherited from colonial currency: it is counted in units of five francs, not one. A trader who says “twenty” in Mooré therefore means one hundred CFA. This fivefold gap is invisible to anyone pars-

²CFA refers to a currency used in several West and Central African countries. In this paper, it specifically denotes the West African CFA franc (ISO 4217: XOF), distinct from the Central African CFA franc (ISO 4217: XAF). See https://en.wikipedia.org/wiki/CFA_franc.

ing the words alone. In voice-first financial applications, such a discrepancy becomes a critical source of error. The system must accurately identify the language of each amount to determine its true value.

```
[TASK] Audio-to-Text Transcription &
      Financial Normalization

[INPUT_AUDIO] "<Audio clip: a trader
              saying 'Moussa [first name] owes me
              2,000 [XOF]' in moore>"

[EXPECTED_VALUE] .....2,000 XOF
[SYSTEM_OUTPUT] .....400 XOF
```

Listing 1: Example of an audio-native LLM transcription failure³. The system maps the acoustics literally, masking the fivefold currency gap.

Our contributions:

- **GG-Abs: Generator-Guided Analysis-by-Synthesis.** To recover Mooré amounts, this framework treats each value as a latent integer, proposes candidates using a bidirectional Mooré numeral generator, and scores them via phonetic re-synthesis against unfused transcripts. Requiring no LLM call, this deterministic method recovers 72.5% of Mooré amounts, improving recall by 10.0 points over parse-only (cf., 5). When used as a hint for `gemini-3.1-flash-lite` (approximately one-sixth the estimated judge-only API cost of `gemini-3.5-flash`), it improves Mooré amount recovery from 53.1% to 70.0%.
- **A Mooré–French voice-to-ledger pipeline for informal credit.** Built on GG-Abs, the pipeline recovers the customer, transaction direction, and amount from monolingual Mooré as well as French–Mooré code-switched speech, without requiring explicit numeral-language diarization within this evaluated language pair.
- **A Mooré–French voice-ledger corpus and rigorous evaluation.** We construct and release a task-specific corpus covering four credit intents, built in three phases, from seed recordings through template-based synthesis to a human-validated gold set, and evaluate on a sealed real-speech holdout. GG-Abs is the only system with strong performance across

both regimes (76.0%/85.0% for Mooré/code-switched amounts, and about 90% for direction in each), improving Mooré amount recovery over the baselines by 16 to 47 points.

2 Related Work

Code-switched speech and cascade pipelines.

Production ASR still handles code-switching poorly. Multilingual training on English-Bantu pairs in South Africa beats monolingual baselines (Yilmaz et al., 2018; Biswas et al., 2019, 2020), and self-supervised representations help under-resourced African languages (Ögúnremí et al., 2023), but most results come from controlled, single-pair settings. Surveys of end-to-end code-switching ASR (Agro et al., 2025; Mustafa et al., 2022) show that Mandarin-English and Hindi-English dominate the literature, with little coverage of African pairs. Turning such speech into structured output typically relies on cascades. ASR followed by machine translation (ASR→Machine Translation (MT)) is the de-facto approach for low-resource speech translation, including African languages and French targets (Kponou et al., 2025), while ASR followed by a language model (ASR→LLM) feeds transcripts to an LLM for correction or understanding (Chen et al., 2023). Beyond standard error propagation, these cascades miss cultural context shifts like the base-five Mooré counting system (Naveen and Trojovský, 2024; Al Sharoufi and Al-Fadhli, 2025), which leaves the numeral correctly transcribed but wrongly valued.

Language identification and diarization. To handle code-switched amounts, a natural approach is to identify the language of each spoken numeral and apply the corresponding counting convention. In audio, this requires spoken language diarization to segment utterances by language. While recent diarization systems show progress (Chua et al., 2023; Dhanush et al., 2025; Baroudi et al., 2026), they primarily target high-resource pairs and require abundant training data. Text-side language identification is no easier at scale (Burchell et al., 2024), and French-African pairs remain barely studied, so this route is not yet viable in our context.

Task-oriented dialogue. While widely studied for high-resource languages and financial domains (Tur and De Mori, 2011; Altinok, 2018), dialogue systems for African languages are only just emerging through datasets like WAXAL (Diack et al.,

³<https://gemini.google.com/share/2ba4289a0b13>

2026) and AfriWOZ (Adewumi et al., 2023). Although cross-lingual transfer helps bridge these gaps (Thangaraj et al., 2024), in-domain financial corpora remain scarce, and no existing system targets a structured financial ledger from code-switched market speech. The output is a structured ledger comprising the amount, direction, and customer. Recent African benchmarks such as INJONGO (Yu et al., 2025) include banking intents across multiple languages, but target text-based intent classification rather than speech-based, value-normalized ledger extraction.

Accessibility and financial inclusion. Financial inclusion depends on low-cost services reaching unbanked populations (Kunjal, 2026). Yet text-centric banking in dominant languages imposes severe literacy barriers on rural communities (Chatterjee and Das, 2021), limiting benefits even where mobile access exists (Ozili, 2025). FinTech providers respond with simplified interfaces (Koefer et al., 2024), and mobile connectivity expands inclusion in West Africa (Senou et al., 2019), but infrastructure alone is insufficient. This literature motivates voice as an accessible interface capable of relaxing literacy constraints (Kunjal, 2026), yet it stops at design and leaves open the technical challenge of recovering correct monetary values from code-switched speech. GG-Abs addresses this gap.

3 Methodology

Figure 1 illustrates the proposed pipeline. It operates in two phases. An offline phase derives correction rules by analyzing ASR drift (Sections 3.1 and 3.2), and an online phase applies these rules to reconstruct amounts from noisy transcripts (Section 3.3).

3.1 Offline: Synthetic Corpus

No existing corpus covers Mooré–French code-switched speech for informal credit ledgers. We built a task-specific corpus in two steps.

Step 1: Intent definition. We fix the task scope with 20 development intents spanning four types, six debt increases, six repayments, five combined multi-transaction cases, and three balance queries, each paired with a structured XOF gold ledger (debt_up, debt_down, queries). Speakers produced free Mooré and Mooré–French versions with French numerals, establishing the schema and linguistic regimes rather than a benchmark.

Step 2: Synthetic augmentation. Bilingual contributors mapped French ledger frames to Mooré templates with entity, product, and amount slots. A generator then fills these slots with regional names, market roles, products, and CFA amounts across the four intent types. Gold JSON ledgers are derived from the source slots and deterministic amount decoding, never from ASR output, yielding 4,570 utterances rendered with the fine-tuned XTTS-v2 Mooré voice (Ouedraogo et al., 2026), which we use as a compositional stress test.

3.2 Offline: ASR drift analysis

Mooré counts money in *wakir* (units of five CFA) rather than single francs. For example, *kobs-naase* means 100×4 *wakir* (2,000 CFA), and *tus pis-nu* means $1,000 \times 50$ *wakir* (250,000 CFA). To invert this convention, we implement a generator-parser pair $G(n) \leftrightarrow \text{parse}(t)$, verified to round-trip ($\text{parse}(G(n)) = n$) up to 10,000,000 CFA. The generator proposes canonical numeral candidates, which are matched against the transcript to recover the amount. This works perfectly on canonical Mooré numerals but fails on raw ASR output. Noise introduces a systematic drift between canonical forms and surface transcripts, corrupting numerals into unparseable fragments or near-homophones.

This drift is patterned rather than random. It exhibits recurring distortions such as compound-boundary loss (*pis-nu* split), final nasal drift (*pisnii* for *pis-nii*), vowel neutralization, and digit fall-back (*tus 14*). Because the corruptions follow predictable patterns, a separate normalization phase reverses them with deterministic rules before matching, mapping the noisy surface form back to a canonical one. The generator itself remains untouched, proposing candidates only, while the rules clean the incoming transcript.

Rule mining occurs strictly offline. The deterministic GG-Abs module makes no LLM call at inference. We use three foundation models (GPT-5, Gemini Pro, Claude Opus 4.8) solely to analyze the drift and propose collision-safe rewrites (i.e., rewrites that never map two distinct amounts to the same form). They are supplied with the canonical numeral lexicon, clean utterances, and paired noisy transcripts. Model consensus ($\geq 2/3$) serves merely as a pre-filter. The definitive gate is a one-rule ablation on the development split, where a rule is kept only if it raises recall or reduces false hints

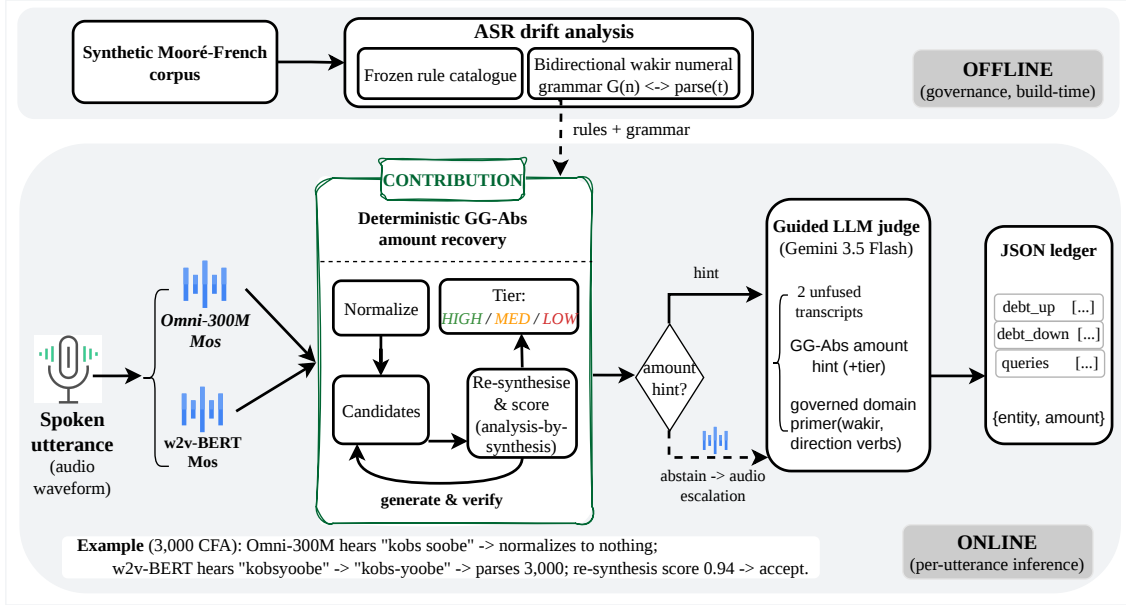


Figure 1: The GG-Abs voice-to-ledger pipeline. *Offline*, we study ASR drift on a synthetic Mooré–French corpus and freeze a rule catalogue and a bidirectional *wakir* numeral grammar $G(n) \leftrightarrow \text{parse}(t)$. *Online*, two unfused ASR channels feed the deterministic GG-Abs module, which normalises each transcript, enumerates candidates, scores them by re-synthesis, and assigns a confidence tier. A candidate amount yields a tagged hint, and the absence of any round-tripping candidate yields abstention. The guided judge (Gemini 3.5 Flash) reads the two transcripts, the hint with its tier, and a governed domain primer, and emits the ledger. On abstention the raw audio is escalated to the judge. The deterministic module (our contribution) recovers the amount without an LLM.

without adding misses, and retired otherwise. All mining and ablation rely exclusively on the development split, never the sealed evaluation holdout. In parallel, an expert track adds rules grounded in descriptive Mooré literature (e.g., discourse-marker strips for *samde* and *wakir*), so every deployed rule carries a precise provenance tag (Table 1).

Rule (prov.)	Drift	→	Canonical
<i>Expert</i>			
<i>samde</i> strip (E)	<i>samde</i> / <i>samda</i>	→	∅
<i>wakir</i> strip (E)	<i>wakir</i>	→	∅ (keep <i>tus</i>)
<i>Heuristic (k/3 consensus)</i>			
final-nii (3/3)	<i>pisnii</i>	→	<i>pis-nii</i>
pis-X (3/3)	<i>pisnaase</i>	→	<i>pis-naase</i>
thousand (3/3)	<i>piir</i>	→	<i>piig</i>
yoobe (3/3)	<i>yogbe</i>	→	<i>yoobe</i>
yopoe (3/3)	<i>yokoe</i>	→	<i>yopoe</i>
final-u (2/3)	<i>nung</i>	→	<i>nu</i>
teen (3/3)	<i>tus 14</i>	→	<i>tus piig la a naase</i>

Table 1: Frozen rules that rewrite recurring ASR drift of Mooré numerals back to the canonical form the generator parses. E marks expert, linguist-validated rules and *k/3* marks rules proposed by *k* of three LLMs.

System robustness is further reinforced by architectural diversity. We analyze drift through two

distinct Mooré systems, Omnilingual ASR and a fine-tuned w2v-BERT ASR (Wav2Vec-BERT 2.0, a multilingual self-supervised model adapted to Mooré through supervised fine-tuning on a dedicated speech corpus (Ouedraogo et al., 2026)). The two systems exhibit complementary failure modes. For 3,000 CFA (*kobs-yoobe*), w2v-BERT elides the boundary as *kobsyoobe*, whereas Omnilingual splits it and drifts the vowel as *kobs soobe*. Both transcripts are normalized by the rules and matched against the generator’s candidates without fusing the channels. Because their error distributions differ, a correct amount routinely survives even when one channel corrupts it.

3.3 Online Inference

neural ASR front ends. Their transcripts then enter the deterministic post-ASR amount-recovery path described in Algorithm 1, before the downstream LLM judge.

At inference, each utterance is processed by two frozen Mooré-targeted ASR front ends, W2V-mos and Omni-mos. Their unfused transcripts enter the deterministic GG-Abs amount-recovery path described in Algorithm 1 and are subsequently passed

Algorithm 1 GG-Abs inference for one utterance

Require: audio x of one utterance

Ensure: ledger ℓ , or abstain

```
1:  $T \leftarrow \text{ASR}(x)$  ▷ two unfused channels
2:  $T \leftarrow \text{NORMALISE}(T)$  ▷ frozen Mooré rules
3:  $C \leftarrow \text{CANDIDATES}(T)$  ▷ runs and windowed spans
4: if  $C$  has no complete run then
5:   return abstain; attach audio  $x$ 
6: end if
7: for all  $c \in C$  do
8:    $s(c) \leftarrow \text{MATCH}(G(c), T)$  ▷ re-synthesis score
9: end for
10:  $\hat{a} \leftarrow \arg \max_{c \in C} s(c)$ 
11:  $\tau \leftarrow \text{TIER}(\hat{a}, T)$  ▷ hint confidence
12:  $\ell \leftarrow \text{JUDGE}(T, \hat{a}, \tau)$ 
13: return  $\ell$  ▷ customer, direction, amount
```

to the downstream judge together with the resulting hint. When GG-Abs abstains, raw audio is additionally attached. The system decodes the audio, normalizes the transcripts, enumerates candidate amounts, verifies them via phonetic re-synthesis, reconciles readings across channels, and finally passes a tagged hint to the judge. When GG-Abs emits a Mooré amount hint, that hint is constructed and round-tripped by the deterministic layer. When GG-Abs abstains, the downstream judge may instead infer an amount from the ASR transcripts and raw audio.

The normalizer applies the frozen rules from Section 3.2 in four steps, running character normalization, then skeleton-based fuzzy matching ($\tau=0.80$), compound repair behind magnitude prefixes ($\tau=0.65$), and sequence cleanup. Fuzzy matching relies on a normalized Levenshtein similarity over pronunciation-stable skeletons, identical to the re-synthesis scoring metric. The pipeline operates with zero learned parameters, and both match thresholds are deterministic constants fixed on the development set. Normalized transcripts are grouped into maximal contiguous runs. A run is valid only if it pairs a magnitude with a quantifier or explicit digits, and a bare unit triggers abstention. To bound the otherwise $O(n^2)$ search, candidates are parseable windows (≤ 12 tokens) within each channel. Each candidate a is scored by re-synthesis,

$$s(a) = \max_c \text{SIM}(\text{skel}(G(a)), \text{skel}(t_c)),$$

where $G(a)$ is the generated spelling, $\text{skel}(\cdot)$ extracts a pronunciation-stable skeleton by folding diacritics and vowel length, and SIM is the normalized Levenshtein similarity (1 minus edit distance over maximum length). Phonetically identi-

cal spellings therefore score equally, so the candidate that best matches the acoustic transcript wins. On the 3,000 CFA example, the correct candidate (*kobs-yoobe*) scores 0.94 against the w2v-BERT skeleton, far above distractors such as 300 (0.63), 600 (0.47), and 60,000 (0.21). The same score resolves merge-versus-split ambiguity, since merged and split renderings compete directly and ties default to the split to preserve genuine compound transactions.

Per-channel readings are then reconciled slot by slot. An amount split across channels is merged only when the merge raises the overall re-synthesis score, which blocks spurious combinations, while single-channel recoveries are corroborated by the windowed enumeration. A stage-attribution audit shows that $\sim 90\%$ of multi-transaction failures are upstream, where the second amount never forms a complete run, so reconciliation is not the bottleneck and these repairs close the rest. Each surviving hint is tiered HIGH when two channels agree and MED for a single channel. Fragile hints are downgraded rather than deleted, so the judge is never deprived of a valid Mooré amount.

4 Experiments

We evaluate GG-Abs on spontaneous market speech, focusing on structured ledger recovery. The full task requires extracting transaction direction and customer identity, but amount recovery remains the primary stress test, compounding ASR noise with non-decimal counting conventions. Systems output a constrained JSON ledger scored against gold records. Amount and direction use exact-set match, while customer extraction is scored by identity equivalence rather than exact string matching, using human labels for Mooré.

To prevent leakage, we enforce a strict three-split regime. All deterministic rules, match thresholds, and LLM prompts were developed and frozen exclusively on a *synthetic dev* pool (Section 3). The end-to-end system was then evaluated exactly once on a *sealed real holdout* of 312 utterances (194 Mooré, 118 code-switched) from seven disjoint speakers, never touched in development, tuning, or calibration. The holdout introduces 60 novel intents, covering debt increases, repayments, multi-transaction scenarios, and balance queries, all distinct from the 20 development intents, so no evaluation data influenced any rule, threshold, or prompt. Table 2 details the holdout subsets used per met-

ric. For joint evaluation we report *transaction-core* (correct amount and direction) and *transaction-exact* (correct amount, direction, and customer). To isolate judge variance, all final scores are the mean $\pm$ std over three runs at temperature 0.1.

Mooré holdout subset	n	Used for
Amount-bearing	160	amount, core, exact
balance query	22	direction (CHECK)
missing-amount	12	no-amount control
No-amount control (22+12)	34	false-insertion rate
All Mooré	194	direction
Code-switch (CS)	118	CS amount and direction

Table 2: Real-speech holdout dataset ($n = 312$).

5 Results

Table 3 reports the sealed real-speech test. The two single-evidence baselines fail in opposite regimes, ASR $\rightarrow$ LLM on code-switching and the audio-native LLM on Mooré and direction. GG-Abs is the only balanced system, so it is not a one-regime amount recogniser but a voice-to-ledger pipeline that keeps the critical fields high in both regimes that break single-evidence systems. For fairness, the LLM baselines in Table 3 receive the same domain-context block as GG-Abs (the *wakir* convention, ledger semantics, and direction cues), so they are evaluated with task knowledge. Amount accuracy is computed only on the 160 of 194 Mooré utterances that carry an amount, the remaining 34 being balance queries or missing-amount cases (Table 2).

System	Mooré ($n=160$)		Code-switch ($n=118$)	
	Amount	Dir.	Amount	Dir.
ASR $\rightarrow$ MT	6.9	—	4.2	—
ASR $\rightarrow$ LLM	60.2	73.5	27.1	72.9
Audio $\rightarrow$ LLM	28.5	58.1	90.1	57.6
GG-Abs (full)	76.0	90.2	85.0	92.7

Table 3: Amount and direction exact-set accuracy (%), averaged over three runs. The audio-native baseline is higher on code-switched amount recovery (90.1 versus 85.0), whereas GG-Abs is higher on the other three reported fields.

Cross-regime performance. The 85.0 versus 90.1 code-switch amount gap is about six utterances on $n=118$, roughly three rows where a spurious Mooré hint blocks the rescuing audio escalation (fixable) and roughly three of genuine residual

(Section B.1). It is a wrinkle in one field, not a hole in the main claim, since GG-Abs wins code-switch direction decisively (92.7 vs. 57.6) and is the better full-ledger system on code-switching. The fixable rows point to a concrete next step, which we leave for future work to keep the sealed holdout scored once. Escalating to audio even when the deterministic module emits a low-confidence (MED) Mooré hint alongside French numeral evidence would stop a spurious hint from suppressing the rescue, and all three blocking rows here carry a MED hint, so this rule would target exactly them.

Mooré amount recovery before the judge. The defining property of GG-Abs is that the critical balance-changing field is recovered *before* the LLM judge. On amount-bearing Mooré utterances it emits a hint in 93.8% of cases (77.3% precision, 72.5% recall). The deterministic post-ASR module requires no additional LLM call. Emitted amounts are structurally valid and traceable. Neither deterministic variant emits an amount on the 34 no-amount controls, while carrying an explicit *wakir* reconstruction and a re-synthesis score that traces the value back to the recognized transcript rather than taking it on trust. Recovery stays consistent across transaction types and does not degrade on compound utterances (Table 4), reaching 69.3% on debt increases, 71.1% on repayments, and 80.0% on combined multi-transaction scenarios, which require every spoken amount to match. The judge therefore receives evidence graded by confidence tier (87.5% accuracy for HIGH two-channel hints, 71.3% for MED single-channel hints) rather than blind guesses.

Intent type	Mooré amount (%)	n
Debt increase (credit)	69.3	75
Repayment	71.1	45
Combined (multi-transaction)	80.0	40
All	72.5	160

Table 4: Deterministic GG-Abs Mooré amount recovery by intent type (exact-set match, $n=160$ amount-bearing utterances, no LLM). Performance stays consistent across types and is strongest on compound multi-transaction utterances.

GG-Abs secures the balance-changing fields, amount and direction, in both regimes. Direction reaches 90.2% on Mooré and 92.7% on code-switched (CS) speech, and transaction-core accuracy (amount and direction jointly correct) is 73.8%

on Mooré and 82.2% on CS, with a single IN/OUT confusion across the 132 single-direction Mooré utterances. The free-text customer name, left to the LLM judge, is the one open field, scored by identity rather than string match (e.g. *moussa* and *musa*) through a human same-person judgment. Strict transaction-exact accuracy (all three fields correct) is therefore 49.4% on Mooré and 50.8% on CS.

Complementary Mooré ASR channels. On Mooré, the deterministic amount module reads two complementary ASR channels, a fine-tuned w2v-BERT Mooré specialist and Omnilingual decoded Mooré-target (Omni-Mooré). Table 5 gives each channel’s effect on the deterministic Mooré amount (exact-set match on the 160 amount-bearing rows, no LLM). The two channels are strongly complementary. The w2v-BERT specialist is the workhorse, and dropping it collapses amount recovery from 72.5 to 38.8 (−33.7), while Omni-Mooré rescues amounts it mishears (−8.1 amount, 10 points of coverage). Keeping the two transcripts unfused lets either channel rescue an amount the other loses, rather than letting one channel’s error contaminate the pool. GG-Abs is therefore not a single recogniser followed by a parser but a generator-guided recovery layer over complementary ASR views.

Channel configuration	Mooré amount (%)	Δ
Full	72.5	—
— w2v-BERT Mooré (w2v)	38.8	−33.7
— Omni-Mooré (omni_mos)	64.4	−8.1

Table 5: Per-channel effect on the deterministic GG-Abs Mooré amount (exact-set match, $n=160$ amount-bearing, no LLM). The two Mooré ASR channels are complementary.

Cheap-judge viability. In constrained deployments, pairing the amount hint with a lightweight judge lifts Mooré amount accuracy from 53.1% to 70.0%, bridging the gap to larger models (Table 6). With a strong judge the hint instead trades a slight accuracy dip (76.0% vs. 78.8%) for strict auditability, since the model already exploits the raw ASR transcripts effectively. Direction resolution stays judge-dependent across all configurations.

The deterministic hint incurs no API cost and negligible CPU latency, so end-to-end latency is dictated by the ASR front-end and the LLM call. On an RTX 6000 Ada, single-utterance ASR latencies (0.30 s for the Omnilingual pass and 0.08 s

for w2v-BERT), and the LLM brings the median end-to-end latency to 1.9 s for the strong judge and 1.5 s for the lightweight one. Switching to the smaller judge therefore speeds up the pipeline and cuts per-call API cost sixfold, from \$2.20 to \$0.36 per thousand calls. The estimates multiply the observed input and output token counts by their respective published prices⁴.

Judge	Hint	Amt (%)	Dir (%)	Latency (s)
Flash	yes	76.0	90.2	1.9
Flash	no	78.8	93.1	1.9
Flash-Lite	yes	70.0	76.2	1.5
Flash-Lite	no	53.1	76.9	1.5

Table 6: Deployment frontier on the Mooré holdout ($n=160$). Flash is `gemini-3.5-flash`, Flash-Lite is `gemini-3.1-flash-lite`. Latency reflects the end-to-end median including the ASR front-end.

6 Conclusion

We presented GG-Abs, a generator-guided analysis-by-synthesis method for voice-based financial record-keeping in Mooré and French–Mooré market speech. Its central idea is to turn a local monetary convention into an executable recovery mechanism. Rather than relying on ASR coverage or free-form LLM interpretation, GG-Abs treats each spoken amount as a latent integer, reconstructs it through the *wakir* counting system, verifies it against noisy ASR outputs, and abstains when no auditable match exists.

On a sealed real-speech holdout of 312 recordings from seven speakers, the full pipeline reaches 76.0% amount accuracy on Mooré and 85.0% on code-switched speech, with direction accuracy above 90% in both regimes. These results show that inclusive voice finance requires more than multilingual recognition. It requires systems that encode local knowledge, protect balance-changing fields with deterministic checks, and make each deterministically recovered amount traceable. We therefore present GG-Abs not as a deployment-ready product but as an assistive, confirmation-gated step toward low-resource financial AI that is accessible, culturally grounded, auditable, and trustworthy in real use.

⁴<https://ai.google.dev/gemini-api/docs/pricing>, accessed 1 Jun 2026.

7 Limitations

Evaluation scope. We evaluate an assistive recovery tool, not a deployment-ready product. The sealed holdout contains 312 recordings from seven speakers in one language pair, Mooré and French. This supports the core hypothesis but leaves demographic, acoustic, and cross-lingual generalization untested, and the single-voice synthetic corpus may miss natural market noise and prosody. Larger field studies with more speakers and sessions remain future work.

Language and rule dependence. GG-Abs relies on a hand-built *wakir* grammar and frozen, linguist-validated ASR-drift rules. This supports auditability but limits portability, since a new language or monetary convention requires a new grammar, parser, and rule catalogue. Our claim is a generator-guided design principle, not a universal rule set. We also note that the drift rules change the final output on only 7 of 388 Mooré-side transcripts on the sealed holdout, so the deployed accuracy rests on re-synthesis scoring and skeleton matching rather than the rules, which serve auditability and validity.

Compute and latency. The two parallel ASR channels improve robustness but raise compute, and a production system would need to learn when fewer suffice. Low-latency use may require GPU access, compression, or an on-device recognizer. We have not yet evaluated deployment on low-end devices or poor connectivity.

Dependencies and open problems. Only the amount is recovered deterministically, while the pipeline still relies on the LLM judge for ledger structure, direction, and customer identity, so a cheaper judge loses direction accuracy (Table 6). Free-text name extraction is the primary bottleneck for strict full-ledger accuracy, and the direction verb lexicon is corpus-induced and awaits full linguistic validation. Routing audio to a commercial API on abstention also introduces privacy and cost concerns. Real-world deployment will therefore require stronger identity resolution, local inference, and explicit user confirmation before any ledger entry is committed.

8 Ethical considerations

Speech data was collected voluntarily in two stages, crowdsourced seeds for template design and record-

ings from native Mooré-speaking colleagues for evaluation. Participants gave informed consent, received no compensation, and may withdraw their data at any time. All transactional content is fictional. The only personal data collected is biometric voice, stored securely under Burkina Faso’s Loi n°010-2004 and GDPR principles, with participants specifically consenting to its collection, research use, and possible public release. To protect privacy, the corpus will be published under strict access control, withholding raw audio while sharing derived text and metadata more openly. Since the system’s audio-escalation path sends raw waveforms to a commercial API, any real-world deployment must disclose this transmission and obtain separate consent. We used large language models only to polish the writing of this paper, for grammar and phrasing. They were not used to generate research ideas, design or run experiments, or produce results. All scientific content and claims are the authors’ own.

Acknowledgments

This work is supported by the Luxembourg Ministry of Foreign and European Affairs through their Digital4Development (D4D) portfolio under the project LuxWAYs (Luxembourg/West-Africa Lab for Higher Education Capacity Building in Cyber-Security and Emerging Topics in ICT4Dev).

This research was also supported by [Google.org](https://www.google.org) through the provision of Google Cloud Platform credits used to conduct our experiments. We are grateful to the Centre d’Excellence Interdisciplinaire en Intelligence Artificielle (CITADEL) for facilitating data collection through its fellows. We also thank all individuals who contributed to the translation and recording tasks.

References

Tosin Adewumi, Mofetoluwa Adeyemi, Aremu Anuoluwapo, Bukola Peters, Happy Buzaaba, Oyerinde Samuel, Amina Mardiyah Rufai, Benjamin Ajibade, Tajudeen Gwadabe, Mory Moussou Koulibaly Traore, Tunde Oluwaseyi Ajayi, Shamsuddeen Muhammad, Ahmed Baruwa, Paul Owoicho, Tolulope Ogunremi, Phylis Ngigi, Orevaoghene Ahia, Ruqayya Nasir, Foteini Liwicki, and Marcus Liwicki. 2023. [Afriwoz: Corpus for exploiting cross-lingual transfer for dialogue generation in low-resource, african languages](#). In *2023 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

Maha Tufail Agro, Atharva Kulkarni, Karima Kadaoui,

- Zeerak Talat, and Hanan Aldarmaki. 2025. [Code-switching in end-to-end automatic speech recognition: A systematic literature review](#).
- Hussain Al Sharoufi and Waleed Al-Fadhli. 2025. [Bridging the gap: Pragmatic and cultural challenges in machine translation](#). *International Journal of Society, Culture & Language*, 13(2):29–43.
- Duygu Altinok. 2018. An ontology-based dialogue management system for banking and finance dialogue systems. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France. European Language Resources Association (ELRA).
- Evershed K Amuzu and John Victor Singler. 2014. Codeswitching in west africa.
- Séverin Baroudi, Yanis Labrak, Shashi Kumar, Joonas Kalda, Sergio Burdisso, Pawel Cyrta, Juan Ignacio Alvarez-Trejos, Petr Motlicek, Hervé Bredin, and Ricard Marxer. 2026. [Doctor or patient? synergizing diarization and asr for code-switched hinglish medical conditions extraction](#).
- Astik Biswas, Febe De Wet, Ewald Van der westhuizen, and Thomas Niesler. 2020. [Semi-supervised acoustic and language model training for English-isiZulu code-switched speech recognition](#). In *Proceedings of the 4th Workshop on Computational Approaches to Code Switching*, pages 52–56, Marseille, France. European Language Resources Association.
- Astik Biswas, Emre Yılmaz, Febe de Wet, Ewald van der Westhuizen, and Thomas Niesler. 2019. [Semi-Supervised Acoustic Model Training for Five-Lingual Code-Switched ASR](#). In *Interspeech 2019*, pages 3745–3749.
- Laurie Burchell, Alexandra Birch, Robert Thompson, and Kenneth Heafield. 2024. [Code-switched language identification is harder than you think](#). In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 646–658, St. Julian's, Malta. Association for Computational Linguistics.
- Amrita Chatterjee and Simontini Das. 2021. [Information communication technology diffusion and financial inclusion: an inter-state analysis for india](#). *Innovation and Development*, 11(1):1–23.
- Chen Chen, Yuchen Hu, Chao-Han Huck Yang, Sabato Marco Siniscalchi, Pin-Yu Chen, and Eng-Siong Chng. 2023. Hyporadise: An open baseline for generative speech recognition with large language models. *Advances in Neural Information Processing Systems*, 36:31665–31688.
- Victoria YH Chua, Hexin Liu, Leibny Paola Garcia Perera, Fei Ting Woon, Jinyi Wong, Xiangyu Zhang, Sanjeev Khudanpur, Andy WH Khong, Justin Dauwels, and Suzy J Styles. 2023. Merlion ccs challenge: A english-mandarin code-switching child-directed speech corpus for language identification and diarization. *arXiv preprint arXiv:2305.18881*.
- Marta R Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.
- G. Mohan Dhanush, J. Gopichand, and B. Balasaigayatri. 2025. [Language diarization model for bilingual code-switched speech analysis](#). In *2025 7th International Conference on Intelligent Sustainable Systems (ICISS)*, pages 656–661.
- Abdoulaye Diack, Perry Nelson, Kwaku Agbesi, Angela Nakalembe, MohamedElfatih MohamedKhair, Vusumuzi Dube, Tavonga Siyavora, Subhashini Venugopalan, Jason Hickey, Uche Okonkwo, Abhishek Bapna, Isaac Wiafe, Raynard Dodzi Helegah, Elikem Doe Atsakpo, Charles Nutrokor, Fiifi Baffoe Payin Winful, Kafui Kwashie Solaga, Jamal-Deen Abdulai, Akon Obu Ekpezu, Audace Niyonkuru, Samuel Rutunda, Boris Ishimwe, Michael Melese, Engineer Bainomugisha, Joyce Nakatumba-Nabende, Andrew Katumba, Claire Babirye, Jonathan Mukibi, Vincent Kimani, Samuel Kibacia, James Maina, Fridah Emmah, Ahmed Ibrahim Shekarau, Ibrahim Shehu Adamu, Yusuf Abdullahi, Howard Lakougna, Bob MacDonald, Hadar Shemtov, Aisha Walcott-Bryant, Moustapha Cisse, Avinatan Hassidim, Jeff Dean, and Yossi Matias. 2026. [Waxal: A large-scale multilingual african language speech corpus](#).
- Gil Keren, Artyom Kozhevnikov, Yen Meng, Christophe Ropers, Matthew Setzler, Skyler Wang, Ife Adebara, Michael Auli, Can Balioglu, Kevin Chan, et al. 2025. Omnilingual asr: Open-source multilingual speech recognition for 1600+ languages. *arXiv e-prints*, pages arXiv-2511.
- Franziska Koefer, Amber Bokkens, Massimo Preziuso, and Michel Ehrenhard. 2024. [Addressing financial and digital literacy challenges for inclusive finance: Insights from microfinance institutions and fintech organisations](#). *EIF Working Paper*, (2024/97).
- D Fortuné Kponou, Salima Mdhaffar, Fréjus AA Laleye, Eugène Cokou Ezin, and Yannick Estève. 2025. Ffstc 2: Extending the fongbe to french speech translation corpus. In *Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025)*, pages 145–152.
- Damien Kunjal. 2026. [From access to ability: The impact of financial inclusion on financial knowledge, capabilities, and literacy among gen z university students in south africa](#). *Journal of Risk and Financial Management*, 19(3).
- Indrani Medhi Thies. 2015. [User interface design for low-literate and novice users: Past, present and future](#). *Found. Trends Hum.-Comput. Interact.*, 8(1):1–72.

- Mumtaz Begum Mustafa, Mansoor Ali Yusoof, Hasan Kahtan Khalaf, Ahmad Abdel Rahman Mahmoud Abushariah, Miss Laiha Mat Kiah, Hua Nong Ting, and Saravanan Muthaiyah. 2022. [Code-switching in automatic speech recognition: The issues and future directions](#). *Applied Sciences*, 12(19).
- Carol Myers-Scotton. 1997. *Duelling languages: Grammatical structure in codeswitching*. Oxford University Press.
- Palanichamy Naveen and Pavel Trojovský. 2024. [Overview and challenges of machine translation for contextually appropriate translations](#). *iScience*, 27(10):110878.
- Tolúlopé Ògúnremí, Christopher D. Manning, and Dan Jurafsky. 2023. [Multilingual self-supervised speech representations improve the speech recognition of low-resource African languages with codeswitching](#). In *Proceedings of the 6th Workshop on Computational Approaches to Linguistic Code-Switching*, pages 83–88, Singapore. Association for Computational Linguistics.
- Fayçal S. A. Ouedraogo, Maimouna Ouattara, Rodrigue Kafando, Abdoul Kader Kabore, Aminata Sabane, and Tegawendé F. Bissyandé. 2026. [Contributing to speech-to-speech translation for African low-resource languages : Study of French-mooré pair](#). In *Proceedings of the Second Workshop on Language Models for Low-Resource Languages (LoResLM 2026)*, pages 623–629, Rabat, Morocco. Association for Computational Linguistics.
- Peterson K. Ozili. 2025. [Financial literacy theory of financial inclusion](#). *SSRN Electronic Journal*.
- Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoeng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, et al. 2024. [Scaling speech technology to 1,000+ languages](#). *Journal of Machine Learning Research*, 25(97):1–52.
- Melain Modeste Senou, Wautabouna Ouattara, and Denis Acclassato Houensou. 2019. [Financial inclusion dynamics in waemu: Was digital technology the missing piece?](#) *Cogent Economics & Finance*, 7(1):1665432.
- T. H. O. da Silva et al. 2024. [How do illiterate people interact with an intelligent voice assistant?](#) *International Journal of Human–Computer Interaction*. Voice assistants and access for low-literacy users (publisher page). Accessed 2026-02-24.
- Harish Thangaraj, Ananya Chenat, Jaskaran Singh Walia, and Vukosi Marivate. 2024. [Cross-lingual transfer of multilingual models on low resource african languages](#).
- Gokhan Tur and Renato De Mori. 2011. *Spoken language understanding: Systems for extracting semantic information from speech*. John Wiley & Sons.
- UNESCO. 2025. [Youth and adult literacy: International literacy day 2025. promoting literacy in the digital era. factsheet](#). UIS education database, September 2025.
- World Bank. 2023. [Literacy rate, adult total \(% of people ages 15 and above\) - burkina faso](#). <https://data.worldbank.org/indicator/SE.ADT.LITR.ZS?locations=BF>. Most recent value: 42% in 2023. Source: UNESCO Institute for Statistics. Accessed: 2026-06-15.
- Hao Yu, Jesujoba Alabi, Andiswa Bukula, Jian Yun Zhuang, En-Shiun Annie Lee, Tadesse Kebede Guge, Israel Abebe Azime, Happy Buzaaba, Blessing Kudzaishie Sibanda, Godson Koffi Kalipe, et al. 2025. [Injongo: A multicultural intent detection and slot-filling dataset for 16 african languages](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9429–9452.
- Emre Yılmaz, Astik Biswas, Ewald van der Westhuizen, Febe de Wet, and Thomas Niesler. 2018. [Building a Unified Code-Switching ASR System for South African Languages](#). In *Interspeech 2018*, pages 1923–1927.

A Judge Prompt and Output Schema

We reproduce the conservative-judge prompt and JSON schema from `gg_abs/judge.py` and `gg_abs/main.py`. The prompt is assembled by `_build_prompt` from a context block, the two raw ASR transcripts, a decision-dependent body, and the schema tail, and is identical across regimes. Every call uses `gemini-3.5-flash` at temperature 0.1 with JSON output.

A.1 Context block

The domain context carries the *wakir* numeral system, the log-odds-induced direction verbs, the phonetic-tolerance note, and the tense rule.

```
Merchant credit-ledger utterance in Moore /
Moore-French code-switch (Burkina Faso,
L2-French speakers). Customers may take
goods/cash on credit (debt INCREASES) or
pay back (debt DECREASES) ; the speaker may
also ask a balance question. A single
utterance may contain several transactions
(compound). All amounts are in CFA.
Entity = customer's name or social role,
in lowercase snake_case, or null when the
speaker does not mention one.
Moore numeral system (only when speaker uses
Moore numerals; not needed for French):
* Speakers count in WAKIR units;
  CFA amount = wakir x 5.
* Magnitudes: tus = 1000 wakir,
  kobs = 100 wakir, piiga = 10 wakir.
* Units 1-9: yembre, yi, ta, naase, nu,
  yoobe, yopoe, nil, wae.
* Compound multiplies magnitude by unit:
  kobs-naase = 100 x 4 = 400 wakir = 2000 CFA.
* Connector "la a": tus la a kobs-yi
  = 1000 + 200 wakir = 6000 CFA.
French numerals are read directly in CFA.
Moore direction verbs (induced from synth via
log-odds; validated direction-discriminative):
* debt DECREASES: yaoo/yaoo/yaoo (pay back),
  koo (customer gives -> repays), reegd.
* debt INCREASES: tara/reega/ket (took credit).
* query: wana ("how much").
ASR Phonetic Tolerance: treat close phonetic
matches of these verbs as valid.
TENSE matters: count debt_down only if payment
ALREADY happened; a promised future payment
stays debt_up only.
```

A.2 Evidence, hint body, and tiering

The two channels follow under a fixed header (W2V-mos, Omni-mos), one transcript per line, empty channels as (none). With a Mooré hint (DECISION_MOORE_HINT) the body adds GG-Abs: amount(s) = {final_amounts} CFA and the judge finalizes direction and entity, one entry per amount. On abstention (DECISION_NO_HINT) the judge self-extracts amounts and returns null if none, the arm that routes to audio. Each emitted hint carries a categorical tier from independent ASR support, not a tunable score: HIGH when at least two channels agree (support_count $\geq$ 2) and MED for a single channel. LOW is the abstention state (no amount, no hint, routes to audio) and

never appears in the tier-precision table. The amount-bearing holdout splits into 56 HIGH and 94 MED (Appendix C).

A.3 Schema tail and escalation

The prompt closes by demanding strict JSON.

```
Return STRICT JSON matching this compound
ledger schema :
{"currency": "XOF",
 "debt_up": [{"entity": <"snake_case"|null>,
 "amount": <int CFA|null>}, ...],
 "debt_down": [{"entity": <"snake_case"|null>,
 "amount": <int CFA|null>}, ...],
 "queries": [{"type": <"who_owes_me"
|"debt_of_entity"|"total_due_to_me"
|"remaining_total">,
 "entity": <"snake_case"|null>}, ...],
 "is_proposal_correct": <bool>}
```

Lists let compound utterances be expressed natively, and output is parsed fence-tolerantly (parse_ledger). Escalation fires exactly when the module abstains (hint.decision == "no_hint"): the same prompt is sent with the raw audio as a second inline_data part (audio/wav), additional evidence rather than a replacement. This produces the headline CS amount 85.0 ± 0.5 and CS direction 92.7.

B Governed Rule Catalogue

The deterministic path is governed regex rewrites with no learned parameters: CPU-only, reproducible from cached transcripts without GPU or API. Table 7 is the deployed Mooré catalogue with a one-rule ablation on the male synthetic development split (intents_mos_fr, pool_dev, $n=800$, baseline amount-exact 77.0%, gold misses 4, false hints 180). The ablation is a deterministic replay with no LLM call: we disable one regex at a time and recompute the hint. Net is helps-minus-hurts, and every kept rule has net ≥ 0 and leaves gold misses unchanged.

The most load-bearing rule is *kobs-nii* (H-3/3, net +32). Expert rules carry literature anchors in the governance file: *samd[e/a/ā]* stripping is expert-validated as non-amount-bearing, and the truncated magnitudes and the phonetic-skeleton fold follow descriptive Mooré sources.

B.1 Code-switch amount gap

On code-switching the module should abstain and route audio to the judge, so the 85.0 versus 90.1 gap (5.1 points, about six utterances on $n=118$) splits in two. In three cases the module fails to abstain and emits a spurious MED hint on a French amount (gold 1000, 2500, 20000 read as 5, 40, 5). Because audio is not attached to a row that already

Rule	Prov.	Fires	Exact	Net	Δ miss
<i>Expert (linguist / literature)</i>					
<i>samd[e/a/ā] strip</i>	EXP	361	78.7	0	0
<i>wakir+tus preserve</i>	EXP	7	71.4	+3	0
<i>wakir strip</i>	EXP	307	73.6	0	0
<i>Heuristic (cross-LLM clusters)</i>					
<i>kobs,nii→kobs-nii</i>	H-3/3	63	63.5	+32	0
<i>pis,nii→pis-nii</i>	H-3/3	1	100.0	+1	0
<i>nung/nuy/num→nu</i>	H-2/3	22	86.4	+2	0
<i>pis[X]→pis-X</i>	H-3/3	120	88.3	0	0
<i>piir(ā)→piig</i>	H-3/3	5	40.0	0	0
<i>tus piir→tus piig</i>	H-3/3	2	100.0	0	0
<i>yogbe→yoobe</i>	H-3/3	0	—	0	0
<i>yokoe→yopoe</i>	H-3/3	0	—	0	0
<i>tus 1d→teen</i>	H-3/3	0	—	0	0

Table 7: Deployed Mooré catalogue, one-rule dev ablation (pool_dev, $n=800$, baseline 77.0%). EXP=expert, H- $k/3$ =proposed by k of three LLM analyses. Fires=utterances matched; Exact=amount accuracy (%) on those rows; Net=helps minus hurts; Δ miss=change in gold misses if removed (all 0).

carries a hint, these lose the rescuing escalation and are all wrong. Attaching audio even when a low-confidence hint fires on a likely-French row would lift CS amount toward 87.5. The other three are genuine residual. The gap is a wrinkle in one field, and GG-Abs still wins CS direction (92.7 vs. 57.6).

C Threshold and Reliability Ablations

The module has two frozen knobs: TAU_FUZZY, the ungated per-token fuzzy threshold (default

0.80), and TAU_GATE, the looser threshold for compound repair behind a matched prefix (default 0.65). Both sweeps are deterministic replays over the $n=800$ development pool, with the sealed hold-out untouched.

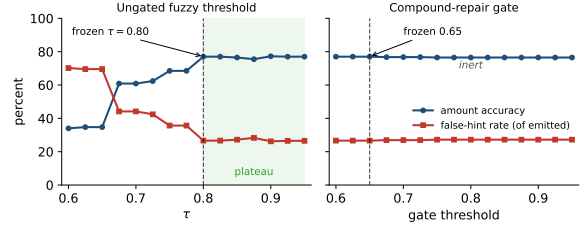


Figure 2: Threshold sweeps on the Mooré development pool ($n=800$, no LLM). Left: amount accuracy plateaus at and above $\tau=0.80$ and collapses below 0.65, where the false-hint rate overtakes accuracy. Right: the compound-repair gate is inert across its range. Both were fixed before the sealed test.

At the deployed $\tau=0.80$, accuracy is 77.0% with a 26.6% false-hint rate, and $\tau=0.85$ is similar (76.5%, 27.2%). Loosening to $\tau=0.65$ collapses the module to 34.8% accuracy and a 69.6% false-hint rate, since indiscriminate fuzzy matching maps noise onto magnitude markers. The frozen design makes magnitudes reachable only by exact token, exact skeleton, or an adjacency-gated rewrite. Sweeping TAU_GATE from 0.60 to 0.95 (with $\tau=0.80$) leaves accuracy in [76.5, 77.0]%, so precision comes from the prefix, not the gate value.