

Same SQL, Different Decision: Aggregate Risk Hides Cross-Lingual Instability in Text-to-SQL Verification

Avinash Goutham Aluguvelly
The University of Texas at Austin
algo@utexas.edu

Abstract

Text-to-SQL systems that use an LLM verifier typically calibrate its execute-or-defer threshold on English requests. We ask whether that threshold still makes the same decisions when the user writes in another language. Our evaluation is paired: the candidate SQL, the schema, the database, and the execution-derived correctness label are held fixed, and only the question changes, across seven languages and two open-weight verifiers. The verifiers rank candidates almost as well in every language, yet 14% to 27% of identical SQL receives a different execute-or-defer decision. The usual aggregates hide this. Selective risk stays near its English value because promotions and demotions offset, and coverage flags only the languages where the shift is one-sided. Underneath, both harms occur: correct queries are needlessly deferred, and incorrect queries that English withheld are executed; under Vietnamese, Llama executes 16.9% of all incorrect candidates, 22.9% of those it had deferred in English. Score-level corrections that equalize coverage barely change which queries run. Translating the request into English before verification helps more, lowering the flip rate and unsafe promotions on both verifiers, though it does not restore the English policy. Audits of multilingual verifiers should therefore report decision transitions, not only aggregate coverage and risk.

1 Introduction

A deployed text-to-SQL system cannot execute every query its generator produces. A query can execute successfully and still return wrong numbers. Practical systems interpose a **verifier** that scores whether a candidate query answers the user’s question, plus a threshold deciding whether the query executes automatically or is escalated to a human. That threshold is an operating point, and every operating point is calibrated on some distribution. We

study policies whose execution threshold is calibrated on English requests.

Zhou et al. (2026) establish that multilingual evaluators assign language-dependent absolute scores to semantically parallel content even when pairwise accuracy remains high, producing acceptance-rate disparities under global thresholds. We investigate a complementary downstream question: what does such score drift do to a risk-calibrated execution policy? Acceptance-rate disparity alone does not tell an operator whether a transported threshold makes the system unsafe, uselessly conservative, or both.

Text-to-SQL makes this answerable: correctness is derived by executing the query and comparing result sets, and the verifier’s score gates a database action.

Contributions.

1. **An execution-grounded counterfactual benchmark.** We hold the SQL, database, schema, and execution-derived label fixed across paired multilingual question renderings.
2. **A decision-level deployment analysis** showing that discrimination is largely stable while 14% to 27% of individual decisions change, with a transition decomposition explaining why selective risk (the error rate among executed queries) does not reveal it and why coverage (the fraction of candidates executed) reveals it only sometimes.
3. **A systematic evaluation of threshold transport.** Coverage parity is cheap; decision agreement is not restored by any score-level correction we tested.
4. **A practical mitigation candidate.** Canonicalizing the request into English improves decision agreement substantially more than any score-level correction, and the gain survives a

benchmark-defect control.

2 Related work

Selective prediction. Abstention under a confidence threshold is classical selective classification (El-Yaniv and Wiener, 2010; Geifman and El-Yaniv, 2017), evaluated by risk-coverage curves, which plot the error rate among retained predictions against the fraction retained. We adopt that framing but apply it to an executable action rather than a label, and we ask what happens when the operating point is transported across a distribution shift the calibration set never contained.

Calibration. Mapping scores to probabilities via Platt scaling (Platt, 1999) or isotonic regression (Zadrozny and Elkan, 2002) is standard, and modern networks are known to be miscalibrated (Guo et al., 2017). One property matters here: any strictly increasing recalibration applied identically to every language, with the threshold mapped through it or refit at the same risk target, leaves the induced decisions unchanged, so language-blind calibration cannot address the failure we document, and a non-strict map that merges distinct scores can change decisions only by discarding information. Distribution-free alternatives such as conformal prediction (Vovk et al., 2005) provide finite-sample risk guarantees but still require labeled calibration data in the target condition.

Text-to-SQL verification. Recent work finds that logic-aware LLM verifiers predict execution correctness better than self-consistency or model log-probability, while requiring calibration and transferring poorly to unseen schemas (Richardson, 2026). We take that verifier-gated setting as given and ask whether its operating point survives a change in the request language.

Multilingual text-to-SQL. MultiSpider (Dou et al., 2023) supplies semantically parallel questions across seven languages over executable Spider databases (Yu et al., 2018), which is what makes a paired counterfactual possible. Section 3.5 documents where its value preservation is incomplete. Spider is also likely contaminated for large models: Ranaldi et al. (2024) find GPT-3.5 drops sharply on the unseen Termite benchmark and under altered schemas. The paired design controls candidate variation, but contamination was not measured and could still affect cross-language verifier comparisons if a verifier has language-specific

familiarity with benchmark questions.

Multilingual evaluator bias. Closest to our work, Zhou et al. (2026) show multilingual reward models and LLM judges assign language-dependent pointwise scores even at >90% pairwise accuracy, with acceptance-rate differences up to 43 points under a global threshold across 23 languages. Their framing emphasizes unsafe acceptance: lower-resource languages receive higher scores, so harmful content passes filters more readily.

We differ in the evaluated artifact, the label source, and the calibration objective. We hold a formal SQL program literally identical across languages, derive labels by execution rather than preference, and calibrate to a target *error rate among executed queries* rather than a target acceptance fraction. That lets us report which failure direction a SQL agent actually suffers, separate language shift from database shift, and measure churn in individual decisions that aggregate rates cannot reveal. Section 5 measures the last of those. Their language-specific percentile thresholds are closely related to the quantile transport we evaluate in Section 7.1; we treat it as a concurrent baseline.

Translate-test. Translating input into a high-resource pivot language before applying an English-trained model is a long-standing cross-lingual baseline (Conneau et al., 2018). Section 7.3 evaluates pivot translation as a mitigation and finds that it improves decision agreement more than any score-level correction on both verifiers, while also lowering unsafe promotions, rather than merely aligning aggregate coverage.

3 Experimental design

3.1 Data and the variant that matters

We use **MultiSpider** (Dou et al., 2023): parallel questions in seven languages (en, de, es, fr, ja, vi, zh) over executable Spider databases.

MultiSpider ships two variants. The default, `with_original_value`, localizes literal values inside questions (a question about “JetBlue Airways” becomes a question about a different airline in Chinese) while leaving the gold SQL in English. The translated question then no longer asks what the gold query answers, and a verifier’s correct low confidence reads as dramatic cross-lingual failure. All results use `with_english_value`.

Section 3.5 audits how completely that variant actually preserves literals; it does not do so everywhere.

We sample 1,200 questions from train and dev, capped at 10 per database, keeping only items whose gold SQL executes and returns a non-empty result set: 1,200 questions over 163 databases (1,053 train / 147 dev). Sampling dev alone would bound the study at approximately 200 questions, since Spider’s dev split has only 20 databases.

3.2 Candidates and labels

Qwen2.5-Coder-7B-Instruct (Hui et al., 2024) generates five candidates ($K = 5$) per question from the English question only (temperature 0.8). Each is executed and labeled by result-set match against the gold query, with row order enforced only when the gold query has `ORDER BY`: 6,000 candidates, 5,053 correct / 947 incorrect (84.2%).

Generating from English only is what makes the design counterfactual. Every language is later shown the *same* SQL with the *same* label.

We call these labels **execution-derived correctness**, or equivalence on the benchmark database instance. Two distinct SQL programs can coincidentally agree on one instance; this is the standard Spider convention rather than unrestricted semantic equivalence.

Execution match follows Spider’s comparison convention. Row order is enforced only when the gold query has `ORDER BY`, and column order is never enforced, so a candidate that returns the requested columns in a different order from the gold query counts as correct. A positional comparison would mislabel such column-reordered `SELECT` lists as incorrect; in this corpus that affects 159 candidates across 45 questions, all labeled correct here. Appendix G traces one candidate through the pipeline.

3.3 Verification

We use two verifiers from different families: **Llama-3.1-8B-Instruct** (Dubey et al., 2024) and **Qwen2.5-7B-Instruct** (Yang et al., 2024). Each sees the schema, one language’s question, and the candidate SQL, and answers `Yes` or `No`. The score is read from the next-token distribution,

$$s_{i,\ell} = \frac{P(\text{Yes})}{P(\text{Yes}) + P(\text{No})}, \quad (1)$$

summed over casing and whitespace variants. This is a forced-choice logit read rather than a generated numeric confidence, so nothing below is at-

tributable to a model’s ability to verbalize a probability. Every candidate is scored under all seven languages by both verifiers: 42,000 scorings per verifier, 84,000 total, no parse failures.

Scoring runs at batch size 1. Batched `bfloat16` forward passes are not numerically identical to unbatched ones. Here batching perturbed confidences by up to 6×10^{-2} , the magnitude of the effects under study (Appendix H).

3.4 Splits, threshold, and decision metrics

One canonical database-disjoint split is used for all primary threshold, decision, transport, and pivot analyses. Corpus-level audits (Section 3.5), repeated-split results (Appendix F), and the screening subset (Section 8) are labeled where they appear. The 163 databases are partitioned 50/50 by database (seed 0): 81 calibration databases (599 questions, 2,995 candidates) and 82 test databases (601 questions, 3,005 candidates: 2,519 correct / 486 incorrect). Every threshold and every fitted correction is estimated on calibration databases; every primary result is measured on test databases.

The English threshold τ_{en} is the most permissive value whose selective risk on calibration-database English scores stays within a 10% target. Feasibility is evaluated only at *distinct* score values, since candidates sharing a score cannot be separated by any threshold. That matters for Qwen, whose $\tau_{\text{en}} = 0.9999971$ sits at the score ceiling because 82% of its English scores exceed 0.99.

Write $d_{i,\ell} = \mathbf{1}[s_{i,\ell} \geq \tau_{\text{en}}]$ for the decision on candidate i under language ℓ , and y_i for its execution-derived label. Then

$$\text{Flip}_\ell = P(d_{i,\ell} \neq d_{i,\text{en}}) \quad (2)$$

$$\text{UnsafePromotion}_\ell = P(d_{i,\text{en}}=0, d_{i,\ell}=1 \mid y_i=0) \quad (3)$$

$$\text{LostAutomation}_\ell = P(d_{i,\text{en}}=1, d_{i,\ell}=0 \mid y_i=1) \quad (4)$$

These are joint rates over all incorrect (resp. correct) candidates, which keeps denominators fixed across languages, so a language cannot look safer merely by withholding more. We also report the conditional form: among incorrect queries that English withheld, the share the target language executes. It is the sharper safety statement and is necessarily larger.

English is the calibrated reference policy, not ground truth. Whether a change is harmful is decided by y_i , not by English agreeing.

Coverage spread is $\max_\ell \text{coverage}_\ell - \min_\ell \text{coverage}_\ell$ over all seven languages including English, on test databases, unless a table states

otherwise. Where a result is averaged over repeated database splits, we say so.

3.5 Literal and numeric fidelity audit

We audit MultiSpider’s renderings against the English source; the full table is in Appendix E. Numeric values survive well (95.8% to 98.6% across languages). Quoted literals do not: preservation ranges from 87.8% (fr) down to 47.3% (vi), where "Pass" becomes 'Đạt' and 'Spices' becomes 'Gia vị'. The `with_english_value` variant therefore does not fully deliver on its name (16 to 69 items per language are affected), and a replication should know this.

Sensitivity analysis. Excluding items with detected literal mismatches changes unsafe-promotion rates by at most 0.6 percentage points. Vietnamese has the worst fidelity and the highest unsafe promotion, and its unsafe-promotion rate moves from 16.9% to 17.0%. The defect is real but does not drive our results. Our paired interpretation assumes semantic equivalence. Full compositional equivalence (preservation of negation, aggregation, ordering, comparison direction, and requested columns) is not manually verified, so our conclusions remain conditional on that assumption.

3.6 Statistics

All intervals are percentile bootstrap. The resampling unit is the question within a language and the database for comparisons spanning database populations. Nonzero decision-rate intervals use a question-clustered percentile bootstrap. For zero-event cells, where that bootstrap collapses to $[0, 0]$, we report an unclustered Clopper-Pearson upper bound as a descriptive sensitivity bound. At this sample scale, negligible differences reach significance, so we report effect sizes throughout.

4 Similar ranking performance, different execution decisions

Per-language AUROC on test databases ranges 0.687 to 0.735 (Llama) and 0.701 to 0.754 (Qwen); the largest gap against English is 0.048 (Llama, vi) and 0.054 (Qwen, ja). The full table is in Appendix B. Cross-lingual changes in discrimination are small relative to the operating-point effects that follow, though not zero; Section 6.1 returns to ja and vi.

Table 1 applies τ_{en} unchanged on test databases, the protocol of every later table; the full-data version is in Appendix B. Llama executes 64.6% of

	en	de	es	fr	ja	vi	zh	spread
Llama	0.544	0.579	0.646	0.423	0.566	0.595	0.565	0.223
Qwen	0.610	0.499	0.537	0.529	0.554	0.462	0.564	0.148

Table 1: Coverage at the English threshold τ_{en} , applied unchanged to every language, on test databases. Llama $\tau = 0.755$; Qwen $\tau = 1.000$. Bold marks each verifier’s minimum.

English-conditioned SQL candidates automatically under Spanish questions and only 42.3% of the identical candidates under French questions; Qwen executes 61.0% under English and 46.2% under Vietnamese.

Figure 1 shows why this is not a ranking failure: the curves largely overlap while the English threshold marker lands at very different coverages.

One might expect the transported threshold to become *unsafe*. On the test split no language on either verifier breaches the 10% target significantly, although point estimates exceed it for Llama under Vietnamese (0.107, CI $[0.081, 0.135]$) and German (0.102, CI $[0.075, 0.130]$), and Llama under French sits entirely below it (0.068, CI $[0.043, 0.095]$). Failure to detect a breach is not evidence of target compliance. (Full corpus, Vietnamese on Llama: 0.118, CI $[0.099, 0.138]$.) Through aggregate risk alone the system appears primarily conservative; Section 5 shows that unsafe promotions nevertheless occur.

5 Decision-level inconsistency and harm

All results in this section are on test databases (3,005 candidates).

5.1 14% to 27% of identical SQL receives a different decision

A flip is any change of decision, and not every flip is harmful; the harmful share is measured by unsafe promotion and lost automation.

Table 2 gives the Llama breakdown (flips 13.9% to 19.5%). Qwen’s table is in Appendix C; its flips run 17.4% to 27.2% and it predominantly over-defers (lost automation 12.5% to 22.1%, unsafe promotion $\leq 10.5\%$). The headline is operating-point dependent: at a 5% target flips fall to 9.2% to 13.8% (Llama) and 11.0% to 14.7% (Qwen), but Qwen then executes only 16.7% of English queries, and its 10%-target threshold sits near the score ceiling (Appendix F).

Which harm dominates is verifier-dependent. French on Llama is almost entirely over-

Same SQL, same selection quality — very different automation rates

Gray curves: each language's achievable risk-coverage trade-off (they overlap). Markers: where one English-calibrated threshold lands.

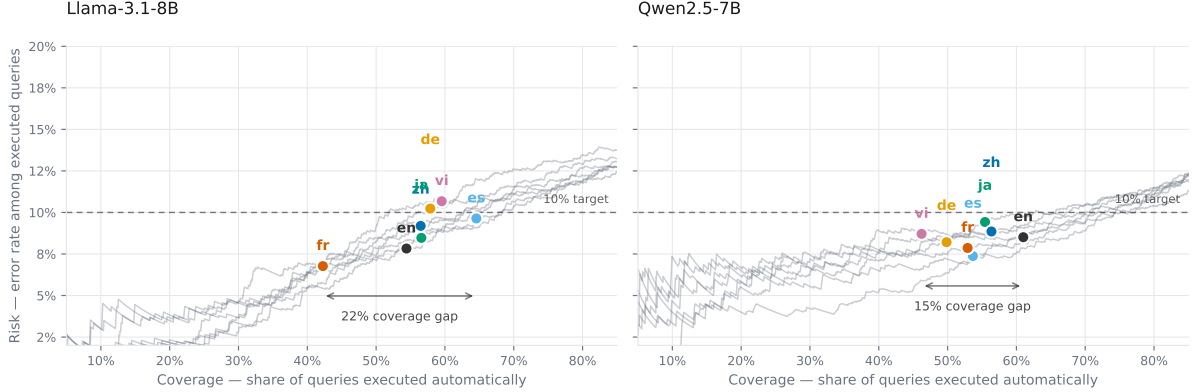


Figure 1: The curves largely overlap, indicating broadly similar selection quality. The markers show where the single English-calibrated threshold actually lands: very different coverages, spreads of 22.3 and 14.8 points, at similar risk. Test databases, one threshold per verifier fitted on English calibration databases at a 10% risk target.

lang	flip	lost	unsafe promotion		95% CI
			all	withheld	
de	13.9%	5.8%	12.8%	17.3%	[7.9%, 17.9%]
es	14.9%	2.6%	13.2%	17.9%	[8.4%, 18.3%]
fr	14.6%	14.0%	1.6%	2.2%	[0.2%, 4.3%]
ja	19.5%	9.0%	10.1%	13.7%	[6.1%, 14.3%]
vi	17.3%	6.5%	16.9%	22.9%	[11.3%, 23.1%]
zh	15.9%	7.1%	11.7%	15.9%	[7.4%, 16.9%]

Table 2: Llama-3.1-8B. *lost* is lost automation. *unsafe* (*all*) divides by all 486 incorrect candidates; *unsafe* (*withheld*) by the 358 that English withheld. Intervals are question-clustered.

conservative (14.0% of correct queries deferred against 1.6% of incorrect promoted, 8 of 486); Vietnamese on the same verifier moves 16.9% of all incorrect candidates from defer to execute, 22.9% of those English withheld. Qwen over-defers in every language.

5.2 Why aggregate risk stays stable

Lost automation and unsafe promotion have different denominators, so quoting them side by side does not establish cancellation. The transition counts do (Table 3).

For Spanish, 447 decisions change. Among incorrect queries, 64 become executed and 5 become deferred, giving $\Delta_{err} = 59$. Among correct queries, 312 become executed and 66 become deferred, giving $\Delta_{cor} = 246$. The accepted set therefore grows by 305 queries, while selective risk rises from 7.8% to 9.6%.

Aggregate selective risk is a ratio, so its change

lang	correct		incorrect		Δ_{err}	Δ_{cor}	risk	
	$e \rightarrow d$	$d \rightarrow e$	$e \rightarrow d$	$d \rightarrow e$			en	ℓ
de	145	199	12	62	+50	+54	0.078	0.102
es	66	312	5	64	+59	+246	0.078	0.096
fr	352	29	50	8	-42	-323	0.078	0.068
ja	227	276	33	49	+16	+49	0.078	0.085

Table 3: Transition decomposition, Llama. Counts of test candidates by true label and by English decision $\rightarrow$ target-language decision, with e = execute and d = defer. Δ_{err} and Δ_{cor} are the net changes in accepted incorrect and accepted correct queries, respectively; the change in the selective-risk denominator is $\Delta_{err} + \Delta_{cor}$. Complete counts for both verifiers in Appendix C.

is far smaller than the churn in its numerator and denominator; it cannot distinguish a stable policy from a heavily reshuffled one. *Aggregate selective risk does not establish counterfactual consistency and does not rule out unsafe promotions.*

Coverage, the other monitored aggregate, is only partly informative. Under Qwen every language lowers coverage (61.0% in English, 46.2% to 56.4% elsewhere), which a coverage monitor would flag. Under Llama, German, Japanese, and Chinese stay within 3.5 points of English's 54.4% while 13.9% to 19.5% of their decisions flip. Risk monitoring misses every case and coverage monitoring only the one-sided shifts; neither identifies which queries changed.

6 What the shift is, and what it is not

6.1 An approximately class-independent offset, with exceptions

Write $\kappa = \delta^{\text{correct}} - \delta^{\text{incorrect}}$ for the difference between a language’s mean score shift on correct and on incorrect candidates. A small κ is consistent with a scale shift that leaves discrimination unchanged; equal mean shifts do not by themselves constrain the score distributions, so discrimination claims rest on the measured AUROC (Table 9). Databases are resampled; the full table is in Appendix D.

For Qwen, κ ’s interval covers zero in all six languages. For Llama, it covers zero in three. Japanese (-0.039 , CI $[-0.057, -0.021]$) and Vietnamese (-0.039 , CI $[-0.061, -0.017]$) shift incorrect-query scores more favorably relative to correct-query scores, and Spanish shows a smaller shift in the same direction (-0.018 , CI $[-0.033, -0.003]$). That is a relative class-dependent shift; these are also Llama’s two lowest test AUROCs (Table 9).

Japanese and Vietnamese are also the two languages with the clearest Llama AUROC degradation on test databases, and Vietnamese has the highest unsafe promotion. The pattern is consistent with resource-related effects reported by Zhou et al. (2026), but our design cannot test that explanation: we did not measure or manipulate resource availability, and per-model pretraining exposure is unknown.

We describe the effect as an approximately additive, class-independent offset *for these verifiers on this task*, an empirical description rather than a causal explanation.

6.2 No evidence that perplexity or database composition explains the shift

Perplexity of each question under the verifier itself correlates with the confidence shift at $r = +0.043$ across 4,500 candidate-language pairs, essentially zero and with the sign opposite to the hypothesis. French and Spanish have indistinguishable mean log-perplexity (3.568, 3.583) yet shifts of -0.091 and $+0.025$ in opposite directions (Appendix E).

Schema shift alone moves coverage by -0.035 , CI $[-0.093, +0.026]$: a small and imprecise effect. The largest language-shift point estimates persist on held-out databases and grow for French on Llama ($-0.122 \rightarrow -0.157$). The database-level intervals are wide, because the resampling unit becomes the database (roughly 80 clusters against

method	Llama	Qwen	fitted on
Raw, A, B	0.223	0.148	none; en only; pooled
C	0.040	0.083	599 q. per language
D	0.040	0.089	599 q. per language*

Table 4: Coverage spread after calibration. A is fitted on English only, B on pooled multilingual labels, C is a separate calibrator per language, D shares one slope and fits a per-language intercept. Every calibrator is fitted on all 599 calibration-split questions (2,995 candidates) per language; *D’s shared slope is additionally fitted on all seven languages pooled. Raw, A and B produce identical decisions. Full four-way table in Appendix B.

roughly 600 questions), which roughly doubles them. So this analysis neither attributes the effect to database composition nor establishes precise cross-schema generalization (Appendix E).

7 Restoring the operating point

Each method is asked two things: whether it restores comparable coverage and whether it restores the English decisions. The answers differ.

7.1 Coverage is easy to restore

We fit four calibrators on calibration databases. Table 4 gives coverage spread over seven languages, test databases, single split.

Calibrators A and B produce decisions identical to raw despite differing data requirements. A needs English labels; B needs labeled data in every language; neither changes a single decision.

A and B cannot work, by arithmetic rather than experiment. Both apply one strictly increasing function to every language, so a threshold refitted afterwards at the same risk target selects the identical set as the raw threshold; their coverage matches raw to three decimals. Calibrating the verifier is therefore not an answer. The correction must be language-aware.

For Llama, a single per-language intercept (D, 0.040) matches full per-language calibration (C, 0.040), consistent with Section 6.1. For Qwen both help less: C reduces the spread only from 0.148 to 0.083, and D to 0.089, roughly halving the spread rather than closing it. Qwen is the saturated verifier, and fitting a logistic map on the logits of near-ceiling scores is ill-conditioned. Language-aware calibration therefore closes most of the gap only when the verifier’s score distribution is well spread.

The coverage spreads above use the full cali-

bration budget. A separate learning-curve check measures expected calibration error (ECE) rather than coverage: with 25 and 50 labeled questions, D’s ECE on Llama is 0.0620 and 0.0588 for German (0.0219 with all 599) and 0.0563 and 0.0511 for French (0.0188 with all 599), lower than C’s at the same budgets but well above the full-data value. This matters because per-language labels are the scarce resource in deployment: D is the cheaper calibrator to fit at small budgets, but the remaining gap to the full-data value is material. D’s shared slope is still fitted on pooled data from all languages, so the per-language count is not its total data requirement. We did not measure coverage recovery at those budgets.

Separately, quantile transport substantially reduces mean coverage spread, to 4.9 and 4.7 percentage points (0.049 ± 0.011 and 0.047 ± 0.023 ; mean over ten database splits, unlike the single-split Table 4). Section 7.2 reports the same transport on the canonical split. But it equalizes coverage without enforcing the target selective risk: the worst language/split combination reaches 13.6% (Llama) and 14.3% (Qwen) against a 10% target, an overshoot that also appears in English on held-out databases and is a separate risk-calibration failure under database shift.

7.2 Decisions are not

We evaluate four transports plus the raw baseline, all fitted on calibration databases without target-language labels. **Offset** shifts target-language logits by the difference between the English and target-language mean logits on calibration data. **Affine** additionally rescales so that the target-language logit mean and standard deviation match English’s, without using the paired scores. **Quantile** aligns the target-language empirical score distribution to English’s. Only **isotonic** uses the paired structure: it fits a monotone map from each candidate’s target-language score to its own English score. Accepted-set overlap is the Jaccard similarity

$$\text{Overlap}_\ell = \frac{|A_\ell \cap A_{\text{en}}|}{|A_\ell \cup A_{\text{en}}|}, \quad (5)$$

where A_ℓ is the set of candidates executed under language ℓ . Table 5 averages over the six non-English languages on test databases.

Coverage spread collapses readily, but decision flips barely move: 16.0% $\rightarrow$ 15.0% at best on Llama and 21.6% $\rightarrow$ 20.5% on Qwen, and on Qwen the paired isotonic transport is worse than

verifier	method	flip	unsafe	lost	overlap	risk
Llama	raw	16.0%	11.0%	7.5%	0.747	0.092
Llama	offset	15.7%	11.7%	6.6%	0.756	0.095
Llama	affine	15.5%	10.9%	7.1%	0.755	0.093
Llama	quantile	15.2%	9.3%	8.1%	0.755	0.089
Llama	isotonic	15.0%	7.5%	9.4%	0.752	0.084
Qwen	raw	21.6%	7.0%	15.7%	0.680	0.084
Qwen	offset	20.7%	10.8%	10.0%	0.712	0.089
Qwen	affine	20.7%	10.7%	10.2%	0.712	0.088
Qwen	quantile	20.5%	10.7%	10.0%	0.714	0.089
Qwen	isotonic	50.2%	0.4%	54.0%	0.180	0.048

Table 5: Score transport, mean over the six non-English languages, test databases. Isotonic is the paired monotone map. Bold marks each verifier’s lowest flip rate. Per-language tables in Appendix B.

verifier	condition	flip	unsafe	lost	overlap	cov.	risk
Llama	native	16.0%	11.0%	7.5%	0.747	0.562	0.092
Llama	bilingual	14.5%	6.6%	9.4%	0.755	0.499	0.073
Llama	pivot	12.2%	2.7%	9.0%	0.791	0.495	0.075
Qwen	native	21.6%	7.0%	15.7%	0.680	0.524	0.084
Qwen	bilingual	16.9%	10.5%	5.9%	0.765	0.657	0.087
Qwen	pivot	17.6%	3.1%	13.8%	0.729	0.513	0.071

Table 6: Native, bilingual, and pivot conditions, mean over the six non-English languages, test databases. *cov.* is coverage. Bold marks each verifier’s best flip, unsafe-promotion, and overlap values.

doing nothing. Qwen’s isotonic failure has a precise cause: its threshold sits at the score ceiling, and isotonic regression averages over ties, so little survives. The tested score transports become ill-conditioned near the ceiling.

Parity in how many queries run does not imply agreement on which ones.

7.3 Input canonicalization improves decision agreement

If score-level correction cannot recover the policy, the input can be canonicalized. We test two additional conditions on the same fixed SQL. The **bilingual** condition pairs the native question with the benchmark’s gold English question, a diagnostic condition since that reference is unavailable in deployment. The **pivot** condition (input canonicalization by machine translation) translates the native question into English using Qwen2.5-7B-Instruct. Both use the same decision-level metrics. Table 6 reports means over the six non-English languages on the canonical test split; English coverage is 0.544 (Llama) and 0.610 (Qwen).

Rising overlap distinguishes recovery from mere conservatism: the accepted set moves *closer to* the English one rather than shrinking. Bilingual lowers

flips on both verifiers, and on Qwen it reaches the lowest flip rate and highest overlap of any condition, but there it raises unsafe promotions (7.0% $\rightarrow$ 10.5%) where pivot lowers them (3.1%). On Llama bilingual reduces unsafe promotions (11.0% $\rightarrow$ 6.6%) but worsens lost automation.

Per-language results. Six-language means can hide language-specific harm, so Appendix B reports pivot per language with paired database-bootstrap intervals. All six languages have lower flip-rate point estimates on both verifiers; the reduction is significant in five of six on each verifier. The worst per-language unsafe-promotion rate falls from 16.9% to 3.7% (Llama) and 10.5% to 4.1% (Qwen). Three caveats remain: Vietnamese on Llama and Japanese on Qwen show no significant flip reduction, although their unsafe promotion falls to 3.7% and to 2.1% (from 8.8%), and French on Llama rises from 1.6% to 2.3% unsafe promotion.

Scope limits. The translator is Qwen2.5-7B-Instruct, also one of the verifiers; the Llama improvement shows the effect is not purely self-model alignment, but Qwen’s gain may partly reflect it. And schemas, identifiers, and gold SQL remain English, so pivot may work partly by making a mixed-language prompt monolingual. Fully localized systems are therefore out of scope. A deployable bilingual condition would pair the native question with a machine-translated English rendering.

Pivot does not recover the English policy: 12.2% and 17.6% of decisions still differ, and lost automation on Llama rises slightly (7.5% $\rightarrow$ 9.0%).

Against quantile, the best score transport on Qwen and within 0.2 points of the best on Llama, pivot’s advantage is significant under a paired database bootstrap: -3.0 points $[-4.3, -1.6]$ on Llama and -2.9 points $[-4.7, -1.1]$ on Qwen. Across ten database splits it holds in all ten on Llama but only seven on Qwen (Appendix F).

Benchmark-repair control. Machine translation restores 37.7% of the Vietnamese literals that MultiSpider’s own translation had altered (Section 3.5), so pivot could be fixing a dataset defect. Restricting to the literal-clean subset (items needing no repair) leaves the improvement essentially unchanged (literal-clean column of Table 13; Qwen/de moves from -6.4% to -6.5%). Pivot is not aliased onto English: only 6.9% of 7,200 translations match the original English up to case and whitespace, all short questions where translation legitimately converges.

8 Robustness to verifier and deployment configuration

On a prespecified database-disjoint screening subset (300 questions, 39 databases, 1,500 candidates), matched-coverage comparisons found no measurable cross-lingual sensitivity difference from three configuration changes: verbalizer Yes/No to CORRECT/INCORRECT, $\Delta +1.5\%$ $[-3.1, +5.6]$; dense to sparse-MoE, $\Delta +0.2\%$ $[-4.1, +4.0]$; bf16 to 4-bit precision, $\Delta -2.4\%$ $[-6.1, +0.8]$. Comparisons are coverage-matched because a verifier that accepts more has more chances to flip (Appendix H).

Independently of language, quantizing the same checkpoint changes 8.4% of English decisions and swapping dense for MoE changes 20.6% under a shared threshold. Most additional verifiers failed the prespecified saturation gate (their scores clustered too heavily at 0 or 1 to admit a usable threshold), including Aya Expanse 8B (Dang et al., 2024) (72.5% of English scores within 10^{-6} of 0 or 1, 19.1% English coverage); we draw no conclusion from them. The gate applied only to additional verifiers; Qwen2.5-7B itself, with 60.8% of English scores within 10^{-6} of an endpoint, would fail it.

9 Conclusion

An English-calibrated SQL verifier can leave aggregate selective risk near its English value while changing 14% to 27% of execution decisions on identical SQL. Promotions and demotions offset, so selective risk does not reveal the change; coverage reveals it only when the shift is one-sided (every language on Qwen, French on Llama) and misses one decision in six flipping at near-English coverage.

Score-level correction reduces coverage disparity without target-language labels but does not recover the English policy. Canonicalizing the request into English improves decision agreement more, significantly in five of six languages on each verifier. Whether canonicalization holds up with a translator independent of the verifier, and over localized schemas, is the natural next experiment.

The inconsistency persists across two verifier families and a prespecified screening of verbalizer, architecture, and precision, though saturation limits those comparisons. Its magnitude and dominant harm direction depend on the verifier and configuration, so it must be measured per deployment at the decision level.

Acknowledgments

We thank the ORACLE 2026 reviewers for comments that led to the threshold-sensitivity and repeated-split analyses, the worked example, and the monitoring discussion in the final version.

Limitations

- **Two verifier families at full scale**, both 7B to 8B open-weights. The *direction* of failure is verifier-dependent; what replicates is that a fixed English operating point fails, not how.
- **Translation fidelity is imperfect** (Section 3.5): quoted literals survive in 47% to 88% of literal-bearing items by language, worst in Vietnamese. Sensitivity analysis shows a ≤ 0.6 -point effect on our conclusions, but the scope claim is conditional on semantic equivalence.
- **Labels are execution-derived** on one database instance, not unrestricted semantic equivalence. They are execution matches against the Spider gold queries under Spider’s comparison convention and were not manually audited for semantic equivalence beyond that.
- **163 databases** is MultiSpider’s ceiling, which limits the schema-shift analysis.
- **Score saturation is widespread** (four of six additional verifiers), limiting Section 8 to screening and making the tested score transports ill-conditioned at ceiling operating points.
- **The language set does not provide a controlled resource-level gradient**, which is why Section 6.1 declines a causal claim.
- **Generation is English-only by design**. The end-to-end setting, where language also affects which SQL is produced, is not evaluated in this controlled study.

Availability. Code, the 6,000 sampled candidates with execution labels, all 84,000 verifier scores plus the mitigation and screening scores, every analysis script, and the generated result tables are released at <https://github.com/tokenledger/same-sql-different-decision>, tagged at the camera-ready version. Every number in this paper is produced by a script in that repository from the released scores; no model inference is needed to reproduce the tables. Model weights are not redistributed; the verifiers and generator are public

checkpoints named in Section 3.3. MultiSpider and Spider are used under their own licenses and are not redistributed.

References

- Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.
- John Dang et al. 2024. Aya expanse: Combining research breakthroughs for a new multilingual frontier. *arXiv preprint arXiv:2412.04261*.
- Longxu Dou et al. 2023. MultiSpider: Towards benchmarking multilingual text-to-SQL semantic parsing. In *Proceedings of the AAAI Conference on Artificial Intelligence*. ArXiv:2212.13492.
- Abhimanyu Dubey et al. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Ran El-Yaniv and Yair Wiener. 2010. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11.
- Yonatan Geifman and Ran El-Yaniv. 2017. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*.
- Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*.
- Binyuan Hui et al. 2024. Qwen2.5-Coder technical report. *arXiv preprint arXiv:2409.12186*.
- John Platt. 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers*.
- Federico Ranaldi, Elena Sofia Ruzzetti, Dario Onorati, Leonardo Ranaldi, Cristina Giannone, Andrea Favalli, Raniero Romagnoli, and Fabio Massimo Zanzotto. 2024. Investigating the impact of data contamination of large language models in text-to-SQL translation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13909–13920, Bangkok, Thailand. Association for Computational Linguistics.
- Robert Richardson. 2026. What predicts correctness in text-to-SQL? a selective-prediction study. *arXiv preprint arXiv:2607.06799*.
- Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. 2005. *Algorithmic Learning in a Random World*. Springer.

An Yang et al. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

An Yang et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Tao Yu et al. 2018. Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-SQL task. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

Bianca Zadrozny and Charles Elkan. 2002. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Ej Zhou, Lucas Resck, Zheng Hui, and Anna Korhonen. 2026. Lower-resource, higher scores: Language bias in LLM evaluators. *arXiv preprint arXiv:2607.14480*.

A Comparison with Zhou et al. (2026)

Table 7 places the two studies side by side. The designs differ in what is held fixed, where the label comes from, and what the threshold is calibrated to.

axis	Zhou et al. (2026)	this work
evaluated artifact	free-text response	a formal SQL program, literally identical across languages
label source	human or model preference	execution against the database
calibration target	acceptance fraction	error rate among executed queries
languages	23	7
headline effect	acceptance-rate differences up to 43 points at >90% pairwise accuracy	14% to 27% of identical SQL decided differently at AUROC gaps ≤ 0.054
harm direction	unsafe acceptance in lower-resource languages	both directions, and which dominates is verifier-dependent

Table 7: Design comparison.

Their language-specific percentile thresholds correspond closely to the quantile transport we evaluate in Section 7.2, which we treat as a concurrent baseline.

B Coverage, discrimination, calibration, and transport detail

Full-data coverage and AUROC. Table 8 repeats the Section 4 analysis on all 6,000 candidates from 1,200 questions rather than on test databases alone. This is a descriptive, in-sample view: the threshold τ is refit on the same 6,000 English scores

it is then evaluated on, so the English risk is 10% by construction, and the numbers do not implement the canonical held-out policy of Section 3.4. They are reported only to show that the pattern is not an artifact of the particular test databases. Intervals are percentile bootstrap over questions, 4,000 resamples.

lang	AUROC	gap en- ℓ	risk	coverage	cov. gap
<i>Llama-3.1-8B</i> , $\tau = 0.7058$					
en	0.711	n/a	0.100	0.609	n/a
de	0.682	+0.029	0.109	0.650	-0.041
es	0.703	+0.008	0.109	0.696	-0.087
fr	0.698	+0.013	0.080	0.492	+0.116
ja	0.682	+0.028	0.110	0.653	-0.044
vi	0.678	+0.033	0.118	0.664	-0.055
zh	0.703	+0.008	0.103	0.639	-0.030
<i>Qwen2.5-7B</i> , $\tau = 1.0000$					
en	0.725	n/a	0.099	0.698	n/a
de	0.694	+0.031	0.093	0.603	+0.095
es	0.714	+0.011	0.094	0.639	+0.059
fr	0.704	+0.021	0.093	0.658	+0.040
ja	0.695	+0.030	0.099	0.644	+0.054
vi	0.685	+0.040	0.093	0.560	+0.138
zh	0.703	+0.022	0.098	0.627	+0.071

Table 8: Full-data version of Table 1, descriptive and in-sample. All 6,000 candidates, with τ refit on the same English scores it is evaluated on; not the held-out policy. AUROC gaps reach significance for de, ja and vi on Llama and de, fr, ja and vi on Qwen; every coverage gap is significant.

AUROC on test databases. Table 9 gives the per-language values summarized in Section 4.

	en	de	es	fr	ja	vi	zh
Llama	0.735	0.710	0.732	0.714	0.704	0.687	0.726
Qwen	0.754	0.716	0.743	0.729	0.701	0.706	0.717

Table 9: Per-language AUROC on the 82 test databases. Largest gap against English: 0.048 (Llama, vi) and 0.054 (Qwen, ja).

Four-way calibration. Table 10 expands Table 4. A and B each apply one strictly increasing map to every language, so their per-language coverage matches raw to three decimals and their spread is raw’s spread.

Per-language score transport. Tables 11 and 12 expand the means in Table 5.

Per-language pivot validation. Table 13 reports pivot against native per language, with paired database-bootstrap intervals, plus the literal-clean control described in Section 7.3.

calibrator	Llama	Qwen
raw (no correction)	0.223	0.148
A English labels only	0.223	0.148
B pooled multilingual labels	0.223	0.148
C per-language slope and intercept	0.040	0.083
D shared slope, per-language intercept	0.040	0.089

Table 10: Coverage spread over seven languages, test databases, single split. Every calibrator is fitted on all 599 calibration questions per language; D’s shared slope pools all seven languages. The 25 to 50 question figure in Section 7.1 refers to a separate calibration-error learning curve, not to these spreads.

C Complete transition counts

Table 14 gives Qwen’s decision-harm breakdown and Table 15 the complete transition counts for both verifiers.

D Full offset decomposition

Table 16 reports the paired score shift against English on identical SQL with identical labels, split by true label, over all 6,000 candidates from 163 databases. The bootstrap resamples databases.

E Translation audit, perplexity, and schema shift

Translation fidelity. Table 17 gives the per-language audit summarized in Section 3.5.

Three Vietnamese examples (quoted verbatim) where the literal was translated rather than preserved:

- **en** Show paragraph details for paragraph with text ‘Korea’ .
vi Hiện thị chi tiết về đoạn văn bản có chủ đề là ‘Hàn Quốc’ .
- **en** Show names, results and bulgarian commanders of the battles with no ships lost in the ‘English Channel’ .
vi Hiện thị tên , kết quả và chỉ huy của tất cả các trận chiến ở ‘Eo biển Măng-sơ’ và không có mất mát nào về tàu .
- **en** What are the notes of the death events which has substring ‘East’?
vi Cho biết những ghi chú có chứa chuỗi con ‘Phía Đông’ và liên quan đến các sự kiện thiệt mạng .

method	lang	flip	unsafe	lost	overlap	cov.
<i>Llama-3.1-8B</i> , $\tau = 0.7549$						
raw	de	13.9%	12.8%	5.8%	0.780	0.579
raw	es	14.9%	13.2%	2.6%	0.778	0.646
raw	fr	14.6%	1.6%	14.0%	0.737	0.423
raw	ja	19.5%	10.1%	9.0%	0.702	0.566
raw	vi	17.3%	16.9%	6.5%	0.736	0.595
raw	zh	15.9%	11.7%	7.1%	0.749	0.565
offset	de	14.7%	13.4%	5.7%	0.770	0.587
offset	es	13.8%	8.2%	5.6%	0.781	0.579
offset	fr	12.1%	9.5%	4.7%	0.805	0.579
offset	ja	20.0%	12.1%	8.6%	0.698	0.580
offset	vi	17.6%	16.0%	7.4%	0.729	0.583
offset	zh	15.7%	11.1%	7.3%	0.750	0.556
affine	de	13.9%	11.1%	6.9%	0.776	0.556
affine	es	13.8%	8.2%	5.6%	0.781	0.579
affine	fr	12.1%	9.5%	4.7%	0.805	0.579
affine	ja	19.9%	9.7%	10.4%	0.690	0.542
affine	vi	17.6%	16.0%	7.4%	0.729	0.583
affine	zh	15.7%	11.1%	7.3%	0.750	0.556
quantile	de	13.8%	8.4%	8.4%	0.771	0.524
quantile	es	13.4%	7.0%	6.2%	0.783	0.560
quantile	fr	11.3%	7.2%	6.5%	0.811	0.537
quantile	ja	19.3%	8.6%	10.7%	0.695	0.529
quantile	vi	17.2%	15.6%	7.7%	0.733	0.573
quantile	zh	16.0%	8.8%	9.1%	0.740	0.526
isotonic	de	13.6%	7.0%	9.1%	0.770	0.505
isotonic	es	13.2%	3.9%	9.4%	0.773	0.492
isotonic	fr	11.4%	8.2%	6.4%	0.809	0.539
isotonic	ja	19.1%	8.2%	11.3%	0.695	0.517
isotonic	vi	16.8%	9.1%	11.8%	0.719	0.487
isotonic	zh	15.6%	8.8%	8.6%	0.747	0.531

Table 11: Per-language transport results on test databases, Llama-3.1-8B.

Perplexity. Table 18 gives mean question log-perplexity under the Llama verifier itself, over 150 questions and 750 candidates, against the paired confidence shift. Across the six language means $r = +0.249$ (six points, descriptive only); across the 4,500 individual candidate-language pairs $r = +0.043$.

Schema shift. Table 19 separates language shift on the databases the threshold was fitted on from language shift plus schema shift on held-out databases. Regime (a) bootstraps over roughly 600 questions; regimes (b) and (c) bootstrap over roughly 80 databases, which is the unit that actually varies under schema shift and which widens the intervals considerably.

AUROC survives schema shift for both verifiers. Every calibration-against-held-out AUROC gap has an interval straddling zero, and for Llama the held-out databases rank marginally better than the calibration databases.

method	lang	flip	unsafe	lost	overlap	cov.
<i>Qwen2.5-7B</i> , $\tau = 1.0000$						
raw	de	21.7%	4.9%	17.3%	0.673	0.499
raw	es	17.4%	4.1%	12.5%	0.737	0.537
raw	fr	20.1%	5.6%	14.5%	0.700	0.529
raw	ja	21.9%	8.8%	14.7%	0.683	0.554
raw	vi	27.2%	8.0%	22.1%	0.595	0.462
raw	zh	21.5%	10.5%	13.3%	0.690	0.564
offset	de	20.7%	9.9%	10.7%	0.709	0.608
offset	es	17.7%	7.0%	8.1%	0.749	0.625
offset	fr	18.2%	7.8%	9.1%	0.741	0.611
offset	ja	21.6%	12.1%	11.2%	0.701	0.616
offset	vi	24.4%	12.1%	13.3%	0.663	0.593
offset	zh	21.5%	15.8%	7.9%	0.712	0.669
affine	de	20.6%	9.7%	10.8%	0.709	0.601
affine	es	17.7%	7.0%	8.1%	0.749	0.625
affine	fr	18.2%	7.8%	9.1%	0.741	0.611
affine	ja	21.4%	11.9%	11.2%	0.702	0.614
affine	vi	24.4%	12.1%	13.3%	0.663	0.593
affine	zh	21.9%	15.4%	8.5%	0.706	0.660
quantile	de	20.7%	9.9%	10.7%	0.709	0.608
quantile	es	17.9%	7.0%	8.1%	0.747	0.626
quantile	fr	17.7%	6.6%	9.8%	0.743	0.591
quantile	ja	21.7%	11.5%	12.1%	0.696	0.601
quantile	vi	23.7%	12.6%	12.4%	0.673	0.604
quantile	zh	21.4%	16.5%	7.1%	0.716	0.682
isotonic	de	55.8%	0.0%	60.7%	0.086	0.053
isotonic	es	40.6%	1.2%	43.0%	0.342	0.217
isotonic	fr	57.2%	0.0%	62.0%	0.063	0.039
isotonic	ja	49.5%	0.8%	53.2%	0.194	0.122
isotonic	vi	52.2%	0.0%	56.3%	0.147	0.092
isotonic	zh	45.8%	0.4%	48.9%	0.251	0.156

Table 12: Per-language transport results on test databases, Qwen2.5-7B.

F Robustness to the operating point and the split

Threshold sensitivity. Table 20 refits the English threshold at 5%, 10%, and 15% risk targets on the canonical calibration split and applies each unchanged to every language on the test split. A target at or above the verifier’s unconditional English error rate on calibration data is met by accepting nearly everything, so the 15% row is trivial for both verifiers and the small flip rates there reflect near-total coverage, not robustness. At 5% Qwen executes only 16.7% of English queries, so its smaller flip rates there reflect abstention.

Qwen’s 10%-target threshold, 0.999997, sits near the score ceiling; rounding every stored probability and the threshold to float32 changes 0.2% of its test decisions (0.0% for Llama) but collapses its 1,534 distinct English scores to 575, so a lower-precision score store would erase most of the ordering the threshold relies on. Scores are stored at full double precision.

lang	flip		Δ flip [95% CI]	unsafe		clean	drop
	nat.	piv.		nat.	piv.	Δ flip	n
<i>Llama-3.1-8B</i> , $\tau = 0.7549$							
de	13.9%	9.8%	-4.1 [-6.3, -1.9]*	12.8%	2.7%	-4.0	95
es	14.9%	9.9%	-5.0 [-7.4, -2.5]*	13.2%	1.9%	-4.9	85
fr	14.6%	10.1%	-4.5 [-7.0, -2.0]*	1.6%	2.3%	-4.6	60
ja	19.5%	13.2%	-6.3 [-9.3, -3.3]*	10.1%	2.3%	-6.0	105
vi	17.3%	17.2%	-0.1 [-3.2, +3.1]	16.9%	3.7%	-1.4	155
zh	15.9%	12.6%	-3.3 [-5.7, -0.7]*	11.7%	3.7%	-3.1	70
<i>Qwen2.5-7B</i> , $\tau = 1.0000$							
de	21.7%	15.3%	-6.4 [-9.6, -3.1]*	4.9%	2.7%	-6.5	95
es	17.4%	14.0%	-3.4 [-6.3, -0.5]*	4.1%	3.9%	-3.4	85
fr	20.1%	14.9%	-5.2 [-8.5, -1.8]*	5.6%	3.3%	-5.1	60
ja	21.9%	20.1%	-1.8 [-4.8, +1.5]	8.8%	2.1%	-2.2	105
vi	27.2%	23.1%	-4.2 [-7.7, -0.6]*	8.0%	2.7%	-3.9	155
zh	21.5%	18.2%	-3.3 [-6.4, -0.4]*	10.5%	4.1%	-3.4	70

Table 13: Pivot against native, per language, test databases. Δ values are in percentage points; * marks an interval excluding zero. *clean* Δ *flip* restricts to items whose English literal survived MultiSpider’s own translation; *drop* n is how many candidates that removes.

lang	flip	lost	unsafe promotion		95% CI
			all	withheld	(all)
de	21.7%	17.3%	4.9%	7.3%	[2.1%, 8.4%]
es	17.4%	12.5%	4.1%	6.1%	[1.5%, 7.4%]
fr	20.1%	14.5%	5.6%	8.2%	[2.4%, 9.3%]
ja	21.9%	14.7%	8.8%	13.0%	[5.2%, 13.0%]
vi	27.2%	22.1%	8.0%	11.8%	[4.4%, 12.2%]
zh	21.5%	13.3%	10.5%	15.5%	[6.8%, 14.8%]

Table 14: Qwen2.5-7B, the counterpart of Table 2. Test databases, 3,005 candidates, 2,519 correct / 486 incorrect, $\tau = 1.0000$.

Repeated database splits. Table 21 reshuffles the 163 databases with seeds 0 to 9, refits the English threshold at the 10% target on the first 81, and evaluates on the remaining 82; seed 0 is the canonical split. Ranges are descriptive, not confidence intervals: the splits overlap and are not independent replications, and unlike the question-clustered intervals in the main tables they include the uncertainty from refitting the threshold on a different calibration set. The instability persists in every split.

On Llama the mean flip rate stays within 14.0% to 16.1%, pivot beats quantile transport in all ten splits, and Vietnamese has the highest unsafe promotion in every split while French has the lowest. On Qwen the mean flip rate ranges 13.1% to 22.3% because the refit threshold lands at different distances from the score ceiling (English coverage 61.0% to 83.5%), and pivot beats quantile in seven splits by up to 2.9 points and loses in three by up to 2.3, so its Qwen advantage is split-dependent. The canonical-split French figure of 1.6% unsafe

lang	correct		incorrect		Δ err	Δ cor	risk	
	e→d	d→e	e→d	d→e			en	ℓ
<i>Llama-3.1-8B</i>								
de	145	199	12	62	+50	+54	0.078	0.102
es	66	312	5	64	+59	+246	0.078	0.096
fr	352	29	50	8	-42	-323	0.078	0.068
ja	227	276	33	49	+16	+49	0.078	0.085
vi	164	255	19	82	+63	+91	0.078	0.107
zh	179	213	29	57	+28	+34	0.078	0.092
<i>Qwen2.5-7B</i>								
de	436	134	57	24	-33	-302	0.085	0.082
es	314	131	57	20	-37	-183	0.085	0.074
fr	366	154	58	27	-31	-212	0.085	0.079
ja	371	203	42	43	+1	-168	0.085	0.094
vi	557	148	74	39	-35	-409	0.085	0.087
zh	336	203	57	51	-6	-133	0.085	0.089

Table 15: Complete transition counts, both verifiers, test databases. e = execute, d = defer.

lang	correct	incorrect	κ	95% CI
<i>Llama-3.1-8B</i>				
de	-0.0010	+0.0163	-0.0173	[-0.0364, +0.0031]
es	+0.0280	+0.0461	-0.0180	[-0.0330, -0.0025]
fr	-0.0875	-0.1002	+0.0127	[-0.0026, +0.0284]
ja	+0.0031	+0.0419	-0.0387	[-0.0572, -0.0208]
vi	+0.0106	+0.0499	-0.0393	[-0.0609, -0.0169]
zh	+0.0033	+0.0173	-0.0140	[-0.0341, +0.0064]
<i>Qwen2.5-7B</i>				
de	-0.0425	-0.0470	+0.0046	[-0.0393, +0.0487]
es	-0.0293	-0.0280	-0.0013	[-0.0443, +0.0429]
fr	-0.0251	-0.0088	-0.0163	[-0.0605, +0.0286]
ja	-0.0496	-0.0324	-0.0172	[-0.0627, +0.0292]
vi	-0.0804	-0.0833	+0.0029	[-0.0538, +0.0573]
zh	-0.0584	-0.0674	+0.0090	[-0.0431, +0.0600]

Table 16: Mean score shift against English on correct and on incorrect candidates, and their difference κ . Section 6.1 reads Spanish, Japanese, and Vietnamese as class-dependent cases; all three are adverse-direction cases in which incorrect-query scores shift more favorably, with Japanese and Vietnamese the largest.

promotion on Llama ranges 0.0% to 2.3% across splits.

G Worked example

We trace one test-split candidate, `train:0#0` on the `department_management` database, through every stage. The gold label was checked by hand against the database.

1. **Question (English only at generation).** *How many heads of the departments are older than 56?* Gold SQL: `SELECT count(*) FROM head WHERE age > 56.`
2. **Candidate generation.** Qwen2.5-Coder-7B-Instruct samples five candidates from the English question and the schema. Can-

lang	numeric	quoted literals	affected	unsafe
de	97.7%	83.2%	22	12.8% → 13.4%
es	98.6%	83.2%	22	13.2% → 13.5%
fr	95.8%	87.8%	16	1.6% → 1.7%
ja	96.8%	73.3%	35	10.1% → 10.5%
vi	98.6%	47.3%	69	16.9% → 17.0%
zh	97.2%	85.5%	19	11.7% → 12.0%

Table 17: Preservation of numeric values (of 216 items carrying one) and of quoted literals (of 131), with the count of affected items out of 1,200. The last column is Llama’s unsafe-promotion rate on all items and on the literal-clean subset.

lang	mean logPPL	Δ vs en	mean Δ conf
en	4.415	+0.000	+0.0000
de	3.460	-0.955	-0.0104
es	3.583	-0.832	+0.0252
fr	3.568	-0.847	-0.0905
ja	3.745	-0.670	+0.0057
vi	4.155	-0.260	-0.0017
zh	4.743	+0.328	+0.0012

Table 18: Question perplexity against confidence shift, Llama-3.1-8B. French and Spanish sit at practically the same perplexity and shift in opposite directions.

didate 0 is `SELECT COUNT(*) FROM management m JOIN head h ON m.head_id = h.head_id WHERE h.age > 56`; candidates 1 to 4 match the gold query.

3. **Execution label.** The `head` table has 10 rows, 5 with age above 56, so the gold query returns 5. Only 5 heads appear in `management`, and 3 of those are older than 56, so candidate 0 returns 3. Result sets differ; the label is *incorrect*. This label is fixed for every language below.
4. **Paired scoring.** Each verifier scores the same candidate under the seven question renderings (Vietnamese: *Có bao nhiêu bộ trưởng có tuổi lớn hơn 56?*), holding SQL, schema, and database fixed.
5. **Threshold.** Fit once on calibration-split English scores at the 10% risk target: $\tau_{\text{en}} = 0.7549$ (Llama), 0.999997 (Qwen).
6. **Decisions.** Table 22 applies each threshold unchanged. Under Llama the English question defers this incorrect query, and German, Spanish, Japanese, and Vietnamese execute it: four unsafe promotions from one candidate. Spanish clears the threshold by 1.1×10^{-5} . Under Qwen only Japanese executes it.

lang	(a) calibration databases		(b) held-out databases	
	cov.	gap [95% CI]	cov.	gap [95% CI]
<i>Llama-3.1-8B</i>				
en	0.579	n/a	0.544	−0.035 [−0.096, +0.025]
de	0.620	+0.041	0.579	−0.001 [−0.061, +0.062]
es	0.650	+0.071	0.646	+0.066 [+0.002, +0.126]
fr	0.458	−0.122	0.423	−0.157 [−0.218, −0.093]
ja	0.616	+0.037	0.566	−0.014 [−0.077, +0.052]
vi	0.611	+0.031	0.595	+0.016 [−0.047, +0.080]
zh	0.610	+0.030	0.565	−0.015 [−0.076, +0.047]
<i>Qwen2.5-7B</i>				
en	0.643	n/a	0.610	−0.033 [−0.095, +0.028]
de	0.536	−0.107	0.499	−0.144 [−0.209, −0.076]
es	0.560	−0.083	0.537	−0.106 [−0.172, −0.038]
fr	0.595	−0.048	0.529	−0.114 [−0.183, −0.046]
ja	0.606	−0.037	0.554	−0.089 [−0.158, −0.018]
vi	0.503	−0.140	0.462	−0.182 [−0.245, −0.117]
zh	0.549	−0.094	0.564	−0.080 [−0.145, −0.015]

Table 19: Coverage under the fixed English threshold. Gaps are against English on the calibration databases. Every (a) gap is significant; in (b) the database-level intervals are roughly twice as wide. Regime (c), schema shift alone with the language held at English (Section 6), closely matches the `en` row of (b) shown here.

verifier	target	τ_{en}	en cov.	flip range	unsafe range
Llama	5%	0.9466	24.6%	9.2% to 13.8%	0.0% to 2.7%
Llama	10%	0.7549	54.4%	13.9% to 19.5%	1.6% to 16.9%
Llama	15%	0.0676	99.1%	0.6% to 1.5%	0.2% to 2.7%
Qwen	5%	1.0000	16.7%	11.0% to 14.7%	0.0% to 1.9%
Qwen	10%	0.999997	61.0%	17.4% to 27.2%	4.1% to 10.5%
Qwen	15%	0.0000	99.0%	1.2% to 2.6%	1.2% to 2.1%

Table 20: Per-language flip and unsafe-promotion ranges (six non-English languages, test split) at three risk targets. *en cov.* is English coverage at that threshold. The 15% target lies at the unconditional error rate and is met by accepting nearly everything.

H Architecture screening and reproducibility

Within-model sensitivity. Each configuration is compared against *its own* English decisions, so baseline disagreement between models cancels. Table 23 reports both regimes. In *risk*, each model uses its own English threshold at the 10% target. In *covmatch*, thresholds are set to a common English coverage, so a saturated model cannot look stable merely by abstaining. Read ΔI only under *covmatch* when either model saturates.

Reproducibility notes. Four defects materially changed results. None raised an error; each produced plausible output.

1. **Dataset variant** (Section 3.1) fabricates a large, convincing cross-lingual effect.
2. **Batched inference** perturbs scores by up to 6×10^{-2} , the magnitude of the effect studied.

	Llama	Qwen
English coverage	54.4% to 73.5%	61.0% to 83.5%
mean flip, six languages	14.0% to 16.1%	13.1% to 22.3%
worst-language flip	15.1% to 19.5%	17.5% to 28.5%
pivot − quantile (pp)	−3.0 to −1.2 (10/10)	−2.9 to +2.3 (7/10)
unsafe promotion, de	8.7% to 17.4%	4.0% to 8.1%
unsafe promotion, es	10.5% to 16.0%	4.1% to 8.1%
unsafe promotion, fr	0.0% to 2.3%	4.2% to 8.5%
unsafe promotion, ja	10.1% to 16.1%	6.2% to 9.9%
unsafe promotion, vi	12.9% to 19.2%	5.0% to 10.5%
unsafe promotion, zh	9.5% to 13.6%	6.8% to 12.3%

Table 21: Ranges over ten database splits (seeds 0 to 9), English threshold refit at the 10% target on each split’s calibration half. Unsafe promotion is the share of all held-out incorrect candidates. The parenthetical count is the number of splits in which pivot flips fewer decisions than quantile transport. Per-seed rows in `results/robustness.md`.

	Llama, $\tau = 0.7549$		Qwen, $\tau = 0.999997$	
lang	score	decision	score	decision
en	0.469	defer	0.99593	defer
de	0.852	execute	0.99966	defer
es	0.755	execute	0.99331	defer
fr	0.593	defer	0.26894	defer
ja	0.777	execute	0.9999994	execute
vi	0.818	execute	0.99990	defer
zh	0.469	defer	0.81758	defer

Table 22: Verifier scores and decisions for candidate `train:0#0`, labeled incorrect. Bold marks an unsafe promotion relative to the English decision. Qwen scores are shown to five decimals because its threshold sits at 0.999997.

Length-sorting and explicit `position_ids` both made it worse. It is batch-shape kernel nondeterminism, not padding.

3. **Score clipping** at $\varepsilon = 10^{-6}$ before a logit transform collapsed 10,845 distinct Qwen scores to 5,619, destroying ordering. Caught only by asserting AUROC invariance under monotone calibration; it moved 0.0196, which is impossible. Use $\varepsilon = 10^{-12}$.
4. **Per-database caps** silently bound sample size: Spider’s dev split has 20 databases, so any cap bounds the sample far below the requested value.

Reading generated artifacts back, and asserting invariants a correct implementation must satisfy, caught more real defects than testing did.

comparison	regime	A	B	ΔI [95% CI]
verbalizer	risk	23.9%	25.0%	+1.1 [−3.4, +5.4]
	covmatch	23.9%	25.3%	+1.5 [−3.1, +5.6]
dense vs MoE	risk	16.1%	16.1%	+0.1 [−4.5, +4.2]
	covmatch	16.1%	16.3%	+0.2 [−4.1, +4.0]
bf16 vs 4-bit	risk	16.1%	15.7%	−0.4 [−3.0, +2.1]
	covmatch	18.1%	15.7%	−2.4 [−6.1, +0.8]

Table 23: Mean flip rate over six non-English languages for configuration A and configuration B, and their difference ΔI in percentage points. Verbalizer compares `Yes/No` against `CORRECT/INCORRECT` on Qwen2.5-7B; the other two rows compare Qwen3-4B against Qwen3-30B-A3B and against its own 4-bit quantization (Yang et al., 2025). Prespecified 300-question subset, 39 databases, 1,500 candidates.