

Silent Locale Corruption: Diagnosing and Mitigating Chinese-Locale Semantic Failures in LLM Tool Calling*

Ruiyang Zhang

Flock.io & Ryonix Labs Inc.

zhangruiyang36@gmail.com

Abstract

When a language-model agent fills tool-call arguments, the argument value is the action: 转一万二 (“transfer twelve thousand”) must become `transfer(12000)`, and 两斤苹果 (“two jin of apples”) must become `order(weight_kg=1.0)`. We show that today’s LLMs silently corrupt such values when Chinese locale semantics must be converted into canonical arguments — producing schema-valid, confidently wrong actions that no validator catches. We introduce ZhTool-Locale, a diagnostic benchmark with machine-computed gold values (lunar-calendar libraries, deterministic numeral and unit parsers), paraphrase triplets, and a pre-registered audit protocol; we report 95% bootstrap confidence intervals and significance tests for every headline claim. Across six models from five families — including two Chinese-trained — we find: (1) lunar-calendar conversion degrades significantly in every model (all $p \leq .031$) and catastrophically in five of six (0–28%, vs. ~100% on fixed-date controls); a pre-registered matched-pair control refuted our initial memorization hypothesis — explicit lunar dates are, if anything, easier than festival names — indicating genuine conversion difficulty rather than recall; (2) Chinese-centric training confers a sizable advantage at the large tier (deepseek-chat: 93.3% overall, 81% lunar — the strongest model tested) but does not rescue the 7B tier (qwen2.5: lunar 3%); (3) paraphrase disagreement detects scattered errors (precision up to 0.92) but is structurally blind to consistent mis-conversion, which we quantify; (4) a deterministic converter restores 94.8–100% accuracy with significant gains for three models (all $p \leq .0012$), but routing is tier-governed: frontier models route voluntarily (93–99%), a mid-tier model improves only under an explicit mandate, and two 7–8B mod-

els cannot be made to route at all (gains non-significant). The sharpest finding is metacognitive rather than arithmetic: that a model cannot compute a lunar conversion inline is unsurprising — the conversion needs astronomical tables — but the models do not know that they cannot, and answer confidently instead of delegating; at the 7–8B tier they decline to route even when the converter is supplied and its use is mandated. Deployable-tier safety must therefore be architectural rather than prompted. We release the benchmark as a generator whose golds are computed on demand, with a per-category deployment card.

1 Introduction

Large language models increasingly act rather than answer: they book tickets, transfer funds, and place orders by emitting structured tool calls. In this setting, the unit of correctness shifts from the fluency of a response to the semantic accuracy of argument values — a date string, an amount, a weight — because the value is executed. An error here is unlike a hallucinated sentence: it is silent. The JSON validates, the API returns 200, and the user discovers the problem when the train ticket is for the wrong day.

This paper studies a specific, deployment-critical source of such errors: locale semantics. Hundreds of millions of users will address agents in Chinese, where everyday expressions require non-trivial conversion into canonical machine values: 一万二 is 12,000 (not 10,002); 两斤 is 1.0 kg (1 斤 = 0.5 kg); 大后天 is “in three days”; and 中秋节 falls on a lunar date whose Gregorian equivalent requires astronomical tables, not arithmetic. Whether models perform these conversions correctly — not whether they answer in the right language — determines whether the executed action matches the user’s intent.

Prior work surrounds this question without answering it. Multilingual tool-calling studies report

*Benchmark, harness, raw outputs and reproducibility manifest: <https://github.com/ryonzhang/zhtoollocale>

that models emit parameter values in the wrong language (Luo et al., 2026); cross-calendar benchmarks test lunar conversion as explicit question answering (Miao et al., 2026); function-calling robustness studies perturb English requests and attribute most resulting failures to surface-form evaluation artifacts (Rabinovich and Anaby-Tavor, 2025). None measures whether the converted value inside an executed tool call is semantically correct. Our evaluation design answers the artifact critique directly: gold targets are canonical values (12000, 1.0, ISO dates), not surface strings, so exact-match scoring cannot be confounded by paraphrase variation — failures that remain are real.

We built ZhToolLocale under three discipline commitments. First, computed golds: every gold value is produced by a deterministic library or parser (lunardate for lunar dates, an ephemeris for solar terms, rule-based numeral and unit converters) or verified by a native speaker; no gold is eyeballed. Second, paraphrase triplets: every probe appears in three meaning-preserving phrasings, making consistency a first-class measured quantity. Third, a pre-registered audit protocol: ambiguous probes are flagged before any model runs; cross-model consensus against a gold triggers re-verification of the gold; API errors are scored as errors, never as model failures. This protocol fired repeatedly during our own study — exposing a task-validity bug in our harness (a missing reference date), confirming one challenged gold (清明 2027), removing one genuinely ambiguous probe (明年 腊八), and, most importantly, refuting our own pre-registered memorization hypothesis (§4.3). We report these events as results about benchmark construction, not embarrassments.

Our contributions:

- ZhToolLocale, a 12-category Chinese-locale tool-calling benchmark with computed golds, paraphrase triplets, layer-tagged probes (argument conversion, tool selection, recovery), and a deterministic harness — released as a generator that mints fresh, gold-verified probes on demand, making it contamination-resistant by construction.
- Findings across six models from five families (incl. two Chinese-trained), all with bootstrap CIs: lunar conversion degrades significantly in every model and catastrophically in five of six, while fixed-date controls pass at $\sim 100\%$; a pre-registered matched-pair control refuted

our initial memorization hypothesis (its decisive cell was zero across all models) — the failure is genuine conversion difficulty, and we report the refuted hypothesis as a first-class result; Chinese-centric training helps at the large tier (deepseek-chat strongest overall) but not at 7B (qwen); and a memorized-constant effect (清明 = April 4, true 2024–25, wrongly applied to 2027) persists in three models.

- Detection and its limits: paraphrase disagreement predicts wrongness with precision up to 0.92 on scattered-error models, and we quantify the structural blind spot — consistent-but-wrong predictions, concentrated almost entirely in the lunar category — that consistency-based detection in the spirit of (Manakul et al., 2023) cannot see.
- Mitigation with significance tests: deterministic converters restore 94.8–100% accuracy, with significant gains for gemini-2.5-flash (+27.1pp, $p < .0001$), gpt-4o-mini (+44.8pp, $p < .0001$) and deepseek-chat (+7.3pp, $p = .0012$). Routing is tier-governed: frontier models route voluntarily (93–99%); the mid tier improves only under a mandate (41%→71%; the unmandated gain is not significant, $p = .067$); two 7–8B models cannot be routed by instruction (qwen 3%→3%, llama 2%→10%; gains non-significant). We also quantify mis-routing per probe: it is rare (7 probe instances across the entire study, 0–3.1% per condition) and confined to one family, so supplying the converter largely removes the error rather than mitigating it. Deployable-tier safety therefore requires architectural enforcement.
- A per-category, per-tier deployment card for practitioners, and a pinned reproducibility manifest.

2 Related Work

Multilingual tool calling. MLCL (Luo et al., 2026) is the closest prior work: it evaluates tool calling under multilingual interaction (including Chinese) and identifies parameter-value language mismatch — values emitted in the user’s language — as the dominant failure. Our question is orthogonal: we hold language handling fixed and ask whether the converted value is semantically

correct; a model can emit a perfectly English, perfectly formatted, wrong number. Multilingual agent benchmarks (MAPS (Hofman et al., 2026), Ticket-Bench (Almeida et al., 2025), ITC (Zhang and Zhu, 2026)) extend coverage across languages and regions but do not isolate conversion correctness with computed golds.

Cross-calendar temporal reasoning. SPAN (Miao et al., 2026) benchmarks calendar conversion (including Chinese Lunar) as explicit QA over a 100-year span, finds average accuracy of 34.5%, and proposes a tool-augmented agent reaching 95%. We complement SPAN: the conversion in our setting is implicit in a natural request and silently executed; and our routing experiment measures whether models voluntarily adopt the tool-based fix SPAN enforces by construction.

Robustness of function calling. Rabinovich and Anaby-Tavor (2025) perturbs English BFCL requests and expanded toolkits, observing 8–19% drops under rephrasing — but attributes 70–90% of those failures to exact-match evaluation artifacts over surface forms. Our design adopts their implied remedy: gold targets are canonical semantic values, immune to surface variation by construction, so the failures we report survive their critique.

Knowing when to use tools. MetaTool (Huang et al., 2024) showed LLMs struggle to recognize when a tool is needed. Our routing experiment specializes this to conversions where the model’s inline ability is illusory — it believes it can convert and silently cannot — and adds the available-vs-forced contrast, finding the awareness failure is tier-governed and, at the 7–8B tier, not correctable by instruction in either model tested.

Consistency-based detection. SelfCheckGPT (Manakul et al., 2023) detects hallucination via sampling disagreement, with a known limitation: confident, systematic errors evade it. We test the related but distinct signal of paraphrase disagreement on tool arguments and quantify the analogous blind spot per category and model; we do not run SelfCheckGPT’s sampling method itself, and frame our detection results accordingly.

Classical temporal and numeric normalization. Deterministic recognition and normalization of temporal expressions is a long-standing NLP task (TimeML (Pustejovsky et al., 2003); HeidelTime (Strötgen and Gertz, 2010); SUTime (Chang and Manning, 2012)). Our converter tools operationalize this tradition as agent-callable functions; our results measure when LLMs fail to perform inline

what these classical systems solved deterministically — and whether they will delegate to them.

Positioning. Each neighbor covers one or two axes — language handling, calendar QA, English paraphrase robustness, tool awareness, consistency detection. ZhToolLocale occupies their intersection: semantic-value correctness of executed tool arguments under Chinese locale semantics, across deployment tiers, with consistency measured and mitigation tested under voluntary and mandated routing.

3 The ZhToolLocale Benchmark

3.1 Design principles

Action-layer evaluation. Each probe is a natural Chinese request plus a tool signature; the model must emit the tool’s arguments. Correctness is judged on canonical argument values — ISO dates, numbers, kilograms — so the score measures the semantics that would execute.

Computed golds. Gold values are produced by code, not judgment: `lunardate` for lunar↔Gregorian conversion, an ephemeris for the 清明 solar term, deterministic parsers for 万/亿 numerals and 市制 units; remaining categories are author-verified (native speaker). The generator can mint unlimited fresh probes with correct golds for any reference date — renewable and contamination-resistant by construction, in the spirit of continuously refreshed benchmarks (Jain et al., 2024) and contamination-free function-calling evaluation (Maekawa et al., 2025).

Paraphrase triplets. Every focused-set probe has three meaning-preserving phrasings (drafted with an LLM writing assistant and author-reviewed for meaning preservation; tooling recorded in the reproducibility manifest), yielding self-consistency as a measured quantity (§5).

Failure attribution and ERR discipline. The harness separates [ERR] (API failure, empty or unparseable output) from FAIL — errors are excluded from both numerator and denominator. A retired model endpoint during our study produced 0% scores that were entirely artifacts; the ERR discipline exists for exactly this reason. The paper’s quantitative analyses focus on the argument-conversion layer; tool-selection and recovery behaviors surface through the routing experiment (§6).

Pre-registration and adversarial gold defense.

Foreseeably-ambiguous probes are flagged before any model runs. Cross-model consensus against a gold triggers library re-verification: this confirmed one challenged gold (清明 2027 = April 5; ephemeris instant 17:29 CST) and removed one probe (明年腊八) whose two readings are both calendar-valid. Hypotheses about mechanisms are likewise pre-registered and tested (§4.3). A full raw-to-canonical changelog is released.

3.2 Probe suite

Twelve categories instantiate three conversion mechanisms: calendar systems (relative dates; lunar dates split by type — famous-festival-with-hint, festival-name-only, pure-lunar-date, leap-month — plus fixed-date and solar-term controls), numeral semantics (万/亿 grouping, internal zeros, elliptical forms, capital 大写 forms, mixed digit-character, long chains), and quantity/format semantics (市制 units incl. fractional 半, classifiers, full-width characters, code-switching, names, addresses, phone formats). Tiers escalate difficulty: tier-1 (23 seed probes), tier-3 (23 hardened), and a focused set (35 groups $\times$ 3 paraphrases = 105 scored probes after audit). The controls are load-bearing: fixed-date holidays (国庆, 圣诞) isolate the lunar mechanism — a model failing lunar but passing controls fails conversion, not dates in general.

3.3 Models and conditions

Six models spanning five families and three tiers: gemini-2.5-flash and gemini-2.5-flash-lite (Google); gpt-4o-mini (OpenAI); deepseek-chat (official api.deepseek.com); qwen2.5:7b and llama3.1:8b (local, Ollama). The two local models are served through Ollama 0.30.4 using the default weights for the qwen2.5:7b and llama3.1:8b tags (content SHAs pinned in the manifest), which are *quantised* rather than full precision; the four hosted models are served at provider precision, which we cannot inspect. We did not run a full-precision control for the local pair, and we flag this as a confound on any small-tier conclusion (Limitations). Two are Chinese-trained (DeepSeek, Qwen). All runs at temperature 0 (documented stability: 95–100% identical outputs across repeats); the reference date is injected into every prompt; exact model versions, endpoints and access dates are pinned in a released reproducibility manifest. Conditions (§6): BASELINE (no tools), AVAILABLE (three

Model	Tier-1 (23)	Focused (105)
deepseek-chat	95.7% [87–100]	93.3% [89–98]
gemini-2.5-flash	100.0% [100–100]	75.2% [67–83]
gemini-2.5-flash-lite	95.7% [87–100]	60.0% [50–70]
qwen2.5:7b	78.3% [61–96]	55.9% [46–65]
gpt-4o-mini	91.3% [78–100]	54.3% [45–64]
llama3.1:8b	47.8% [30–70]	25.7% [18–34]

Table 1: Master accuracy with 95% bootstrap CIs. deepseek-chat (Chinese-trained, large) is the strongest model tested; qwen2.5-7B (Chinese-trained, small) is not rescued by training language.

deterministic converter tools offered — lunar, numeral, unit — executed by real code on call), FORCED (converters plus a one-line mandate), and HINT (one line of conversion rules, no tools). All intervals below are 95% bootstrap CIs (10,000 resamples).

4 Findings

4.1 Both tier and training matter — and neither saves the small tier

Two regularities emerge. First, a capability-tier gradient: large models outperform small ones, and llama3.1-8B fails broadly (25.7% [18–34]). Second, training matters at the top: deepseek-chat (93.3% [89–98]) significantly outperforms the strongest Western model tested (gemini-2.5-flash, 75.2% [67–83]; flash-lite gap itself +15.2pp, $p = .023$ raw, which does not survive a joint correction over all 13 tests and is therefore reported descriptively, §4.5) — Chinese-centric training confers a real advantage at the large tier. But it does not transfer down: qwen2.5-7B (55.9%) sits beside gpt-4o-mini (54.3%). Locale competence is not a property a family inherits; it is a property a model earns with scale and exposure together.

Is the problem already solved? deepseek-chat’s strong numbers (93.3% overall; 99% voluntary routing) invite the conclusion “just use deepseek-chat.” Three facts resist it. First, deployment reality: many production stacks — Western platforms, vendor-constrained enterprise procurement, on-device or cost-capped settings — do not, and sometimes cannot, use deepseek-chat; whether a given deployed agent is locale-safe is not answered by the existence of one safe model. Second, even deepseek-chat degrades significantly on lunar conversion (81% vs. 100% on fixed-date controls, $p = .0002$), and the failures are silent and transaction-stakes; “81% on cultural dates” is not a property one ships a payment agent on. Third, the de-

playable tier — where cost forces many real deployments — fails regardless of training family (qwen2.5-7B lunar 3%; llama3.1-8B broadly) and cannot be repaired by instruction (§6). The contribution is a map of where silent locale corruption lives and how to contain it, not a claim that no model handles any case.

4.2 Lunar conversion degrades in every model — catastrophically in five of six

Lunar conversion is the dominant failure surface: significant degradation in all six models, catastrophic (0–28% family-level accuracy) in five. Even the strongest model, deepseek-chat, is significantly worse on lunar dates than on fixed-date controls ($p = .0002$). Failures are silent and plausible — schema-valid tool calls whose dates are wrong by one day to, in small-model outliers, several years (most errors fall within 1–40 days), with off-by-one (端午 as 06-20 vs 06-19) the most insidious: no validator or casual human review catches it before execution.

4.3 A pre-registered memorization hypothesis, refuted

Our initial decomposition suggested an attractive mechanism: that apparent lunar competence reflects recall of famous festival dates rather than conversion ability. We pre-registered this hypothesis and built a matched-pair control: for five famous festivals, paired probes share an identical computed gold, one phrased with the festival name only (中秋节那天...), its twin with the identical lunar date and no name (农历八月十五那天...). The decisive cell — name correct but identical date wrong — would indicate recall without conversion.

The hypothesis is not supported: the decisive cell is zero in every model able to attempt it. We flag a conditional-power limitation — the decisive cell can fire only when the name version is correct, and name-accuracy is 0% for three of the five models, so the control is informative (a genuine null) only for gemini-2.5-flash and deepseek-chat and uninformative below them. Where the pair members differ, the explicit lunar date is the easier one (4 reverse cases): the festival name adds a year-anchoring step that can trigger recall of a past year’s instance (e.g., 元宵 resolved to an already-passed 2026-03-03 by name, but correctly to 2027-02-20 by explicit date). We therefore drop the memorization framing entirely: the lunar failure

is genuine, uniform conversion difficulty. We report this refutation as a first-class result — pre-registered mechanisms should be tested, and the negative outcome materially changes what mitigation can work (no prompt can supply an ephemeris; §6).

4.4 Numerals and units: largely solved at the top, fragile below

Capital-form (大写), mixed-digit, and elliptical numerals are at or near ceiling for every model except llama. The residual danger concentrates in internal-zero numerals (一亿零五十万 = 100,500,000) and fractional 市制 units (二两半 = 0.125 kg) at the smaller tiers. Locale conversion is not uniformly hard: the danger lives in specific, identifiable sub-categories and tiers — the granularity the deployment card captures (§7).

4.5 Statistical robustness

We report 13 hypothesis tests in two pre-registered families. Under Holm–Bonferroni *within* family every result survives: all six lunar-versus-control tests stay significant (largest adjusted $p = .031$) and all three significant routing gains stay significant (largest adjusted $p = .0048$). Under the most conservative alternative — one Holm family over all 13 tests — 8 of 13 survive, and two claims do not: the flash–lite focused gap ($p = .023 \rightarrow .115$) and llama3.1:8b’s lunar contrast ($p = .031 \rightarrow .123$). We therefore report the flash–lite gap descriptively, and state the lunar result as significant in five of six models with llama3.1:8b directionally consistent but not robust to a joint correction. Effects at the ~ 100 pp scale are secure at this n ; the smaller contrasts are not, which is why they are demoted. Provider errors (22 across the twelve routing cells) are excluded from numerator and denominator; scoring every one as *incorrect* instead moves no cell by more than 3.4pp and changes no direction or significance. Fragmentation does not explain the failures either: within any one tokeniser the lunar and fixed-date-control probes fragment almost identically, yet accuracy between them differs by up to 100pp. Full correction tables, the error-sensitivity recomputation and the tokenisation measurement are in Appendices B–E.

4.6 A stale memorised constant

清明 is the one calendar probe in our suite that is *not* lunar: it is a solar term, fixed by the ephemeris, and it fell on April 4 in 2024 and 2025. It therefore

Lunar probe type	dschat	25-flash	lite	4o-mini	qwen	llama
famous + hint	100%	17%	0%	0%	0%	0%
festival-name-only	81%	33%	5%	0%	0%	0%
pure-lunar-date	50%	33%	0%	0%	17%	0%
leap-month	100%	0%	0%	0%	0%	0%
control: 清明 (solar)	100%	100%	0%	100%	0%	0%
control: 国庆 (fixed)	100%	100%	100%	100%	100%	0%
control: 圣诞 (fixed)	100%	100%	100%	100%	100%	100%

Table 2: Lunar decomposition (baseline). Every model degrades significantly on lunar versus its own fixed-date controls ($p \leq 0.0308$ in all six models; e.g., gpt-4o-mini 0% vs 100%, $p < .0001$; deepseek-chat 81% vs 100%, $p = .0002$). Three models also carry a stale 清明 constant (April 4, correct 2024–25, wrong for 2027); because they fail this control, the conversion-isolation argument rests on the fixed-date controls, which every model except llama passes.

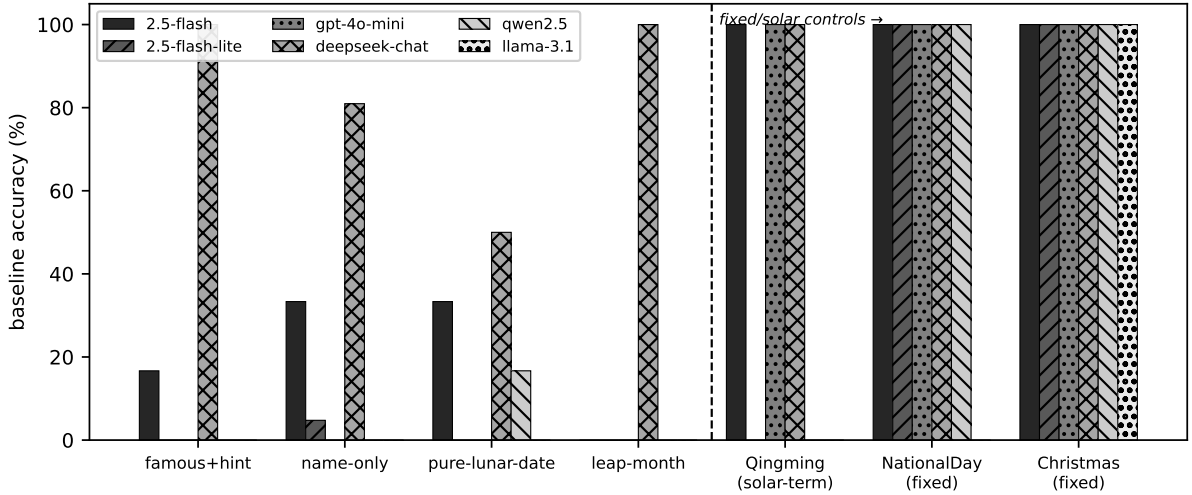


Figure 1: Lunar decomposition by probe type and model (baseline accuracy). Fixed-date and solar-term controls are visually separated at right; no model clears 50% on the conversion-requiring types except deepseek-chat.

Model	N✓D✓	N✓D✗(dec.)	N✗D✓	both ✗
gemini-2.5-flash	3	0	3	4
deepseek-chat	3	0	1	6
gemini-2.5-flash-lite	0	0	0	10
gpt-4o-mini	0	0	0	10
qwen2.5:7b	0	0	0	10

Table 3: Matched-pair memorization control (baseline, no tools). N = festival-name version, D = date version; ✓/✗ denote correct/incorrect. The pre-registered decisive cell (N✓D✗) is zero in every model; the reverse cell (name wrong, identical date right) is non-zero.

separates two failure modes that the lunar probes conflate. Three models — gemini-2.5-flash-lite, qwen2.5:7b and llama3.1:8b — score 0% on it, while gemini-2.5-flash, gpt-4o-mini and deepseek-chat score 100% (Table 2). The three failures reproduce the April 4 constant on a target year where it is wrong — a different pathology from the lunar failures. Those are an inability to compute; this is a memorised surface constant applied past its

validity window, and the model is confident precisely because it once was right. The two need different guards: forcing a converter catches the first, whereas the second would survive any amount of extra reasoning and is caught only by binding the value to a dated source. We observe this on one probe family and do not claim a general measurement of constant staleness.

5 Detection and Its Structural Blind Spot

We test whether paraphrase disagreement — the three phrasings of a probe not all yielding the same prediction — flags wrong answers at runtime with no ground truth.

It works for scattered-error models: a deployer can rephrase-and-check, escalating on disagreement. But it is structurally blind to confident systematic errors: counting consistent-but-wrong groups (all three paraphrases identical and wrong) yields a blind spot concentrated almost entirely in

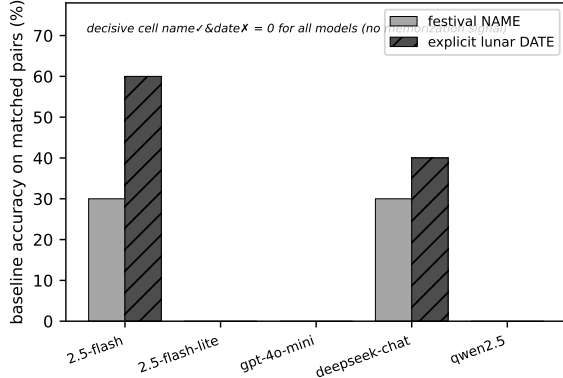


Figure 2: Memorization matched-pair control — a null result. Per model, baseline accuracy on the festival-name phrasing vs. the identical explicit-lunar-date phrasing. The pre-registered decisive cell (name correct, identical date wrong) is zero for every model; where the two differ, the explicit date is easier — refuting the recall hypothesis and indicating genuine conversion difficulty.

Model	Prec (WRONG DISAGREE)	Rec (DISAGREE WRONG)
gemini-2.5-flash	0.78	0.78
deepseek-chat	0.00	0.00
gemini-2.5-flash-lite	0.80	0.29
gpt-4o-mini	0.62	0.31
qwen2.5:7b	0.92	0.73
llama3.1:8b	0.89	0.62

Table 4: Disagreement-as-detector. High precision on scattered-error small models (qwen 0.92, llama 0.89). deepseek-chat’s 0.00/0.00 reflects that its only two wrong groups are consistent-wrong — invisible to disagreement by construction.

the lunar category — 10 of gpt-4o-mini’s 11 blind-spot groups are lunar; 8 of lite’s 9 — and the pattern persists at the very top: deepseek-chat’s only two wrong groups are both consistent-wrong lunar, so disagreement detection catches none of them. This blind spot is not an artifact of near-identical paraphrases. Measuring within-triplet dissimilarity directly, the suite averages 0.556 normalised character edit distance (0.611 Jaccard) between the three phrasings of a probe, and the lunar family — where the blind spot concentrates — is *more* diverse than the rest (0.616 vs 0.522), not less. The three phrasings differ substantially and the model converges on the same wrong value anyway, which is precisely what makes the error invisible to consistency checking. Models that systematically mis-convert defeat consistency-based detection precisely where conversion is hardest. Majority-vote repair is correspondingly marginal (−2 to +2 points): errors are consistent, so there is

no minority-correct vote to recover. Detection, not voting, is the value of the triplets; the cure must come from elsewhere.

6 Mitigation: Deterministic Tools Work — If the Model Calls Them

Frontier models route voluntarily and are fixed. gemini-2.5-flash, gpt-4o-mini and deepseek-chat route at 93–99% unprompted and reach 94.8–100% (all gains significant, $p \leq .0012$); AVAILABLE $\approx$ FORCED, so the mandate adds nothing — these models “know” they cannot convert in-line.

The mid tier under-routes from overconfidence. gemini-2.5-flash-lite routes 41% voluntarily; its unmandated gain is not significant (+6.5pp, $p = .067$). The mandate raises routing to 71% and accuracy to 81.5% [73–89] (forced-vs-baseline: +21.7pp, 95% CI [+13, +32], $p < .0001$) — mandating helps, but does not fully recover.

The 7–8B tier cannot be routed by instruction — in either model tested. qwen2.5-7B routes 3% whether the converter is available or mandated (available gain −1.1pp, forced −4.3pp, both NS); llama3.1-8B routes 2–10% (available +1.1pp, forced −1.2pp, both NS). The harness confirms qwen uses the native tool interface without format errors — it simply will not call the converters. At the deployable tier the fix cannot be prompted in; safety must be enforced architecturally, e.g., a deterministic pre-parser outside the model.

Mis-routing, quantified. Counting *calls* overstates mis-routing, because a model may legitimately chain — parse the numeral, then convert the unit — so a correct first step scores as a wrong tool. We therefore score per probe: a probe is mis-routed only when the converter its family requires is never called on it. On that definition, mis-routing is rare. Across all twelve routing cells there are 7 mis-routed probe instances in total: 6 for deepseek-chat (3 AVAILABLE, 3 FORCED) and 1 for gpt-4o-mini (FORCED); every other cell is zero. Per-condition rates are 0–3.1%. Every one of the 7 falls in `units_half`, where the model calls the numeral parser and never the unit converter (二两半 → 2, the numeral without the 斤 → kg step). Supplying the converter therefore largely removes the error rather than migrating it; the residual is a single-family slot-decomposition failure, and it is the case a deterministic pre-parser should own.

Model	BASELINE	AVAILABLE	FORCED	route% av→fo	$\Delta(\text{avail}-\text{base}), p$
gemini-2.5-flash	72.9% [64–81]	100.0%	100.0%	93→96%	+27.1, $p < .0001$
deepseek-chat	92.7% [88–98]	100.0%	100.0%	99→100%	+7.3, $p = .0012$
gpt-4o-mini	50.0% [40–59]	94.8% [90–99]	94.8% [90–99]	97→99%	+44.8, $p < .0001$
gemini-2.5-flash-lite	59.4% [50–69]	66.3% [57–76]	81.5% [73–89]	41→71%	+6.5, $p = .067$ (NS)
qwen2.5:7b	54.8% [44–65]	54.2% [44–64]	51.0% [42–60]	3→3%	-1.1, $p = .867$ (NS)
llama3.1:8b	25.0% [17–33]	26.6% [18–36]	26.2% [17–36]	2→10%	+1.1, $p = .883$ (NS)

Table 5: Routing experiment on locale-sensitive probes (95% CIs; paired bootstrap p for the AVAILABLE gain). The fix is near-total where models route; whether they route is tier-governed. Baseline here is the locale-sensitive subset ($n \approx 96$), so figures differ slightly from the full focused set in Table 1 (e.g., deepseek-chat 92.7% vs 93.3%). A few cells exclude [ERR] outputs (lite: 4; deepseek-chat forced: 3; llama forced: 12); see released results.

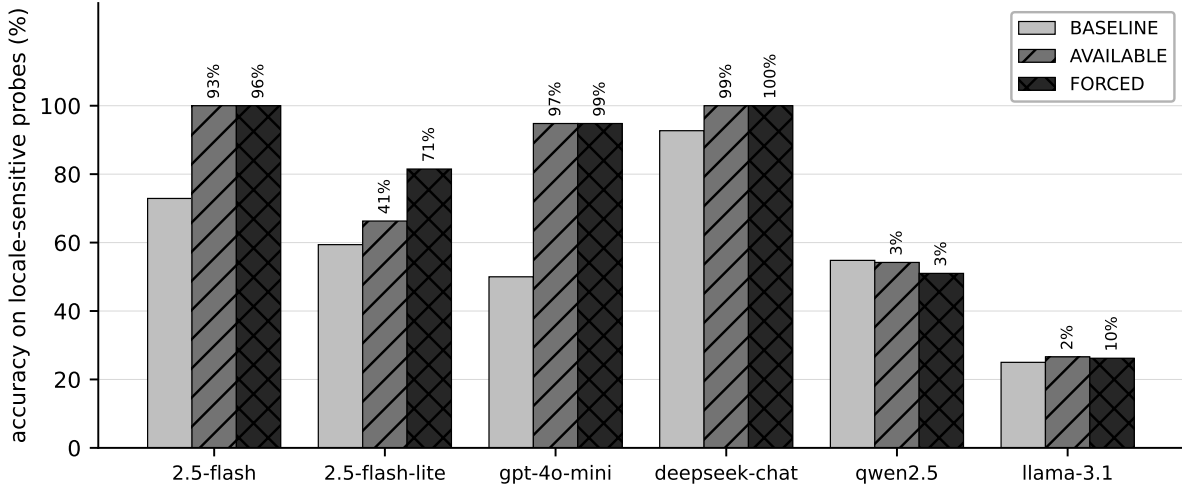


Figure 3: Routing gradient: BASELINE vs. converter-AVAILABLE vs. converter-FORCED accuracy on locale-sensitive probes, with the voluntary routing rate annotated above the AVAILABLE/FORCED bars. The fix is near-total where models route; the 7–8B tier does not route.

Prompt hints are tier-dependent and can harm the frontier. A single line of conversion rules helps weak tiers on units (gpt-4o-mini +17, qwen +7) but is net-negative for the strongest Western model (gemini-2.5-flash: mixed-digit -33, elliptical -11) and does not repair lunar conversion (per-model lunar deltas -3 to +7, all small); for deepseek-chat hints are negligible throughout (+3 lunar, 0 elsewhere) — consistent with §4.3: no reminder can substitute for an ephemeris. Prompt-level mitigation trades frontier accuracy to raise the floor.

7 Deployment Guidance

- Lunar / cultural-calendar slots: unsafe or cautionary for every model tested. Force deterministic tool routing; do not rely on the model, hints, or tool availability alone.
- Internal-zero and fractional-unit slots: safe at the top tiers, caution/unsafe below — shim or route at the deployable tier.

- Capital, mixed-digit, elliptical numerals: safe for most tiers; verify only on sub-8B models.
- Runtime monitoring: paraphrase-disagreement detection catches scattered errors cheaply but not systematic lunar mis-conversion — pair it with mandatory routing on calendar/quantity slots; never use it as the sole guard.
- Tier rule: prompt-level fixes are a top-tier privilege; at the deployable ($\leq 8B$) tier, locale safety must be architectural — enforced outside the model.

8 Conclusion

When agents act in Chinese, converting locale expressions into canonical tool arguments is a silent failure surface: schema-valid, plausible, executed. The failure is statistically robust — worst for calendar conversion, where every model degrades

Category	dschat	25-flash	lite	4o-mini	qwen	llama
lunar	CAUTION	UNSAFE	UNSAFE	UNSAFE	UNSAFE	UNSAFE
internal_zero	SAFE	SAFE	SAFE	UNSAFE	CAUTION	UNSAFE
elliptical	SAFE	SAFE	SAFE	UNSAFE	SAFE	UNSAFE
capital	SAFE	SAFE	SAFE	SAFE	SAFE	UNSAFE
mixed_digit	SAFE	SAFE	SAFE	SAFE	SAFE	UNSAFE
long_chain	SAFE	SAFE	UNSAFE	UNSAFE	SAFE	UNSAFE
units_half	SAFE	SAFE	CAUTION	CAUTION	UNSAFE	UNSAFE

Table 6: Deployment card (SAFE $\geq 95\%$ / CAUTION 80–95% / UNSAFE $< 80\%$) across all six models. deepseek-chat is the only model with no UNSAFE family; its sole below-ceiling family is lunar (81%, CAUTION). **The 95% and 80% cut-offs were chosen after the runs, for presentation, and — unlike the rest of the protocol — were not pre-registered**; read the card as a summary of measured accuracies binned for legibility, not as a calibrated risk model.

significantly, five of six catastrophically; a pre-registered control refuted memorization, establishing conversion difficulty; Chinese-centric training helps the large tier but not small models; and the deterministic fix works where models route to it, which the 7–8B tier will not do. Consistency-based detection catches scattered errors but misses systematic ones. We release ZhToolLocale as a renewable, computed-gold benchmark with a per-tier deployment card, framing locale-reliability — not just language — as a measurable, mitigable concern.

Limitations

Scale. Per-category cells are small (≈ 6 probes/group); CIs are reported everywhere and claims are framed as patterns. The released generator mints additional gold-verified probes for larger- n studies.

Single annotator. Human-verified gold categories (names, addresses, classifiers) were authored and verified by one native speaker; no inter-annotator agreement is reported, so for those categories the gold rests on a single judgement. The computed-gold categories — which carry every headline claim in this paper, including all lunar, numeral and unit results — are produced by deterministic libraries and are unaffected.

Harness realism. The harness is single-turn, the tools are mocked, and the reference date is injected into every prompt. It therefore measures raw model behaviour under favourable conditions, not an end-to-end agent stack: we show that the models get these conversions wrong, not that any particular deployed system fails this way in production. A multi-turn demonstration with real tool schemas and no injected reference date is the natural next experiment, and we have not run it. Our deployment

framing should be read accordingly.

Quantisation. The two local models were served through Ollama at its default quantised weights for these tags, while the four hosted models run at provider precision. Part of the “the deployable tier cannot be rescued” result may therefore be a quantisation effect rather than a pure scale or capability effect. A full-precision replication of the 7–8B pair would separate the two; we did not run one.

One model per training $\times$ scale cell. The claim that Chinese-centric training helps at the large tier but not at 7B rests on deepseek-chat and qwen2.5:7b — one model per cell, from different families rather than a scaled pair within one family. The design cannot separate training origin from family identity, and we state the pattern as an observation, not a general law.

Calendar generality. We do not test other computationally hard calendars (Hebrew, Ethiopian, Hijri). We therefore cannot say whether this is specific to Chinese locale semantics or an instance of a general weakness at calendar arithmetic that happens to surface here; the generator is built to support that comparison and it is the cleanest follow-up.

Scope. One language (Chinese), single-turn requests, mocked tools, temperature 0. We claim a case study with a generalizable recipe; a multi-locale extension (Japanese era names, German half-hour and decimal-comma conventions, French vigesimal numerals) is the planned next step and the generator is built for it.

Coverage asymmetry. deepseek-chat was added in the final experimental round and is now characterized in all tables (detection, deployment card, routing, hints). Raw outputs for all models are released. One excluded endpoint (gemini-2.0-flash, retired mid-study) and earlier third-party-served

DeepSeek variants are documented in the released appendix; no third-party-served results appear in this paper’s tables.

Mitigation realism. Converters are idealized (exact libraries); the residual `units_half` mis-routes show that real deployments still face slot-decomposition errors our tools do not model, even though mis-routing itself is rare.

Negative result. The memorization control (§4.3) is a five-festival, ten-pair design; its decisive cell fires only when the festival-name version is answered correctly, which holds for just two of five models — so the null is informative for those two and the control is uninformative below them; subtle recall effects below this design’s sensitivity cannot be excluded.

Ethics and Reproducibility Statements

Ethics. This study involves no human subjects or personal data; all probes are synthetic. The identified failure modes concern reliability rather than misuse, and disclosing them supports safer deployment; mitigation is released alongside diagnosis. AI-use disclosure: AI assistants were used for drafting, code, and experiment orchestration; all research decisions, data, analyses and claims are the author’s, and every citation was human-verified against its primary source.

Reproducibility. All runs use temperature 0 (95–100% repeat-identical outputs); exact model identifiers, endpoints and access dates are pinned in the released reproducibility manifest, together with library versions, the probe generator, harness, raw outputs, and the full raw-to-canonical changelog, at <https://github.com/ryonzhang/zhtoollocale>.

References

- Thales Sales Almeida, João Guilherme Alves Santos, Thiago Laitz, and Giovana Kerche Bonás. 2025. *Ticket-Bench: A kickoff for multilingual and regionalized agent evaluation*. *Preprint*, arXiv:2509.14477.
- Angel X. Chang and Christopher D. Manning. 2012. SUTime: A library for recognizing and normalizing time expressions. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC’12)*, pages 3735–3740.
- Omer Hofman, Jonathan Brokman, Oren Rachmil, Shamik Bose, Vikas Pahuja, Toshiya Shimizu, Trisha Starostina, Kelly Marchisio, Seraphina Goldfarb-Tarrant, and Roman Vainshtein. 2026. MAPS: A multilingual benchmark for agent performance and security. In *Findings of the Association for Computational Linguistics: EACL 2026*. ArXiv:2505.15935.
- Yue Huang, Jiawen Shi, Yuan Li, Chenrui Fan, Siyuan Wu, Qihui Zhang, Yixin Liu, Pan Zhou, Yao Wan, Neil Zhenqiang Gong, and Lichao Sun. 2024. *Meta-Tool benchmark for large language models: Deciding whether to use tools and which to use*. *Preprint*, arXiv:2310.03128.
- Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fan-jia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. 2024. *Live-CodeBench: Holistic and contamination free evaluation of large language models for code*. *Preprint*, arXiv:2403.07974.
- Zheng Luo, T Pranav Kutralingam, Ogochukwu N Okoani, Wanpeng Xu, Hua Wei, and Xiyang Hu. 2026. *Lost in execution: On the multilingual robustness of tool calling in large language models*. *Preprint*, arXiv:2601.05366.
- Seiji Maekawa, Jackson Hassell, Pouya Pezeshkpour, Tom Mitchell, and Estevam Hruschka. 2025. *Towards reliable benchmarking: A contamination free, controllable evaluation framework for multi-step LLM function calling*. *Preprint*, arXiv:2509.26553.
- Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. 2023. *SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models*. *Preprint*, arXiv:2303.08896.
- Zhongjian Miao, Hao Fu, and Chen Wei. 2026. SPAN: Benchmarking and improving cross-calendar temporal reasoning of large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*. ArXiv:2511.09993.
- James Pustejovsky, José Castaño, Robert Ingria, Roser Saurí, Robert Gaizauskas, Andrea Setzer, and Graham Katz. 2003. TimeML: Robust specification of event and temporal expressions in text. In *New Directions in Question Answering*.
- Ella Rabinovich and Ateret Anaby-Tavor. 2025. On the robustness of agentic function calling. In *Proceedings of the 5th Workshop on Trustworthy NLP (TrustNLP) at NAACL 2025*. ArXiv:2504.00914.
- Jannik Strötgen and Michael Gertz. 2010. HeidelTime: High quality rule-based extraction and normalization of temporal expressions. In *Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval-2010)*, pages 321–324.
- Zuoyu Zhang and Yancheng Zhu. 2026. *Enhancing tool calling in LLMs with the international tool calling dataset*. *Preprint*, arXiv:2603.05515.

A Condition and status labels

BASELINE — locale-sensitive probes, no converter offered. **AVAILABLE** — converters supplied as tools, use not required. **FORCED** — converters supplied *and* their use mandated in the system prompt. **HINT** — one line of conversion rules added to the prompt, no tools. **ERR** — the provider returned an error rather than a completion; excluded from numerator and denominator throughout (Appendix B).

B Multiple comparisons and error sensitivity

family	test	raw p	Holm p
lunar vs. fixed control	gemini-2.5-flash	< .0001	< .0006
	gemini-2.5-flash-lite	< .0001	< .0006
	gpt-4o-mini	< .0001	< .0006
	qwen2.5:7b	< .0001	< .0006
	deepseek-chat	.0002	< .0006
	llama3.1:8b	.0308	.0308
routing (AVAIL—BASE)	gemini-2.5-flash	< .0001	< .0006
	gpt-4o-mini	< .0001	< .0006
	deepseek-chat	.0012	.0048
	gemini-2.5-flash-lite	.0668	.2004
	qwen2.5:7b	.8666	1.000
	llama3.1:8b	.8828	1.000

Table 7: Holm–Bonferroni within each pre-registered family. Every result survives. Under a single joint family over all 13 reported tests, the flash–lite focused gap (.023 $\rightarrow$.115) and llama3.1:8b’s lunar contrast (.031 $\rightarrow$.123) do not.

model	cond	ERR	excl.	ERR=wrong
gemini-2.5-flash	AVAIL	0	100.0%	100.0%
gemini-2.5-flash	FORCED	0	100.0%	100.0%
gemini-2.5-flash-lite	AVAIL	4	66.3%	63.5%
gemini-2.5-flash-lite	FORCED	4	81.5%	78.1%
gpt-4o-mini	AVAIL	0	94.8%	94.8%
gpt-4o-mini	FORCED	0	94.8%	94.8%
deepseek-chat	AVAIL	0	100.0%	100.0%
deepseek-chat	FORCED	0	100.0%	100.0%
qwen2.5:7b	AVAIL	0	54.2%	54.2%
qwen2.5:7b	FORCED	0	51.0%	51.0%
llama3.1:8b	AVAIL	2	26.6%	26.0%
llama3.1:8b	FORCED	12	26.2%	22.9%

Table 8: Accuracy under the paper’s convention (errors excluded) versus the worst case (every provider error scored incorrect). Accuracies use the paper’s canonical scoring (variant group $\mathbb{f}\text{-L07}$ excluded, tool-call envelopes unwrapped), so the left column reproduces Table 5 exactly. Maximum shift 3.4pp (llama3.1:8b FORCED); no comparison changes direction or significance.

C Mis-routing, per probe

Scored per probe — mis-routed iff the converter the probe’s family requires is never called on it — there are 7 mis-routed probe instances across all twelve routing cells: deepseek-chat 3 (AVAILABLE) + 3 (FORCED) and gpt-4o-mini 1 (FORCED); all other cells are zero, and every instance is `units_half`. Per-condition rates are 0–3.1%. Counting *calls* rather than probes inflates this substantially, because models legitimately chain the numeral parser before the unit converter; deepseek-chat makes 16 (AVAILABLE) and 23 (FORCED) such chained extra calls alongside the correct converter.

D Paraphrase diversity

Within-triplet dissimilarity between the three phrasings of each probe, character level, over the 36 variant groups of the focused set: mean normalised edit distance 0.556, mean Jaccard distance 0.611. By family, edit distance is 0.639 (`units_half`), 0.617 (control), 0.616 (lunar), 0.584 (elliptical), 0.509 (mixed_digit), 0.432 (internal_zero), 0.385 (capital), 0.310 (long_chain). The lunar family, where the consistency-detection blind spot concentrates, is more diverse than the non-lunar mean (0.616 vs 0.522) — so the blind spot is not an artifact of near-identical phrasings.

E Tokenisation

Tokens per character over the 108 focused probes: DeepSeek-V3 0.710, Qwen2.5-7B 0.746, GPT-4o (`o200k_base`) 0.920, Llama-3.1-8B 0.962. Chinese-optimised vocabularies are therefore $\approx 25\%$ more efficient here, as one would expect. That difference cannot explain our results. Within a single tokeniser, lunar probes and fixed-date-control probes fragment almost identically — lunar/control ratios 0.97 (GPT-4o), 1.06 (Qwen2.5), 0.98 (Llama-3.1) and 1.09 (DeepSeek-V3) — while their accuracies differ by up to 100pp; a quantity constant across the contrast cannot produce the contrast. Across models, qwen2.5:7b carries a tokeniser nearly as efficient as DeepSeek’s (0.746 vs 0.710) and still scores 3% on lunar against DeepSeek’s 81%. Fragmentation does not order the effect. (Gemini’s tokeniser is not public and is omitted.)