

To “Ba” or Not to “Ba”? LLM Surprisal as a Diagnostic of Context-Sensitive Language Understanding in L2 Chinese

Shengjie Zhang

International Cultural Exchange School
Fudan University
shengjiezhang22@m.fudan.edu.cn

Yan Cong

School of Languages and Cultures
Purdue University
cong4@purdue.edu

Abstract

Some difficulties in learning Chinese as a second language (L2 Chinese) involve choosing a sentence pattern that is grammatically well formed but less natural in context. Mandarin *ba*/SVO choice is one such case: learners may avoid *ba* when the discourse favors it or extend it to contexts where SVO is more appropriate. Because edit-based correction systems may leave both alternatives unchanged, diagnosing these choices requires sensitivity to discourse context, making the task a targeted test of context-sensitive natural language understanding (NLU). Using 200 controlled discourse contexts, we examine whether autoregressive LLM surprisal can serve as a graded diagnostic of this distinction. Across 21 open-weight models, sentence-level mean surprisal identifies the contextually preferred continuation with a mean model-level accuracy of 87.93%, whereas a MacBERT-based correction baseline misses most contextually non-preferred continuations. Relevant context generally lowers surprisal for the intended continuation, and sentence-level scoring is more reliable than first-token scoring, whose interpretation varies with tokenizer granularity. The results suggest that surprisal can complement edit-based correction in L2 pragmatic assessment, while cross-model asymmetries, structural preference biases, and tokenizer effects underscore the need for construction-specific controls and human interpretation when evaluating context-sensitive NLU.¹

1 Introduction

Some difficulties in L2 Chinese concern not whether a sentence is grammatical, but whether a particular sentence pattern is natural in its discourse context. The Mandarin *ba* (把) construction provides a clear example. It places an af-

fected object before the verb and is associated with disposal, affectedness, and event boundedness. L2 learners may avoid *ba* when the discourse favors it, or extend it to contexts where a subject-verb-object (SVO) continuation is more natural (Schachter, 1974; Odlin, 1989; Xing, 2006). In both cases, the resulting continuation may be syntactically well formed while remaining contextually non-preferred.

This distinction is difficult for conventional edit-based Chinese Grammatical Error Correction (CGEC). Such systems are primarily designed to identify overt form problems, including incorrect characters, missing or redundant words, and sentence-internal grammatical errors. When both BA and SVO alternatives are well formed, an editor-accept system may leave both unchanged even though one fits the preceding discourse less well. A complementary diagnostic is therefore needed for cases in which the relevant problem is contextual preference rather than grammatical well-formedness.

Before LLM surprisal can be used for this purpose, it is necessary to establish that model probabilities are sensitive to the relevant discourse distinction. This makes L2 pragmatic diagnosis a targeted test of context-sensitive natural language understanding (NLU). The model must distinguish between two locally well-formed continuations on the basis of the preceding context. Benchmarks such as BLiMP and CLiMP have shown how model probabilities can be used to evaluate controlled linguistic contrasts (Warstadt et al., 2020; Xiang et al., 2021), but their contrasts primarily distinguish acceptable from unacceptable sentences. The present task instead asks whether a model prefers the more natural member of a context-dependent BA/SVO contrast.

We operationalize this sensitivity using autoregressive surprisal, the negative log probability of a continuation, $S = -\ln P$. Lower surprisal in-

¹To facilitate reproducibility, our dataset and code are available at: <https://github.com/HHHHoly1219/llm-l2-ba-surprisal>

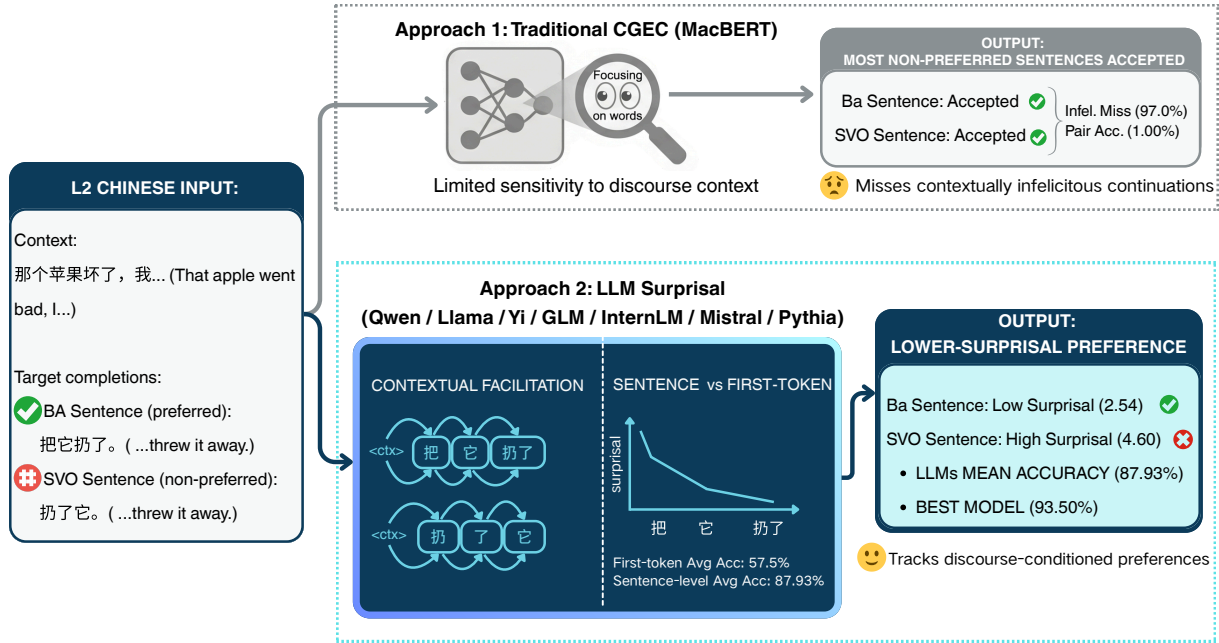


Figure 1: Overview of the evaluation paradigm. The MacBERT baseline is evaluated as an edit-based correction system, while LLM surprisal is evaluated through controlled BA/SVO contrasts. The comparison tests whether a model can prefer the continuation that is more natural in context.

dictates that the model finds a continuation more expected in the supplied context (Hale, 2001; Levy, 2008; Goodkind and Bicknell, 2018). We conduct three linked analyses. **First**, we test whether sentence-level surprisal favors the contextually preferred BA or SVO continuation and compare this graded diagnostic with a MacBERT-based correction baseline. **Second**, we compare the surprisal of the intended continuation with and without the preceding context to measure contextual facilitation. **Third**, we compare sentence-level and first-token surprisal to determine which scoring granularity provides the more reliable diagnostic (see Figure 1).

This study makes three contributions:

- **A diagnostic formulation of L2 pragmatic choice.** We show why contextually non-preferred but syntactically well-formed BA/SVO choices require a graded diagnostic that complements edit-based correction.
- **A targeted evaluation of context-sensitive NLU.** We introduce controlled BA/SVO contrasts that test whether LLM probabilities align with human-reviewed contextual preferences relevant to avoidance-like and overgeneralization-like choices in L2 Chinese.

- **A thorough empirical analysis of diagnostic reliability.** Across 21 autoregressive models, we examine contextual preference, contextual facilitation, and scoring granularity. The results show that observed context-sensitive NLU varies across models and structural preferences, while its measurement is also sensitive to tokenization and scoring choices.

2 Related Work

2.1 L2 Acquisition of the “Ba” Construction

The *ba* (把) construction is a well-studied but pedagogically difficult construction in Mandarin. Classic descriptions analyze *ba* as a construction associated with disposal, affectedness, definiteness of the object, and a bounded event outcome (Chao, 1968; Li and Thompson, 1981; Liu, 1997; Sybesma, 1999). These properties make *ba* especially natural in contexts where an action brings about a change of state, physical displacement, or resultative/directional completion. They also make the construction difficult to characterize as a syntactic alternation with the SVO alternative. The choice between the *ba* construction and an SVO sentence depends on how the event is packaged in discourse.

For L2 learners, the difficulty is therefore not only the surface *ba* frame (*ba*–Object–Verb–

Complement), but also the semantic and discourse conditions that license it. Learners may avoid a difficult target construction and turn to a less marked alternative (Schachter, 1974; Odlin, 1989); for *ba*, this corresponds to using SVO even when the context favors a *ba* construction (Xing, 2006). Conversely, learners may overgeneralize an acquired form after intensive training (Selinker, 1972; Ellis, 1994), extending *ba* to contexts that do not satisfy its usual semantic or discourse conditions. In the present study, BA-preferred contexts model avoidance-like cases, while SVO-preferred contexts model overgeneralization-like cases. Since both alternatives can be syntactically well formed, the relevant contrast is about pragmatic naturalness in context.

2.2 From Form Correction to Context-Sensitive Assessment

Chinese Grammatical Error Correction (CGEC) is commonly formulated as an edit-based task in which a system either rewrites an erroneous learner sentence or leaves it unchanged. Resources such as MuCGEC support this formulation by providing multiple reference corrections for Chinese learner texts (Zhang et al., 2022). This approach is well suited to overt form problems, including incorrect characters, missing or redundant words, and sentence-internal grammatical errors.

Contextually non-preferred BA/SVO choices pose a different problem. Both alternatives may be syntactically well formed, while differing in how naturally they fit the preceding discourse. An SVO continuation may be less natural when the context highlights disposal, affectedness, or a bounded result. Conversely, a locally well-formed *ba* continuation may be less appropriate in a habitual or unbounded event description. Because neither alternative necessarily contains an overt form error, an edit-or-accept system may leave both unchanged.

We therefore use a MacBERT-based correction model as an edit-based baseline (Cui et al., 2020). It illustrates what an edit-based interface explicitly detects when the relevant distinction concerns discourse fit rather than sentence-internal well-formedness. This motivates a complementary diagnostic that represents contextual preference as a graded signal.

2.3 Surprisal as a Diagnostic of Context-Sensitive NLU

Surprisal quantifies how strongly a language model expects a continuation. It is defined as the negative log probability of a token sequence given its preceding context, so lower surprisal indicates greater model expectation. In psycholinguistics, surprisal has been used to relate probabilistic language expectations to human processing difficulty (Hale, 2001; Levy, 2008). Surprisal estimates from stronger language models have also been found to better predict aspects of reading behavior (Goodkind and Bicknell, 2018; Wilcox et al., 2020).

Surprisal does not by itself demonstrate language understanding. However, it can provide an observable measure of context sensitivity. If a model uses the preceding discourse to distinguish between BA and SVO, it should assign lower surprisal to the continuation judged more natural in that context. This makes surprisal relevant to L2 assessment: a well-formed learner production may still receive a weaker probability-based preference because it does not fit the discourse. Recent work has similarly used model-derived surprisal to examine L2 development and sensitivity to form–meaning mappings (Cong, 2025a,b).

The present study tests whether this probability-based signal is suitable for diagnosing discourse-conditioned BA/SVO choice. We ask whether models systematically favor the human-reviewed preferred continuation, whether adding the relevant context facilitates that continuation, and whether the resulting signal remains reliable across sentence-level (the mean negative log probability over all scored tokens in the target) and first-token (the negative log probability of the first scored token of the continuation) scoring. These analyses separate three issues: whether the model distinguishes the contextual preference, whether that preference is sensitive to discourse information, and whether it can be measured consistently across models and tokenizers.

3 Experimental Design

3.1 Overall Hypotheses

The experimental design evaluates whether LLM surprisal can provide a context-sensitive diagnostic for BA/SVO choices relevant to L2 Chinese assessment. **H1** (Hypothesis 1) concerns diagnostic discrimination: sentence-level surprisal should be lower for the contextually preferred continua-

tion, whereas the CGEC baseline may leave both syntactically well-formed alternatives unchanged. **H2** concerns contextual facilitation: adding the relevant discourse context should reduce the surprisal of the intended continuation. **H3** concerns diagnostic reliability across scoring granularities: sentence-level surprisal over the full continuation should provide a more stable preference signal than first-token surprisal at the initial BA/SVO choice point (*ba* marker in BA continuations or the initial verb in SVO continuations).

3.2 Dataset

We constructed a controlled evaluation dataset with 800 scored target continuations. The dataset contains 200 discourse contexts, each paired with a *ba* continuation and an SVO continuation. This yields 400 contextualized target continuations. We also created a contextless counterpart for each continuation by removing the preceding discourse context, yielding another 400 scored continuation instances.

The contextualized subset is divided into two types. **BA-preferred contexts** include 100 contexts that favor the *ba* continuation: 50 involve physical state change and 50 involve spatial displacement. In these items, the *ba* continuation is contextually preferred and the SVO continuation is non-preferred. **SVO-preferred contexts** include 100 contexts where SVO is more natural: 50 involve habitual actions and 50 involve narrative or routine event descriptions. In these items, the SVO continuation is contextually preferred and the *ba* continuation is non-preferred. Representative examples are shown in Table 1. Detailed pinyin, glosses, and translations are provided in Appendix A.

The contextless condition is used for H2. It allows us to compare the same continuation with and without the licensing context, and therefore to test whether surprisal changes when discourse information is supplied.

All items were first screened with a T5 Chinese correction model (Raffel et al., 2020; Xu, 2022) as an auxiliary check. T5 left 78.0% of sentences unchanged. Because many suggested edits were unnecessary surface changes, final validation was conducted by 6 graduate student volunteers in Teaching Chinese to Speakers of Other Languages (TC-SOL). The review checked the intended pragmatic category, the syntactic well-formedness of both continuations, and the absence of typographical errors.

3.3 Model Selection

We evaluate 21 autoregressive LLMs and one MacBERT-based correction baseline. The LLM set covers Chinese-oriented and English-centric model families, with parameter sizes ranging from 410M to 14B.

Correction baseline. We use a MacBERT-based correction model as an edit-based CGEC baseline (Cui et al., 2020). This baseline is not intended as a discourse-level pragmatic evaluator. It is included to test what an edit-or-accept correction interface detects or misses when both BA and SVO alternatives are syntactically well formed.

Autoregressive LLMs. The autoregressive LLMs are among the most powerful models as of writing, including Chinese-oriented model families such as Qwen2.5 (0.5B, 1.5B, 3B, 7B, 14B) (Qwen Team, 2024), Qwen3 (0.6B, 1.7B, 4B, 8B, 14B) (Qwen Team, 2025), InternLM2.5-7B-Base (Cai et al., 2024; InternLM Team, 2024), GLM-4-9B-Base (THUDM, 2024), and Yi-1.5 (6B, 9B) (Young et al., 2024; 01.AI, 2024). For comparison, we also include widely used English-dominant or multilingual model families: Llama-2-7B-Base (Touvron et al., 2023), Llama-3-8B-Base and Llama-3.1-8B-Instruct (Dubey et al., 2024), Mistral-7B-v0.3-Base (Jiang et al., 2023; Mistral AI, 2024), and Pythia (410M, 1.4B, 2.8B) (Biderman et al., 2023).

3.4 Surprisal Metrics

For each item, the model receives a prefix and a target continuation. In the contextualized condition, the prefix is the discourse context C ; in the contextless condition, the prefix is empty $\emptyset$. Surprisal is computed only over the target continuation tokens, not over the context itself. Thus, in the contextualized condition we estimate $S(T | C)$, where T is either the BA or SVO continuation. Since the implementation uses natural-log cross-entropy, all surprisal values are reported in nats.

Sentence-level mean surprisal (S_{sentence}) is the average negative log probability over all scored tokens in the target continuation. This metric is used for H1 and H2:

$$S_{\text{sentence}}(T | C) = \frac{1}{N} \sum_{i=1}^N -\ln P(w_i | C, w_{<i}). \quad (1)$$

We use the mean surprisal so that longer continuations are not automatically penalized for containing more tokens.

Context Type	Category	Context	Preferred Completion	Non-preferred Completion
BA-preferred	Physical change	state 那个苹果坏了, 我.....(The apple has gone bad, I...)	✓ Ba: 把它扔了。(threw it away.)	# SVO: 扔了它。(threw it away.)
BA-preferred	Spatial displacement	抽屉里的那份合同很重要, 我.....(The contract in the drawer was very important, I...)	✓ Ba: 把它拿出来。(took it out.)	# SVO: 拿出它来了。(took it out.)
SVO-preferred	Habitual actions	弟弟非常喜欢运动, 每天下午都去操场.....(My younger brother loves sports; he goes to the playground every afternoon to...)	✓ SVO: 踢足球。(play soccer.)	# Ba: 把足球踢了。(play soccer.)
SVO-preferred	Narrative / schedule routine	昨晚停电了, 我在黑暗中坐了一会儿, 然后去厨房.....(The power went out last night. I sat in the dark for a while, then went to the kitchen and...)	✓ SVO: 点了那根蜡烛。(lit that candle.)	# Ba: 把那根蜡烛点了。(lit that candle.)

Table 1: Representative controlled BA/SVO contrasts used in the evaluation. ✓ indicates the contextually preferred continuation; # indicates a syntactically well-formed but pragmatically infelicitous alternative. Literal translations provided.

First-token surprisal (S_{first}) is the negative log probability of the first scored model token in the target continuation. This metric is used for H3:

$$S_{\text{first}}(T | C) = -\ln P(w_1 | C). \quad (2)$$

For BA continuations, this token usually corresponds to the *ba* marker; for SVO continuations, it usually corresponds to the first main verb or the first token of the SVO continuation. All contextualized stimuli were constructed so that this choice point is preceded by the discourse context, ensuring that first-token surprisal is computed conditionally rather than as an unconditioned sentence-initial score. The underlined portions in Table 1 illustrate these positions, such as 把 in a *ba* continuation or the first verb in an SVO continuation.

3.5 Specific Analyses by Hypothesis

3.5.1 H1: Correction Baseline and LLM Surprisal

To address H1, we test whether an edit-based correction interface can diagnose context-conditioned BA/SVO choices, and whether sentence-level LLM surprisal can select the contextually preferred continuation. For the MacBERT baseline, we evaluate 400 full contextualized sentences, corresponding to 200 minimal pairs. Each input is the preceding context concatenated with one target continuation. We treat any non-empty correction output as an

error flag and an unchanged output as acceptance, and compute the miss rate for non-preferred continuations, the acceptance rate for preferred continuations, pair-level edit/accept accuracy, and strict construction-level correction accuracy.

For the 21 autoregressive LLMs, we compute sentence-level mean surprisal for the BA and SVO continuations in each contextualized minimal pair. A model is counted as correct if it assigns lower surprisal to the contextually preferred continuation. In a BA-preferred context, this means:

$$S_{\text{sentence}}(\text{BA} | C) < S_{\text{sentence}}(\text{SVO} | C). \quad (3)$$

In an SVO-preferred context, the inequality is reversed. We report BA-preferred accuracy, SVO-preferred accuracy, and overall accuracy for each model.

3.5.2 H2: Context and No-context Comparison

H2 tests whether the preceding discourse context reduces the surprisal of the intended continuation. For each contextually preferred continuation T , we compute sentence-level surprisal reduction as

$$\Delta S = S_{\text{sent}}(T | \emptyset) - S_{\text{sent}}(T | C). \quad (4)$$

In BA-preferred contexts, T is the *ba* continuation; in SVO-preferred contexts, T is the SVO continuation. A positive ΔS indicates that the intended

continuation becomes more predictable when the relevant context is supplied. We interpret this quantity as contextual facilitation.

As a stricter check, we also compute SVO-to-BA preference reversals in BA-preferred contexts. A reversal occurs when a model assigns lower surprisal to the SVO continuation without context, but lower surprisal to the *ba* continuation after the BA-licensing context is supplied.

3.5.3 H3: Sentence-level and First-token Surprisal

H3 asks whether the contextual naturalness of BA/SVO continuations is better captured by sentence-level mean surprisal over the full target continuation or by first-token surprisal at the initial choice point. Prior work on language model evaluation and L2 surprisal has shown that local target-word or critical-word surprisal can be informative for syntactic and developmental analysis (Marvin and Linzen, 2018; Hu et al., 2020; Cong, 2025b). This makes the initial BA/SVO choice point a plausible diagnostic candidate.

However, we hypothesize that sentence-level surprisal will be more reliable for the present phenomenon. The felicity of *ba* is not determined by the marker alone, but by the larger constructional sequence involving the affected object, verb, and resultative, directional, or otherwise bounded predicate. In addition, first-token scores may be sensitive to Chinese tokenization. We therefore evaluate both S_{sentence} and S_{first} in the contextualized condition using the same minimal-pair decision rule as in H1: a model is correct when the contextually preferred continuation receives lower surprisal. We then compare model-level accuracy under the two granularities and report a tokenizer diagnostic in Appendix D.

4 Results

4.1 Correction Baseline and LLM Surprisal for Pragmatic Discrimination

H1 compares an edit-based CGEC baseline with sentence-level LLM surprisal on context-conditioned BA/SVO choice. Table 2 summarizes the MacBERT baseline results.

MacBERT missed 97.0% of the pragmatically infelicitous continuations and produced the correct pair-level edit/accept pattern in only 1.0% of cases. None of its edits constituted a construction-level BA/SVO correction. This pattern reflects a

Scope	Infel. Miss	Fel. Accept	Pair Acc.	Const. Corr.
All contexts	97.0%	97.5%	1.0%	0.0%
BA-preferred	99.0%	99.0%	1.0%	0.0%
SVO-preferred	95.0%	96.0%	1.0%	0.0%

Table 2: MacBERT baseline performance. Infel. Miss = infelicitous continuations accepted without edits; Fel. Accept = felicitous continuations accepted without edits; Pair Acc. = preferred accepted and non-preferred flagged within the same minimal pair; Const. Corr. = non-preferred continuation rewritten into the paired preferred BA/SVO alternative.

mismatch: both alternatives are syntactically well formed, and the infelicity lies in the relation between the continuation and the preceding discourse, not in the sentence string alone.

By contrast, sentence-level LLM surprisal provides a usable preference signal. Across the 21 autoregressive LLMs, the mean model-level accuracy was 87.93%, with higher accuracy in BA-preferred than in SVO-preferred contexts (92.05% vs. 83.81%). Qwen3-14B-Base achieved the highest overall accuracy (93.50%), while Pythia-410M-Base was lowest (77.50%). The full ranking and paired surprisal-differential plots are provided in Appendix B.1 and Appendix B.2.

The model-level pattern is not reducible to parameter scale. As shown in Appendix B, Qwen2.5-1.5B-Base and Qwen2.5-3B-Base both reach 90.50%, matching Qwen2.5-14B-Base, and Qwen3-0.6B-Base reaches 87.00%. A more informative pattern is the BA-SVO balance gap: Pythia-410M-Base and Llama-2-7B-Base reach 100.00% in BA-preferred contexts but only 55.00% and 65.00% in SVO-preferred contexts, while Pythia-1.4B-Base and Pythia-2.8B-Base show similar gaps (97.00%/71.00% and 96.00%/70.00%). These asymmetries suggest that high BA-preferred accuracy can reflect a bias toward the marked *ba* continuation rather than balanced sensitivity to both avoidance and overgeneralization, classic cases in L2 development.

The contrast between Qwen (Chinese-oriented) and the Llama-2/Pythia models (English-centric or general-purpose) suggests that stronger performance depends on balanced calibration of BA and SVO preferences, not simply on assigning low surprisal to well-formed *ba* sentences. This points to the need for construction-sensitive evaluation that tests both the licensing and overextension of marked constructions.

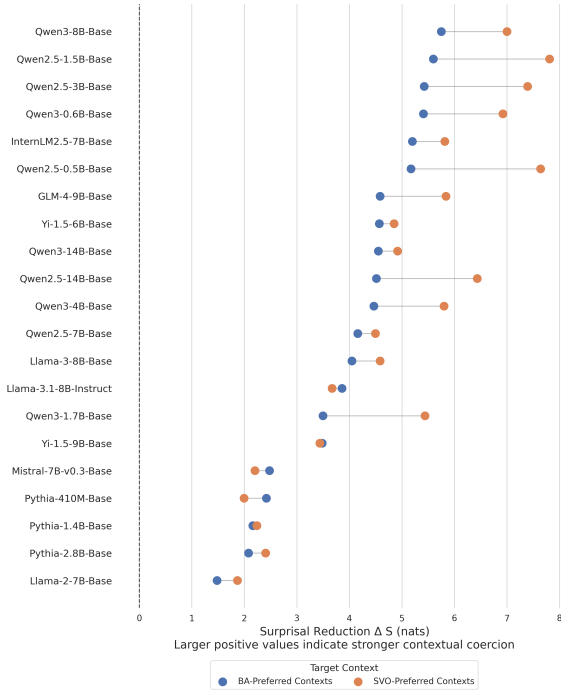


Figure 2: Mean surprisal reduction after adding context, computed as $\Delta S = S_{\text{sent}}(T \mid \emptyset) - S_{\text{sent}}(T \mid C)$.

4.2 Contextual Facilitation

To address H2, we ask whether adding the relevant discourse context makes the licensed continuation less surprising. Figure 2 shows positive mean surprisal reductions across models, indicating that the supplied context generally facilitates the intended continuation. The size of this effect varies by model and by context type. Qwen3-8B-Base shows large reductions in both BA-preferred and SVO-preferred contexts, suggesting relatively balanced contextual facilitation. Qwen2.5-0.5B-Base shows an especially large reduction in SVO-preferred contexts, but its H1 accuracy remains less balanced than the strongest Qwen models, indicating that a large ΔS alone does not guarantee balanced pragmatic discrimination. By contrast, Llama-2-7B-Base and the Pythia models show smaller reductions and low reversal rates, suggesting weaker context-driven reweighting or a stronger context-independent structural preference.

As a stricter check, we also computed preference reversals in BA-preferred contexts: cases where a model preferred SVO without context but preferred the *ba* continuation after the BA-licensing context was supplied. The highest reversal rates occur for Yi-1.5-6B-Base and Qwen2.5-1.5B-Base, while Llama-2-7B-Base, Llama-3-8B-Base, and several Pythia models show near-zero reversal rates.

Low reversal should be interpreted cautiously: it can reflect weak contextual coercion, but it can also occur when a model already prefers *ba* before context is added. Model preferences between the BA and SVO continuations without context are reported in Appendix C.1. The full reversal-rate plot is provided in Appendix C.2.

4.3 Granularity of Pragmatic Evaluation

The results support H3: sentence-level surprisal is a more reliable metric than first-token surprisal for BA/SVO pragmatic discrimination. Across the 21 autoregressive LLMs, mean sentence-level accuracy was 87.93%, compared with 57.50% for first-token accuracy. A two-sided paired *t*-test across models confirmed the difference ($p = 1.05 \times 10^{-4}$). As shown in Figure 3, the models fall below the diagonal, indicating consistently higher accuracy for sentence-level evaluation. Several models illustrate this pattern. Qwen3-14B-Base achieved the highest sentence-level accuracy (93.50%), followed by Llama-3-8B-Base (92.00%) and GLM-4-9B-Base and Mistral-7B-v0.3-Base (both 91.50%). These high sentence-level scores do not translate into equally high first-token accuracy. For example, Qwen3-14B-Base reaches 75.50% first-token accuracy, while Mistral-7B-v0.3-Base falls to 0.00%. The strongest first-token results are Qwen2.5-3B-Base (78.50%), Qwen2.5-14B-Base (78.00%), Llama-3-8B-Base (77.50%), and Qwen3-0.6B-Base (76.00%), but these still remain below their corresponding sentence-level accuracies. The full model-level breakdown is provided in Appendix D.

This gap is linguistically informative. Although local target-word surprisal can be useful in other settings, the Mandarin *ba* construction is not licensed by the marker *ba* alone. Its felicity depends on a larger constructional configuration: *ba* plus a definite affected object, a transitive verb, and typically a resultative, directional, or otherwise bounded predicate. First-token surprisal only measures the probability of the initial structural choice, before the model observes the *ba*-object-verb-complement sequence that determines whether the event is compatible with a *ba* construal. Sentence-level surprisal is therefore better aligned with the linguistic unit being evaluated.

The first-token metric is also less stable across tokenizers. Four models (Mistral-7B-v0.3-Base, Yi-1.5-6B-Base, Yi-1.5-9B-Base, and Llama-2-7B-Base) show 0.00% first-token accuracy despite

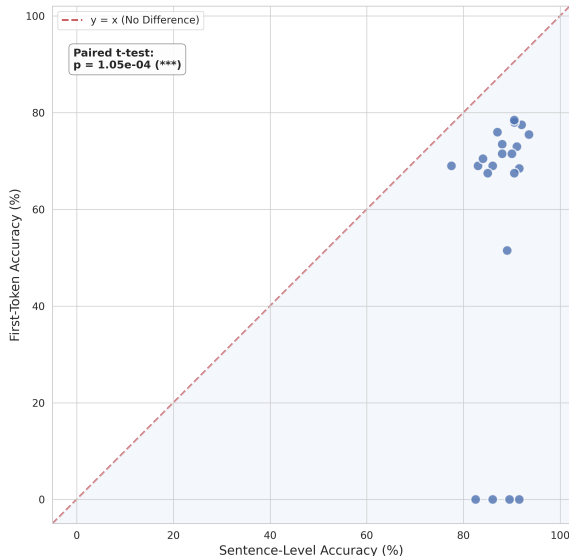


Figure 3: Granularity comparison. The x-axis is sentence-level accuracy and the y-axis is first-token accuracy. Points below the dashed diagonal indicate higher accuracy for sentence-level surprisal.

much higher sentence-level accuracy. A tokenizer diagnostic over all 400 contextualized continuations (Appendix D, Table 7) shows that the first model token is not always a linguistically meaningful Chinese character. In such cases, the measured “first token” is not necessarily the linguistic marker under analysis. This pattern is consistent with our expectation that first-token surprisal is vulnerable to tokenization artifacts in cross-model evaluation of Chinese constructions (Haslett, 2025). Sentence-level surprisal, by averaging over the full continuation, is less sensitive to this local segmentation problem and provides the more reliable diagnostic signal for the present context-sensitive assessment task.

5 Discussion and Implications

The findings provide qualified evidence that LLM surprisal captures the context-sensitive NLU required for diagnosing L2-relevant BA/SVO choices. Here, context-sensitive NLU is defined narrowly as ranking two locally well-formed continuations by their fit with the preceding discourse. The MacBERT-based baseline leaves most contextually non-preferred alternatives unchanged, whereas autoregressive surprisal generally favors the human-reviewed intended continuation.

However, this sensitivity remains uneven and measurement-dependent. Relevant context generally lowers surprisal for the intended continua-

tion, and preference reversals show that context can alter the relative BA/SVO preference. Yet surprisal reduction may also reflect greater lexical or event predictability, so it does not by itself establish construction-specific pragmatic understanding. Sentence-level scoring is more reliable than first-token scoring, suggesting that the relevant signal extends beyond the initial choice point. Cross-model results further show that parameter count alone does not explain performance. Chinese-oriented families, particularly Qwen, tend to show smaller BA–SVO gaps, whereas several Llama-2 and Pythia models display a stronger preference for *ba* across context types. Because training data, architecture, scale, instruction tuning, and tokenization vary together, the observed family-level differences cannot be attributed to any single model property.

For L2 Chinese pedagogy, this research is important because avoidance-like and overgeneralization-like cases require different feedback. A learner who uses SVO in a BA-favoring context may need help noticing affectedness, disposal, or bounded result. A learner who overuses *ba* may need help identifying contexts where SVO is the more natural event description. Surprisal can help locate where context-sensitive feedback is needed, but it cannot by itself provide an adequate pedagogical explanation. Practical use would require validation on authentic learner productions, human-informed decision criteria, and continued teacher oversight.

6 Conclusion

This study begins from an L2 Chinese assessment problem: BA/SVO choices may be grammatically well formed but less natural in a particular discourse context. We ask whether LLM surprisal is sufficiently sensitive to this distinction to serve as a complementary diagnostic. Across 21 autoregressive models, sentence-level surprisal generally favors the intended continuation, while a MacBERT-based correction baseline leaves most contextually non-preferred alternatives unchanged. Contextual facilitation and scoring-granularity analyses further show that the diagnostic depends on both discourse information and the linguistic region over which surprisal is measured.

These findings suggest that context-sensitive NLU can support L2 pragmatic diagnosis, but they do not establish human-like pragmatic understanding or a ready-to-use correction system. Cross-

model asymmetries, BA preference biases, and tokenizer effects require construction-specific controls and human interpretation. Surprisal is therefore best treated as a graded supplement to edit-based correction and teacher judgment. More broadly, this study illustrates how probabilities from LLMs can be used to investigate language behavior beyond text generation and automatic classification.

Limitations

Controlled rather than authentic learner data

The dataset consists of controlled BA/SVO contrasts rather than authentic learner productions. The contexts are short and single-turn. This keeps the conditions clear, but it does not test longer discourse dependencies. This design isolates contextual preference, but real learner texts contain interacting lexical, orthographic, morphosyntactic, and discourse-level problems. The present findings therefore establish diagnostic potential under controlled conditions, not performance on naturally occurring learner errors.

Human validation of contextual preference

The intended preferences were established through expert review rather than a formal acceptability experiment. They should therefore be interpreted as designed contextual preferences rather than population-level estimates of native-speaker judgment. Future work should obtain blinded ratings from multiple native speakers and examine whether surprisal differences correlate with graded human judgments.

Interpretation of contextual facilitation H2 defines contextual facilitation as the difference between the surprisal of the same intended continuation without and with its preceding context. A positive value shows that the continuation becomes more predictable when context is supplied, but the reduction may reflect both construction-related information and more general lexical or event predictability. The BA and SVO continuations were not uniformly matched for lexical material, aspectual marking, or boundedness, and cues such as *le* (了) may therefore contribute to the observed alignment asymmetry. Preference reversals provide a stricter, although narrower, indication that context changes relative BA/SVO preference.

Pedagogical validation The study does not test whether surprisal-based diagnostics improve teacher decisions, learner feedback, or learning

outcomes. Practical application would require validation on authentic learner data, criteria for distinguishing strong preferences from acceptable variation, and an explanation component that converts model probabilities into linguistically and pedagogically appropriate feedback.

Ethics Statement

This study does not involve the collection or processing of any sensitive personal data, personally identifiable information (PII), or real-world user text. The evaluation dataset comprising 800 scored target continuations was constructed for this study and manually verified by graduate student volunteers in Teaching Chinese to Speakers of Other Languages (TCSOL) for the sole purpose of controlled computational evaluation. The primary intended use of this surprisal-based diagnostic methodology is to serve as a supplementary, formative tool to assist language educators and L2 learners in understanding implicit pragmatic constraints. Human oversight remains necessary when interpreting probabilistic model outputs for pedagogical use.

Acknowledgments

We acknowledge LUHME (The 3rd Workshop on Language Understanding in the Human-Machine Era) reviewers for their constructive comments. We thank Lena Rotteveel, the Computation and Linguistic Meaning Lab, Fudan University for the helpful discussion.

References

- 01.AI. 2024. [Yi-1.5 model card](#). Hugging Face model card.
- Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, and 1 others. 2023. [Pythia: A suite for analyzing large language models across training and scaling](#). *Preprint*, arXiv:2304.01373.
- Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, and 1 others. 2024. [Internlm2 technical report](#). *Preprint*, arXiv:2403.17297.
- Yuen Ren Chao. 1968. *A Grammar of Spoken Chinese*. University of California Press, Berkeley.
- Yan Cong. 2025a. [Demystifying large language models in second language development research](#). *Computer Speech & Language*, 89:101700.

- Yan Cong. 2025b. [Second language learning of degree expressions: A computational approach](#). *Natural Language Processing*, 31(5):1187–1209.
- Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. [Revisiting pre-trained models for Chinese natural language processing](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 657–668, Online. Association for Computational Linguistics.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, and 11 others. 2024. [The Llama 3 herd of models](#). *Preprint*, arXiv:2407.21783.
- Rod Ellis. 1994. *The Study of Second Language Acquisition*. Oxford University Press, Oxford.
- Adam Goodkind and Klinton Bicknell. 2018. [Predictive power of word surprisal for reading times is a linear function of language model quality](#). In *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2018)*, pages 10–18, Salt Lake City, Utah. Association for Computational Linguistics.
- John Hale. 2001. [A probabilistic Earley parser as a psycholinguistic model](#). In *Second Meeting of the North American Chapter of the Association for Computational Linguistics*.
- David A. Haslett. 2025. [Tokenization changes meaning in large language models: Evidence from Chinese](#). *Computational Linguistics*, 51(3):785–814.
- Jennifer Hu, Jon Gauthier, Peng Qian, Ethan Wilcox, and Roger Levy. 2020. [A systematic assessment of syntactic generalization in neural language models](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1725–1744, Online. Association for Computational Linguistics.
- InternLM Team. 2024. [Internlm2.5-7b model card](#). Hugging Face model card.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, and 1 others. 2023. [Mistral 7b](#). *Preprint*, arXiv:2310.06825.
- Roger Levy. 2008. [Expectation-based syntactic comprehension](#). *Cognition*, 106(3):1126–1177.
- Charles N. Li and Sandra A. Thompson. 1981. *Mandarin Chinese: A Functional Reference Grammar*. University of California Press, Berkeley.
- Feng-hsi Liu. 1997. [An Aspectual Analysis of ba](#). *Journal of East Asian Linguistics*, 6(1):51–99.
- Rebecca Marvin and Tal Linzen. 2018. [Targeted syntactic evaluation of language models](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202, Brussels, Belgium. Association for Computational Linguistics.
- Mistral AI. 2024. [Mistral-7b-v0.3 model card](#). Hugging Face model card.
- Terence Odlin. 1989. *Language Transfer: Cross-Linguistic Influence in Language Learning*. Cambridge University Press, Cambridge.
- Qwen Team. 2024. [Qwen2.5 technical report](#). *Preprint*, arXiv:2412.15115.
- Qwen Team. 2025. [Qwen3 technical report](#). *Preprint*, arXiv:2505.09388.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. [Exploring the limits of transfer learning with a unified text-to-text transformer](#). *Journal of Machine Learning Research*, 21(140):1–67.
- Jacquelyn Schachter. 1974. [An error in error analysis](#). *Language Learning*, 24(2):205–214.
- Larry Selinker. 1972. [Interlanguage](#). *International Review of Applied Linguistics in Language Teaching*, 10(3):209–231.
- Rint Sybesma. 1999. *The Mandarin VP*. Kluwer Academic Publishers, Dordrecht.
- THUDM. 2024. [Glm-4-9b model card](#). Hugging Face model card.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruiti Bhosale, and 1 others. 2023. [Llama 2: Open foundation and fine-tuned chat models](#). *Preprint*, arXiv:2307.09288.
- Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. [BLiMP: The benchmark of linguistic minimal pairs for English](#). *Transactions of the Association for Computational Linguistics*, 8:377–392.
- Ethan Gotlieb Wilcox, Jon Gauthier, Jennifer Hu, Peng Qian, and Roger Levy. 2020. [On the predictive power of neural language models for human real-time comprehension behavior](#). *Preprint*, arXiv:2006.01912.
- Beilei Xiang, Changbing Yang, Yu Li, Alex Warstadt, and Katharina Kann. 2021. [CLiMP: A benchmark for Chinese language model evaluation](#). *arXiv preprint arXiv:2101.11131*.
- Janet Zhiqun Xing. 2006. *Teaching and Learning Chinese as a Foreign Language: A Pedagogical Grammar*. Hong Kong University Press, Hong Kong.

Ming Xu. 2022. [Mengzi-t5-base chinese correction model](#). Hugging Face model card.

Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Guoyin Wang, Heng Li, Jiangcheng Zhu, Jianqun Chen, and 1 others. 2024. [Yi: Open foundation models by 01.ai](#). *Preprint*, arXiv:2403.04652.

Yue Zhang, Zhenghua Li, Zuyi Bao, Jiacheng Li, Bo Zhang, Chen Li, Fei Huang, and Min Zhang. 2022. [MuCGEC: a multi-reference multi-source evaluation dataset for Chinese grammatical error correction](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3118–3130, Seattle, United States. Association for Computational Linguistics.

A Annotated Stimulus Examples

This appendix provides fully annotated examples of the contextual minimal pairs used in the evaluation. Table 3 and Table 4 present a compressed version with pinyin, glosses, or translations in Appendix A.

Category	Context	Target Completions
Physical State Change	那个苹果坏了，我..... (<i>That apple went bad, I...</i>)	BA Sentence ✓ 把 它 扔 了。 bǎ tā rēng le ba it throw PFV ‘...threw it away.’ SVO Sentence # 扔 了 它。 rēng le tā throw PFV it ‘...threw it away.’
Spatial Displacement	抽屉里的那份合同很重要，我..... (<i>The contract in the drawer was very important, I...</i>)	BA Sentence ✓ 把 它 拿 出 来 了。 bǎ tā ná chū lái le ba it take out come PFV ‘...took it out.’ SVO Sentence # 拿 出 它 来 了。 ná chū tā lái le take out it come PFV ‘...took it out.’

Table 3: Annotated examples of BA-preferred contexts. ✓ indicates the contextually felicitous continuation; # indicates the syntactically well-formed but pragmatically infelicitous alternative.

Category	Context	Target Completions
Habitual Actions	弟弟非常喜欢运动，每天下午都去操场..... (<i>My younger brother loves sports; he goes to the playground every afternoon to...</i>)	SVO Sentence ✓ 踢 足球。 tī zúqiú kick soccer ‘...play soccer.’ BA Sentence # 把 足球 踢 了。 bǎ zúqiú tī le ba soccer kick PFV ‘...play soccer.’
Narrative / Schedule Routine	昨晚停电了，我在黑暗中坐了一会儿，然后去厨房..... (<i>The power went out last night. I sat in the dark for a while, then went to the kitchen and...</i>)	SVO Sentence ✓ 点 了 那 根 蜡烛。 diǎn le nà gēn làzhú light PFV that CL candle ‘...lit that candle.’ BA Sentence # 把 那 根 蜡烛 点 了。 bǎ nà gēn làzhú diǎn le ba that CL candle light PFV ‘...lit that candle.’

Table 4: Annotated examples of SVO-preferred contexts. ✓ indicates the contextually felicitous continuation; # indicates the syntactically well-formed but pragmatically infelicitous alternative.

B Supplementary Results for H1

This appendix reports supplementary evidence for H1. It includes the full model-wise sentence-level accuracy results, the corresponding model ranking figure, and paired surprisal-differential checks. The main text reports only the aggregate H1 results.

B.1 Model-wise LLM Surprisal Accuracy

This subsection reports the model-wise H1 accuracy results for sentence-level LLM surprisal. Figure 4 visualizes the model ranking, while Table 5 provides the corresponding BA-preferred, SVO-preferred, overall accuracy, and balance-gap values.

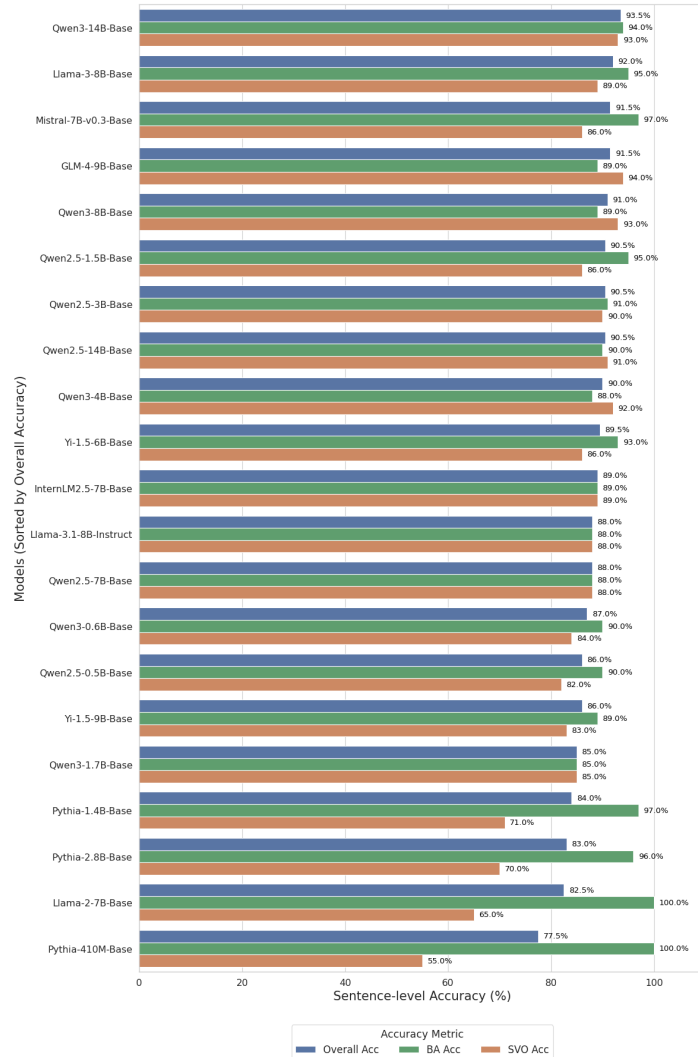


Figure 4: Model-wise sentence-level pragmatic discrimination accuracy. Blue bars show overall accuracy; green and orange bars show performance in BA-preferred and SVO-preferred contexts.

Model	BA	SVO	Overall	Gap
Qwen3-14B-Base	94.0	93.0	93.5	1.0
Llama-3-8B-Base	95.0	89.0	92.0	6.0
Mistral-7B-v0.3-Base	97.0	86.0	91.5	11.0
GLM-4-9B-Base	89.0	94.0	91.5	5.0
Qwen3-8B-Base	89.0	93.0	91.0	4.0
Qwen2.5-1.5B-Base	95.0	86.0	90.5	9.0
Qwen2.5-3B-Base	91.0	90.0	90.5	1.0
Qwen2.5-14B-Base	90.0	91.0	90.5	1.0
Qwen3-4B-Base	88.0	92.0	90.0	4.0
Yi-1.5-6B-Base	93.0	86.0	89.5	7.0
InternLM2.5-7B-Base	89.0	89.0	89.0	0.0
Llama-3.1-8B-Instruct	88.0	88.0	88.0	0.0
Qwen2.5-7B-Base	88.0	88.0	88.0	0.0
Qwen3-0.6B-Base	90.0	84.0	87.0	6.0
Qwen2.5-0.5B-Base	90.0	82.0	86.0	8.0
Yi-1.5-9B-Base	89.0	83.0	86.0	6.0
Qwen3-1.7B-Base	85.0	85.0	85.0	0.0
Pythia-1.4B-Base	97.0	71.0	84.0	26.0
Pythia-2.8B-Base	96.0	70.0	83.0	26.0
Llama-2-7B-Base	100.0	65.0	82.5	35.0
Pythia-410M-Base	100.0	55.0	77.5	45.0
Average	92.05	83.81	87.93	9.57

Table 5: Full H1 sentence-level surprisal accuracy by model. BA and SVO indicate accuracy in BA-preferred and SVO-preferred contexts, respectively. Gap is the absolute BA-SVO accuracy difference. Values are percentages.

B.2 Item-Wise Paired Surprisal Differentials

To examine the item-level evidence underlying the H1 accuracy results, we computed a paired surprisal differential for each contextual minimal pair i :

$$d_i = S_{\text{nonpreferred},i} - S_{\text{preferred},i},$$

where S denotes sentence-level mean surprisal in nats. Specifically, the differential was calculated as

$$d_i = \begin{cases} S_{\text{SVO},i} - S_{\text{BA},i}, & \text{for BA-preferred contexts,} \\ S_{\text{BA},i} - S_{\text{SVO},i}, & \text{for SVO-preferred contexts.} \end{cases}$$

Accordingly, $d_i > 0$ indicates that the model assigns lower surprisal to the contextually preferred continuation, whereas $d_i < 0$ indicates a preference reversal.

Figures 5 and 6 show the distributions of these item-wise differentials separately by model and contextual preference. Each point represents one contextual minimal pair, and the percentage annotation above each distribution reports the proportion of items for which $d_i > 0$. Positive differentials predominate across all model-condition combinations, providing descriptive item-level support for H1. However, the distributions also reveal variation in effect magnitude and consistency: the proportion of positive differentials ranges from 85% to 100% in BA-preferred contexts, compared with 55% to 94% in SVO-preferred contexts. The latter variation is particularly pronounced for Llama-2-7B and the smaller Pythia models. Beyond the binary accuracy measure reported in Section B.1, these distributions reveal the magnitude, dispersion, and negative tails of the item-wise effects.

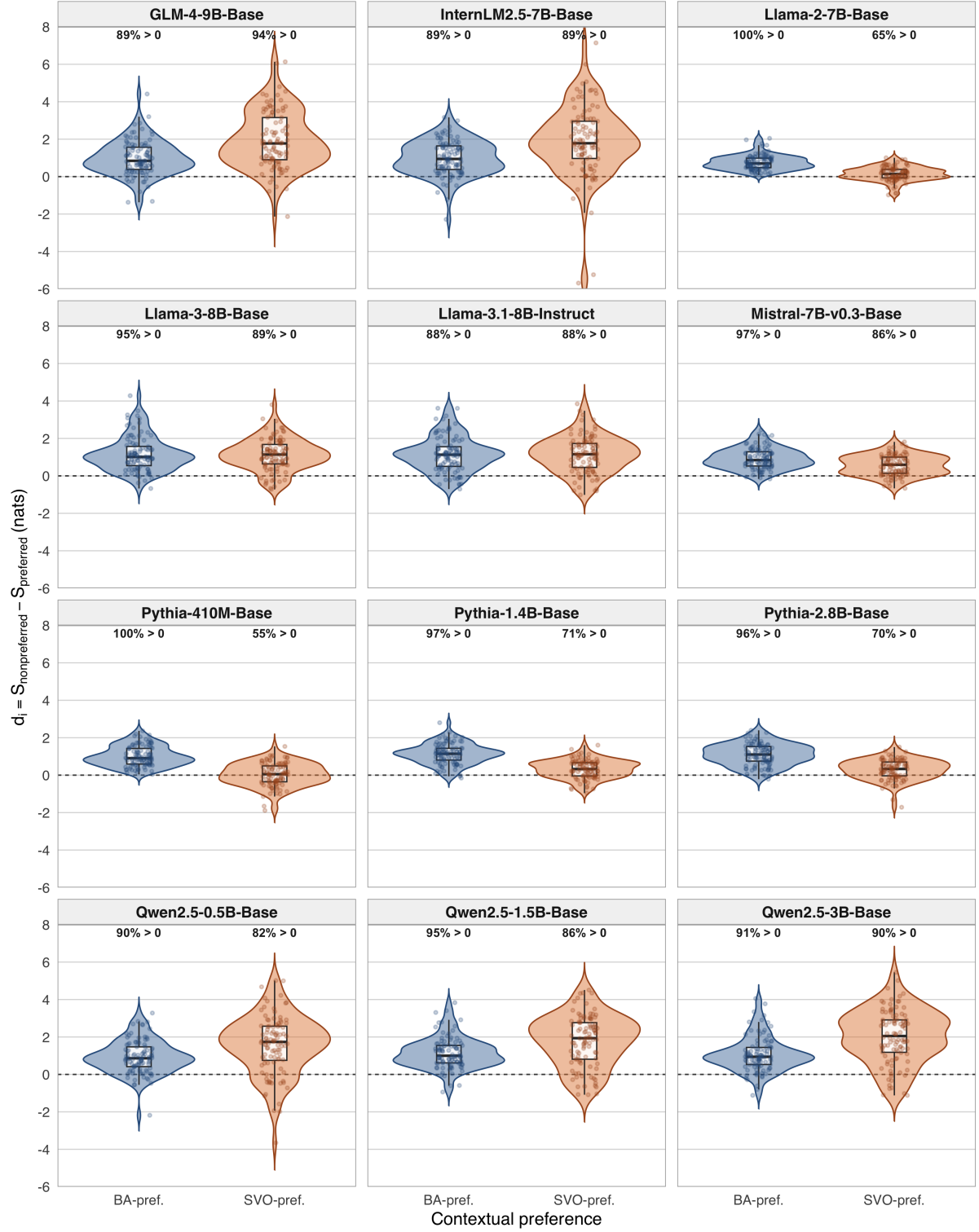


Figure 5: Item-wise paired surprisal differentials for H1 (Part 1 of 2). For each contextual minimal pair i , $d_i = S_{\text{nonpreferred},i} - S_{\text{preferred},i}$, measured in nats. Positive values indicate lower surprisal for the contextually preferred continuation. Points represent individual minimal pairs, violin widths show distributional density, and the embedded boxplots summarize the median and interquartile range. The dashed horizontal line marks $d_i = 0$, and the annotations report the percentage of items with $d_i > 0$. Blue and orange denote BA-preferred and SVO-preferred contexts, respectively.

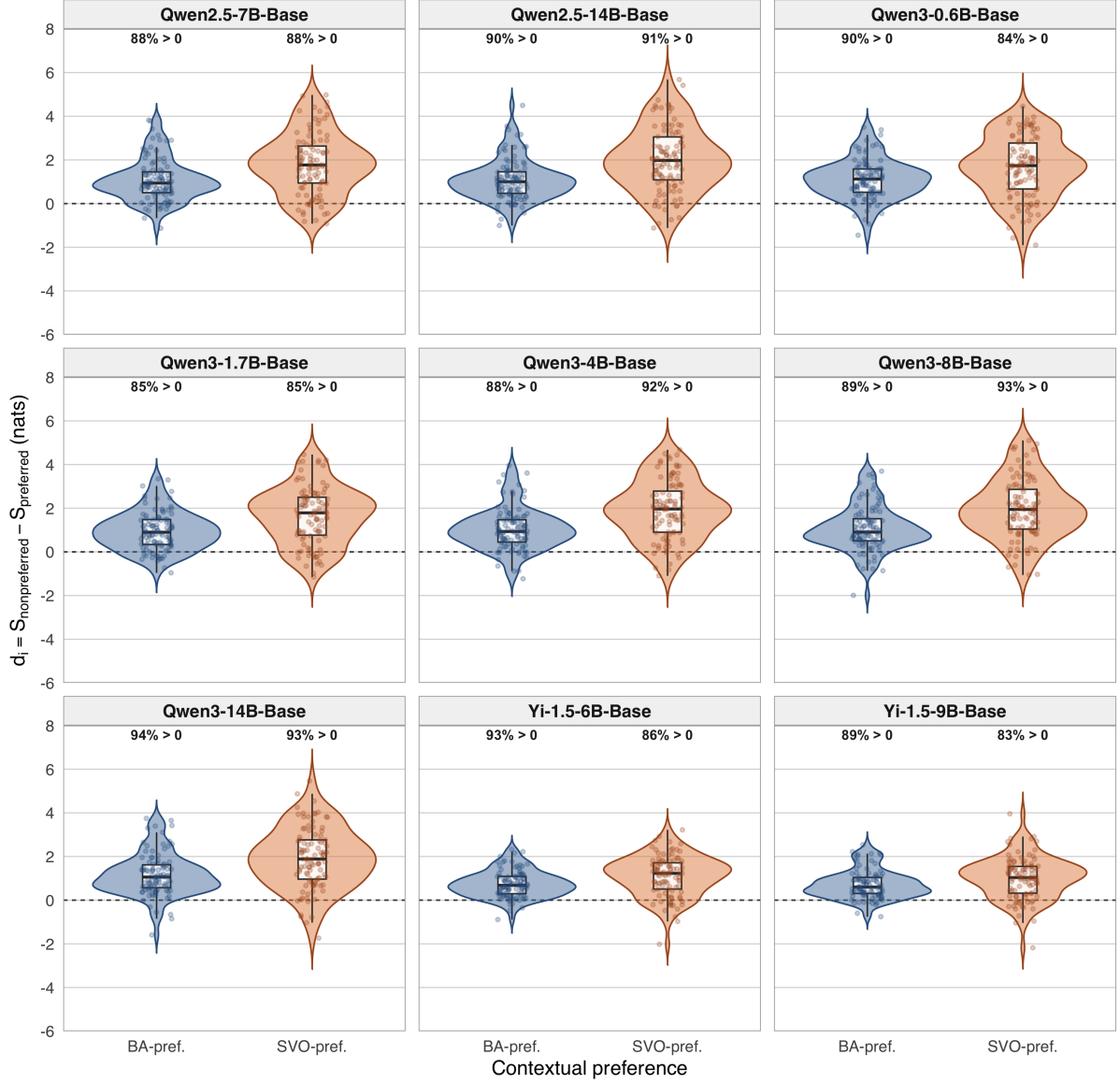


Figure 6: Item-wise paired surprisal differentials for H1 (Part 2 of 2). For each contextual minimal pair i , $d_i = S_{\text{nonpreferred},i} - S_{\text{preferred},i}$, measured in nats. Positive values indicate lower surprisal for the contextually preferred continuation. Points represent individual minimal pairs, violin widths show distributional density, and the embedded boxplots summarize the median and interquartile range. The dashed horizontal line marks $d_i = 0$, and the annotations report the percentage of items with $d_i > 0$. Blue and orange denote BA-preferred and SVO-preferred contexts, respectively.

C Supplementary Results for H2

C.1 Source-Label Alignment Without Context

To examine model preferences without preceding textual context, we applied the H1 decision rule to the original empty-prefix scores. For model m and pair j ,

$$\hat{c}_{mj}^{\emptyset} = \begin{cases} \text{BA}, & S_{\text{sent},m}(T_j^{\text{BA}} \mid \emptyset) < S_{\text{sent},m}(T_j^{\text{SVO}} \mid \emptyset), \\ \text{SVO}, & \text{otherwise}, \end{cases}$$

$$A_{mj}^{\emptyset} = \mathbf{1}\{\hat{c}_{mj}^{\emptyset} = y_j\},$$

where y_j is the intended preference label from the original discourse context. BA was selected only when its sentence-level mean surprisal was strictly lower than that of SVO; ties were assigned to SVO. For each model, the label alignment rate was the mean of $A_{mj}^{\emptyset}$ over the relevant pairs. The model-level rates were then averaged equally across the 21 models and reported overall and separately for BA-preferred and SVO-preferred pairs (Table 6).

Mean alignment was 61.28% overall, 81.38% for BA-preferred pairs, and 40.83% for SVO-preferred pairs. Every model showed higher alignment for BA-preferred pairs, consistent with a tendency to select BA without preceding textual context. This pattern may reflect a general preference for BA, continuation-internal cues, or both. For example, aspectual marking such as *le*($\overline{\text{了}}$) remains available in some BA continuations even when the preceding context is removed.

Model	Overall		BA-preferred		SVO-preferred	
	Valid n	Alignment (%)	Valid n	Alignment (%)	Valid n	Alignment (%)
GLM-4-9B-Base [†]	194	65.46	100	82.00	94	47.87
InternLM2.5-7B-Base	200	59.50	100	67.00	100	52.00
Llama-2-7B-Base	200	55.50	100	100.00	100	11.00
Llama-3-8B-Base	200	60.50	100	100.00	100	21.00
Llama-3.1-8B-Instruct	200	71.50	100	97.00	100	46.00
Mistral-7B-v0.3-Base	200	76.00	100	98.00	100	54.00
Pythia-1.4B-Base	200	65.00	100	98.00	100	32.00
Pythia-2.8B-Base	200	58.00	100	100.00	100	16.00
Pythia-410M-Base	200	70.00	100	94.00	100	46.00
Qwen2.5-0.5B-Base [†]	196	58.67	100	76.00	96	40.63
Qwen2.5-1.5B-Base [†]	196	50.00	100	64.00	96	35.42
Qwen2.5-14B-Base [†]	196	63.27	100	87.00	96	38.54
Qwen2.5-3B-Base [†]	196	55.61	100	69.00	96	41.67
Qwen2.5-7B-Base [†]	196	76.02	100	78.00	96	73.96
Qwen3-0.6B-Base [†]	196	58.16	100	69.00	96	46.88
Qwen3-1.7B-Base [†]	196	60.71	100	82.00	96	38.54
Qwen3-14B-Base [†]	196	64.29	100	71.00	96	57.29
Qwen3-4B-Base [†]	196	62.76	100	79.00	96	45.83
Qwen3-8B-Base [†]	196	46.94	100	71.00	96	21.88
Yi-1.5-6B-Base	200	56.50	100	60.00	100	53.00
Yi-1.5-9B-Base	200	52.50	100	67.00	100	38.00
All 21 models	4,154	61.28	2,100	81.38	2,054	40.83

Table 6: Source-label alignment under the original empty-prefix scoring procedure: descriptive estimates by model.

Note. Valid n includes only pairs with finite scores for both continuations; undefined scores were excluded without imputation. For models marked [†], the original empty-prefix implementation lacked a usable sequence-start token, so mean surprisal was computed only over tokens after the first continuation token. Consequently, single-token SVO continuations yielded undefined scores in 46 model-pair observations. Counts in the final row are summed across models and do not represent unique pairs. These approximate descriptive alignment estimates should not be interpreted as grammaticality accuracy.

C.2 Supplementary diagnostics for H2

This appendix provides two supplementary diagnostics for H2. Figure 7 reports SVO-to-BA preference reversal rates in BA-preferred contexts, while Figure 8 provides an additional bar-chart view of mean surprisal reduction by context type. These figures complement the main-text dumbbell plot: the reversal analysis tests whether context can change a model’s preference, whereas the ΔS bar chart summarizes the magnitude of contextual surprisal reduction.

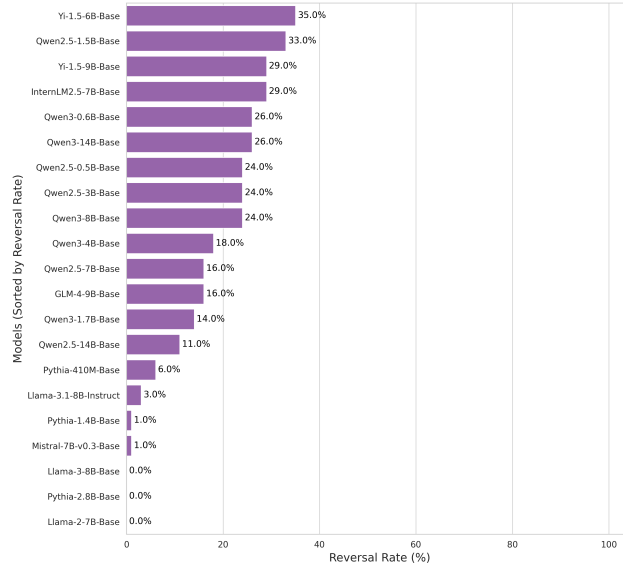


Figure 7: SVO-to-BA preference reversal rates in BA-preferred contexts. A reversal occurs when a model prefers the SVO continuation without context but prefers the *ba* continuation after the BA-licensing context is supplied. Low reversal rates can indicate weaker contextual reweighting, but they may also arise when a model already prefers *ba* in the no-context condition.

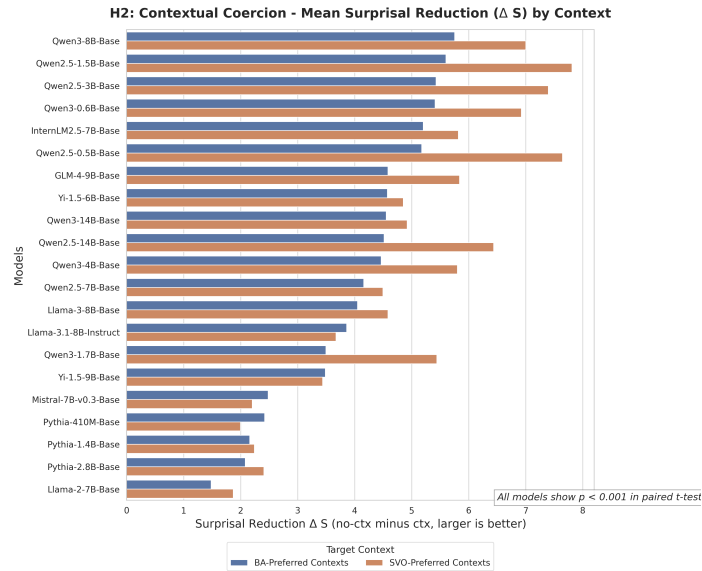


Figure 8: Mean surprisal reduction by context type. Values are reported in nats and computed as no-context sentence-level surprisal minus contextualized sentence-level surprisal.

D Supplementary Results for H3

This appendix provides the full model-level granularity comparison for H3, together with a tokenizer diagnostic for the first-token metric. The diagnostic checks whether the first model token of each target continuation begins with the expected first Chinese character. It does not change the computed surprisal values; it only assesses whether the first-token score is linguistically interpretable across model families.

Model	Sent. Acc.	First Acc.	Start Rate	Warn Rate	Tokenizer Status
Llama-2-7B-Base	82.5	0.0	0.000	1.000	Unstable
Mistral-7B-v0.3-Base	91.5	0.0	0.000	1.000	Unstable
Yi-1.5-6B-Base	89.5	0.0	0.000	1.000	Unstable
Yi-1.5-9B-Base	86.0	0.0	0.000	1.000	Unstable
Pythia-1.4B-Base	84.0	70.5	0.075	0.925	Unstable
Pythia-2.8B-Base	83.0	69.0	0.075	0.925	Unstable
Pythia-410M-Base	77.5	69.0	0.075	0.925	Unstable
Llama-3-8B-Base	92.0	77.5	0.915	0.085	Mostly stable
Llama-3.1-8B-Instruct	88.0	73.5	0.915	0.085	Mostly stable
GLM-4-9B-Base	91.5	68.5	1.000	0.000	Stable
InternLM2.5-7B-Base	89.0	51.5	1.000	0.000	Stable
Qwen2.5-0.5B-Base	86.0	69.0	1.000	0.000	Stable
Qwen2.5-1.5B-Base	90.5	67.5	1.000	0.000	Stable
Qwen2.5-3B-Base	90.5	78.5	1.000	0.000	Stable
Qwen2.5-7B-Base	88.0	71.5	1.000	0.000	Stable
Qwen2.5-14B-Base	90.5	78.0	1.000	0.000	Stable
Qwen3-0.6B-Base	87.0	76.0	1.000	0.000	Stable
Qwen3-1.7B-Base	85.0	67.5	1.000	0.000	Stable
Qwen3-4B-Base	90.0	71.5	1.000	0.000	Stable
Qwen3-8B-Base	91.0	73.0	1.000	0.000	Stable
Qwen3-14B-Base	93.5	75.5	1.000	0.000	Stable
Mean	87.93	57.50	—	—	—

Table 7: H3 granularity results and tokenizer diagnostic. Sent. Acc. is sentence-level surprisal accuracy; First Acc. is first-token surprisal accuracy. Start Rate is the proportion of contextualized continuations whose first model token begins with the expected first Chinese character. Warn Rate is the proportion of cases where the first model token does not begin with the expected first Chinese character and is therefore less stable for linguistic interpretation. The diagnostic was computed over 400 contextualized continuations for each model.

Note. A high Start Rate indicates that the first-token score is relatively interpretable as the onset of the BA or SVO continuation. A high Warn Rate indicates that the first model token may correspond to byte-level or sub-character fragments rather than to a linguistically meaningful Chinese character. Exact character matching is not required here, because a stable tokenizer may still encode a multi-character token that begins with the correct first character. For Sent. Acc. and First Acc., a pair was counted as correct only when the source-context-preferred continuation had strictly lower first-token surprisal than the alternative; equal or higher values were counted as incorrect. The **zero** accuracies reported for some models include tied BA and SVO scores and therefore do not necessarily indicate a consistent preference for the context-dispreferred continuation.