Prompt as Stratified Interface: Decomposing Label-Mediated Effects in Chinese-English NLI

Anonymous ACL submission

Abstract

Multilingual prompting studies often compare English instructions with target-language instructions. For classification, however, this comparison bundles more than instruction wording: prompts also define categories, name them, bind them to answer symbols, and specify an output form. We study this label interface in Chinese-English XNLI with examples, gold labels, and decoding fixed. Under ordinary semantic-label prompting, Qwen3-8B shows large CN-interface gains on both input sides, larger Qwen3 models are near zero, and Llama-3.1-70B shows a smaller gain concentrated on Chinese inputs. Removing visible category names with all six A/B/C permutations nearly eliminates the Qwen3-8B gains, while Llama-3.1-70B retains a Chinese-input residual. Restoring names inside A/B/C options improves accuracy and reduces the largest mapping instabilities, but signed contrasts remain model-specific. Direct semantic-label output adds only small gains once category names are visible. We use *semantic label anchoring* as a descriptive shorthand for this verbalizer-like interface pattern: meaningful category names stabilize NLI prompting relative to arbitrary option symbols. The contribution is a scoped interface decomposition showing that label names, option mappings, and output contracts can change what is measured as a prompt-language effect in generative NLI classification.

1 Introduction

Multilingual LLM evaluations often ask whether non-English inputs should be paired with English instructions or with instructions written in the language of the input (Ahuja et al., 2023; Lai et al., 2023; Lin et al., 2022). Recent work has made this question more precise by showing that prompt-language comparisons depend on translationese, task construction, and which prompt components are translated (Enomoto et al., 2025; Mondshine et al., 2025). We follow that component view but focus on the label interface, a central part of classification prompts often treated as formatting.

Natural language inference (NLI) makes the issue concrete. A standard NLI prompt does not merely ask a model to classify. It presents a premise and hypothesis, defines *entailment*, *neutral*, and *contradiction*, names those categories, and constrains the answer form. In XNLI, the underlying label space is stable across languages (Conneau et al., 2018), but the prompt surface is not. Translating an NLI prompt between English and Chinese can change the directive, relation definitions, category names, and output labels together. A measured difference between Chinese-language and English-language prompting may therefore be a whole-interface effect rather than an instruction-language effect.

This ambiguity is not resolved by replacing labels with A/B/C. Prompt-learning work has long shown that label words and verbalizers influence classification (Schick and Schütze, 2021; Gao et al., 2021; Hu et al., 2022). Multiple-choice work likewise shows that option identifiers and answer positions can influence model behavior (Zheng et al., 2024; Pezeshkpour and Hruschka, 2024; Yang et al., 2025b; Xue et al., 2024). These findings create a puzzle for multilingual classification prompting: semantic labels may carry task meaning, but abstract symbols introduce their own binding problem. What appears to be a language effect may depend on how category names, symbols, and output contracts are arranged.

We treat the prompt as a stratified task interface. Here, a *stratum* is an operational layer of the prompt that serves a distinct function in the classification interaction: input language, directive wording, relation definitions, lexical category names, symbolic answer identifiers, and output contract. The term draws on linguistic accounts of language as layered meaning-making (Halliday, 1978; Halliday and Matthiessen, 2014), formal semantic accounts of entailment and contradiction as structured rela-

tions (Montague, 1970; Heim and Kratzer, 1998; Janssen and Zimmermann, 2025), and evidence that labels can guide categorization (Waxman and Markow, 1995; Lupyan et al., 2007; Lupyan, 2012; Goldstone, 1994). Our use is deliberately behavioral: the strata are prompt components that can be reorganized while examples and gold labels are fixed.

The empirical scope is narrow by design. We study Chinese-English XNLI as a controlled bilingual testbed, not as a universal benchmark of multilingual reasoning. We cross input language and instruction language in four conditions: Chinese input with Chinese instruction (ZH-ZH), Chinese input with English instruction (ZH-EN), English input with English instruction (EN-EN), and English input with Chinese instruction (EN-ZH). The main metric is *CN-interface gain*: for Chinese inputs, Acc(ZH-ZH) − Acc(ZH-EN); for English inputs, Acc(EN-ZH) − Acc(EN-EN). Positive values mean that the Chinese-language task interface outperforms the English-language task interface on the same input side.

The experiments form a diagnostic ladder. First, we measure the ordinary semantic-label interface across a Qwen3 scale ladder and Llama-3.1 models. Second, we remove visible category names and require A/B/C output while preserving relation definitions. Third, we restore category names inside A/B/C options while keeping symbolic output. Fourth, we keep the anchored options but ask for semantic-label output in selected conditions. This sequence does not identify every prompt stratum separately: directive wording and relation definitions remain translated together. It tests how the label-interface layers studied here–category-name visibility, A/B/C binding, and output form–change the measured prompt-language contrast.

The results are model-specific. Qwen3-8B shows large CN-interface gains under semantic labels; those gains nearly vanish under ABC-only prompting and partly return when category names are restored. Llama-3.1-70B shows a smaller semantic-label gain concentrated on Chinese inputs; after label removal, a Chinese-input residual remains, and anchoring improves the English-instruction condition more strongly than the Chinese-instruction condition. Across both families, meaningful category names reduce the largest A/B/C mapping instabilities but do not eliminate all symbol sensitivity.

The contribution is not the general claim that label words matter, which is already established by the verbalizer literature. Nor is *semantic label anchoring* introduced as a new mechanism. The contribution is the controlled multilingual combination: input language and instruction language are crossed while category-name visibility, option mappings, and output form are reorganized on the same Chinese-English NLI examples. The central finding is that the label interface can change what is measured as a prompt-language effect.

2 Prompt Interfaces as Strata

2.1 Operational layers in NLI prompting

Prompt-language comparisons often treat a prompt as if it had a single language attribute. Classification prompts are compositionally structured: examples, directives, definitions, label names, and output strings can be aligned to the input language or mixed across languages. Both choices are natural, but they modify several layers of the task interface at once.

Table 1 defines the operational strata studied here. The table is an experimental map, not a complete linguistic theory. It also clarifies the scope of the paper: input language is crossed with instruction language; directive wording and relation definitions are translated together; category-name visibility, A/B/C mappings, and output form are directly manipulated.

Two distinctions drive the experiments. First, definitions and category names are different layers. A definition such as "if Sentence 1 is true, Sentence 2 must be true" describes a relation; a label such as *entailment* or 蕴含 names it. Second, category names and output form are different layers: a model may see *entailment* inside an option while still being required to output A.

2.2 Label words, verbalizers, and semantic label anchoring

NLI labels are both task identifiers and meaningful words. The English label *contradiction* and the Chinese label 矛盾 carry lexical-semantic content that can help bind a definition to a category. Replacing them with A/B/C removes that cue but adds a different burden: the model must track which arbitrary symbol corresponds to which relation in the current prompt.

We use *semantic label anchoring* as a descriptive term for the observed prompt-surface pattern in which meaningful category names improve accuracy or reduce option-mapping instability relative

Stratum	Role in NLI prompting	Handling in this paper
Input	Presents premise and hypothesis	Chinese vs. English XNLI inputs, crossed with instruction language
Directive	Tells the model to classify	Chinese vs. English wording; bundled with definitions
Relation definition	Defines entailment, neutral, contradiction	Translated with the instruction; visible in all regimes
Category name	Supplies meaningful labels	Present, hidden, or restored
Symbol binding	Maps relations to A/B/C	All six permutations in ABC-only and anchored regimes
Output contract	Specifies required answer form	A/B/C vs. semantic-label output; semantic-output control run only in selected conditions

Table 1: Operational prompt strata. Direct diagnostic interventions are category-name visibility, A/B/C binding, and output form; directive wording and relation definitions remain bundled.

to abstract symbols. The term does not specify an internal mechanism, nor imply that label semantics explain every prompt-language difference. It names a measurable behavior of the interface.

This term should be read as an interface-level instance of the broader verbalizer effect, not as a competing theory. Prior prompt-learning work establishes that class-to-word mappings influence classification, that label-word choices can introduce bias and variance, and that richer label-word semantics can stabilize performance (Schick and Schütze, 2021; Gao et al., 2021; Shin et al., 2020; Hu et al., 2022). Our question is different. We do not ask whether one verbalizer is better than another in isolation. We ask how verbalizer-like category names interact with instruction language, arbitrary answer symbols, and output contracts in a multilingual generative NLI prompt.

The stratified framing adds four ingredients beyond a bilingual verbalizer comparison. First, the measured outcome is not only task accuracy, but the signed Chinese-versus-English interface contrast on the same input side. Second, anchored ABC separates category-name visibility from final output form: the model can see a semantic category name while still returning A/B/C. Third, all six A/B/C permutations make symbol-binding instability measured rather than hidden. Fourth, residual effects after label removal are interpreted cautiously as evidence that bundled non-label strata may still contribute. Thus, semantic label anchoring describes how verbalizer-like semantics behave inside a stratified multilingual prompt interface.

This framing combines existing concerns rather than replacing them. Multilingual prompting work shows that English and target-language prompts can behave differently, and component-level analyses show that prompt translation is not a single intervention (Ahuja et al., 2023; Lai et al., 2023; Lin et al., 2022; Enomoto et al., 2025; Mondshine et al., 2025). Multiple-choice studies show that option identifiers and option order can affect LLM behavior (Zheng et al., 2024; Pezeshkpour and Hruschka, 2024; Yang et al., 2025b; Xue et al., 2024). NLI is a useful testbed because its label space is small, interpretable, and standardized across SNLI, MultiNLI, and XNLI (Bowman et al., 2015; Williams et al., 2018; Conneau et al., 2018), while still requiring caution because NLI datasets contain artifacts (Gururangan et al., 2018). We bring these threads together in a Chinese-English generative classification interface.

2.3 Identification scope

The design supports behavioral localization of prompt-surface effects, not internal-mechanism claims. If a CN-interface gain collapses under ABC-only prompting, the broad contrast depends on the visible semantic-label interface under the tested prompts. If a residual remains, the design indicates a remaining contribution from bundled non-label strata, most plausibly directive wording, relation-definition wording, or their interaction with input language. Because the current prompts translate directives and definitions together, those two layers are not separately identified.

The contrasts should therefore be read as diagnostic interventions. Semantic-label prompting estimates a broad interface contrast. ABC-only prompting tests what remains after category names and semantic output labels are removed, while also measuring the A/B/C binding burden. Anchored ABC tests whether restoring category names changes accuracy or mapping variance without changing the required answer form. Semantic-output anchoring tests whether final answer strings add anything once category names are already visible. This framing assigns each result to a tested prompt-surface in-

tervention rather than to "Chinese instruction" or "English instruction" alone.

3 Experimental Design

3.1 Data and Conditions

We use the XNLI test split (Conneau et al., 2018). XNLI is appropriate for this study because it provides the same three-class NLI interface in Chinese and English. The choice of a single dataset and two languages is intentional: it keeps the task label space, examples, and gold labels fixed while the prompt interface changes. The study is therefore a controlled bilingual interface analysis.

The main sample contains 600 Chinese-input examples and 600 English-input examples, balanced as 200 entailment, 200 neutral, and 200 contradiction cases per language. The same examples are used across prompt regimes. Appendix C reports a disjoint alternative-sample check for Qwen3-8B and Llama-3.1-8B.

The four prompt-language conditions are ZH-ZH, ZH-EN, EN-EN, and EN-ZH. CN-interface gain is computed separately for Chinese and English inputs:

$$\text{Gain}_{\text{ZH-input}} = \text{Acc}(\text{ZH-ZH}) - \text{Acc}(\text{ZH-EN}),$$
$$\text{Gain}_{\text{EN-input}} = \text{Acc}(\text{EN-ZH}) - \text{Acc}(\text{EN-EN}).$$

3.2 Model Selection

We use two open model families because the study depends on controlled reruns across prompt regimes, model sizes, and family profiles. Qwen3 is a multilingual model family with strong Chinese coverage; its technical report describes dense models from 0.6B to 32B parameters and MoE models up to Qwen3-235B-A22B, with expanded multilingual support (Yang et al., 2025a). We evaluate Qwen3-8B, Qwen3-32B, and Qwen3-235B-A22B-Instruct-2507 in the semantic-label baseline. This within-family ladder asks whether a Chinese-English prompt-interface effect attenuates in larger Qwen3 models.

Llama-3.1 supplies a cross-family contrast from a different open-weight release lineage. The Llama 3 Herd report describes Llama-3.1 8B and 70B Instruct models as instruction-tuned multilingual long-context models (Grattafiori et al., 2024). We evaluate Llama-3.1-8B-Instruct and Llama-3.1-70B-Instruct in the semantic-label baseline. The goal is not to rank models; it is to test whether label-interface effects have the same profile across two release families.

The diagnostic regimes focus on Qwen3-8B and Llama-3.1-70B because the semantic-label baseline makes them contrastive and informative profiles. Qwen3-8B shows the largest CN-interface gains; Llama-3.1-70B shows a smaller but reliable Chinese-input gain with a different family profile. Larger Qwen3 models serve as semantic-label boundary cases; Appendix F adds an ABC-only boundary check for Llama-3.1-8B, but not a full anchored-ABC ladder. The ladder should therefore be read as an intensive analysis of two model profiles, not as evidence that the same decomposition holds for every model in the baseline set.

3.3 Prompt Regimes

The experiments form a four-step ladder. In the semantic-label baseline, the system message identifies the model as a sentence-relation classifier. The user message defines entailment, neutral, and contradiction, then asks the model to return one semantic label. Chinese-instruction conditions use Chinese definitions and Chinese label words; English-instruction conditions use English definitions and English label words. This approximates a practical semantic-label NLI prompt.

In ABC-only abstraction, the model sees three options, A/B/C, each paired with a relation definition, and must return the option letter. Semantic category names in both languages are removed. Relation definitions remain in the instruction language. We run all six permutations between the three NLI relations and A/B/C, so every relation appears twice under each option and every option order is observed.

In anchored ABC, the output remains A/B/C, but each option is written with a category name and a definition, such as "A (*entailment*): ..." or "A (蕴含): ...". English-instruction conditions use English anchors; Chinese-instruction conditions use Chinese anchors. This regime tests whether category names stabilize the symbolic interface when the required answer remains abstract.

The semantic-output anchor keeps the same anchored options but asks for the semantic label rather than A/B/C. Because this control is run only in selected Chinese-instruction conditions, it is analyzed as a condition-level output-contract control, not as a full CN-interface-gain component.

3.4 Decoding and Statistical Tests

All runs use deterministic decoding with temperature 0. We report accuracy, macro-F1, invalid-output rate, CN-interface gain, and mapping vari-

ance. Invalid outputs and API errors are zero in all main results; supplementary checks use latest-row deduplication and report repair details in Appendix F. We compute 95% paired bootstrap confidence intervals over examples using 2,000 resamples. Paired condition comparisons use McNemar tests over example-level correctness. For multi-mapping regimes, condition estimates average across the six fixed A/B/C permutations; mapping ranges and standard deviations are reported separately.

4 Results

4.1 Semantic-label prompts produce model-specific interface effects

The semantic-label baseline asks whether Chinese and English task interfaces behave as interchangeable ways of expressing the same classification request. They do not. Table 2 shows that the whole-interface effect varies by model family, model size, and input side, visualized in Figure 1.

Model	ZH in.	EN in.	Mean
Qwen3-8B	+.085	+.090	+.087*
Qwen3-32B	-.007	-.010	-.008
Qwen3-235B	+.002	-.008	-.003
Llama-8B	+.022	-.015	+.003
Llama-70B	+.067	+.023	+.045*

Table 2: Semantic-label gains. Mean averages the two input sides; * marks a 95% paired bootstrap interval excluding zero. Full intervals are in Appendix D.1.

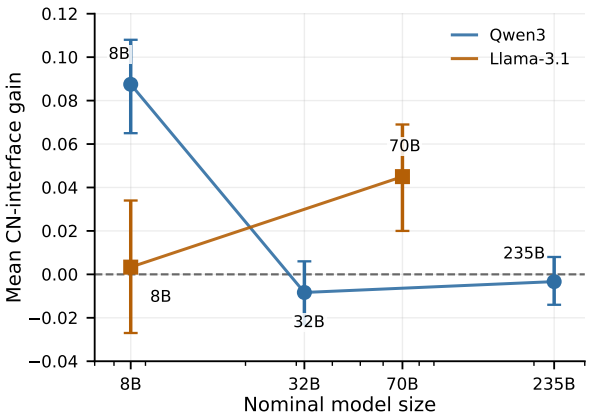

Figure 1: Semantic-label gains by model. Points show mean CN-interface gain with 95% paired bootstrap intervals. Lines are descriptive within family, not fitted scaling trends.

Qwen3-8B shows a large gain on both input sides. In the same family, Qwen3-32B and Qwen3-235B are near zero, so the pattern should not be described as a general scaling law. The Llama pair has a different profile: Llama-3.1-8B is near zero, while Llama-3.1-70B has a smaller positive gain concentrated on Chinese inputs. The baseline therefore establishes the empirical puzzle: under ordinary semantic labels, prompt-language effects are strong in some model-family and size combinations and weak in others.

At this stage, the effect is still a broad interface contrast. Directive wording, relation definitions, category names, and output labels all move together. The next regimes test how much the observed contrast changes when the label interface is reorganized.

4.2 ABC-only removes Qwen3-8B gains but leaves a Llama residual

ABC-only prompting removes semantic category names while preserving relation definitions. If directive and definition language are sufficient to preserve the baseline contrast, the gain should remain. If visible category names and output labels are central to the observed contrast, the contrast should shrink. Table 3 summarizes the resulting ABC-only residuals, averaging over all six A/B/C permutations.

Model	Side	Sem.	ABC	Red.
Qwen3-8B	ZH	+.085	+.002	98%
Qwen3-8B	EN	+.090	.000	∼100%
Llama-70B	ZH	+.067	+.031*	53%
Llama-70B	EN	+.023	-.007	69%

Table 3: ABC-only residuals. ABC values average over all six A/B/C permutations; * marks a 95% interval excluding zero. Full intervals are in Appendix D.1.

For Qwen3-8B, removing category names nearly eliminates both CN-interface gains. Accuracy drops especially in Chinese-instruction conditions: ABC-only is .140 lower than the semantic baseline in ZH-ZH and .132 lower in EN-ZH. Under these prompts, the Qwen3-8B baseline contrast strongly depends on the visible category-name/output-label layer.

Llama-3.1-70B has a different profile. Its English-input gain becomes effectively zero, but its Chinese-input gain remains +.031. This residual shows that removing visible category names is not enough to eliminate the prompt-language contrast. Because directive and definition language remain bundled, the current design does not assign the residual to a specific non-label layer.

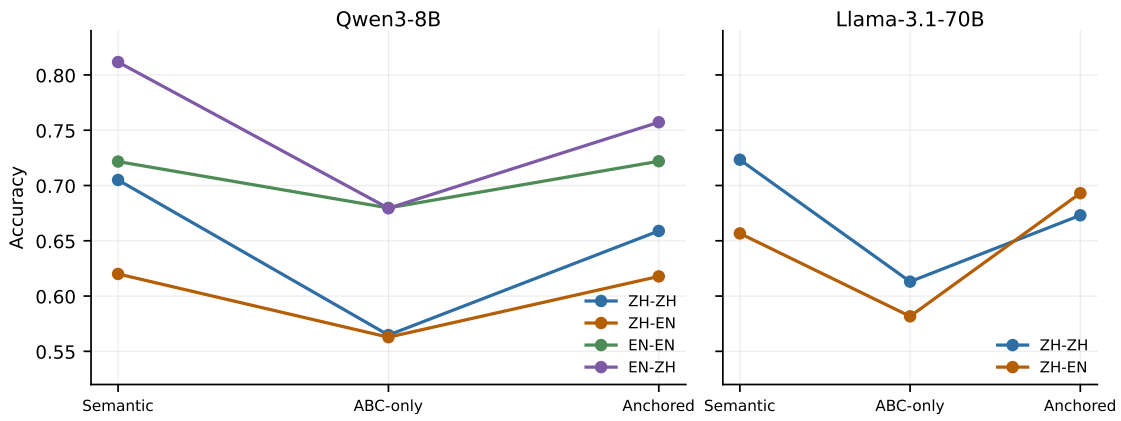

Figure 2: Diagnostic ladder. Qwen3-8B is shown for all four prompt-language conditions; Llama-3.1-70B is shown for Chinese-input conditions with anchored diagnostics.

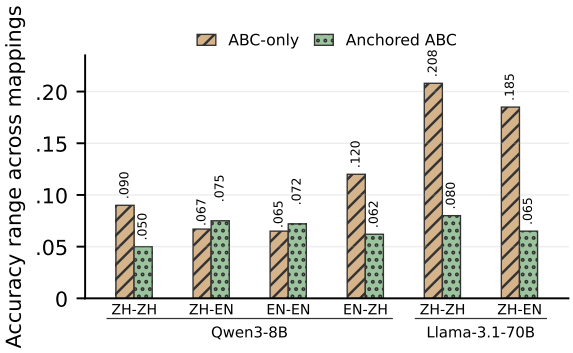

Figure 3: Mapping variance. Bars show accuracy range across all six permutations; hatch patterns distinguish ABC-only and anchored ABC.

ABC-only prompting also exposes option-binding instability. Because all six mappings are tested, the A/B/C condition is not treated as a neutral control. It is an additional interface layer whose mapping variance can be measured.

4.3 Anchoring changes accuracy, variance, and signed contrast

Anchored ABC keeps the A/B/C output contract but restores category names inside the options. This condition asks whether semantic names stabilize the symbolic interface without changing the final answer format. Table 4 reports the anchored contrasts and the shift from ABC-only to anchored ABC.

For Qwen3-8B, anchoring raises accuracy in every condition and restores positive CN-interface gains: +.041 on Chinese inputs and +.035 on English inputs. This is not a full reconstruction of the semantic-label gains, but it shows that making category names visible again changes the contrast while the model still outputs A/B/C. Anchoring also improves the conditions that suffered most when labels were removed; the average anchor improvement is .037 larger in Chinese-instruction conditions than in English-instruction conditions. Figure 2 shows the same movement at the condition level.

Model	Side	ABC	Anchor	Shift
Qwen3-8B	ZH	+.002	+.041*	+.039
Qwen3-8B	EN	.000	+.035*	+.036
Llama-70B	ZH	+.031	-.020*	-.051

Table 4: Anchored contrasts. Shift is anchored gain minus ABC-only gain; * marks a 95% interval excluding zero. Full intervals are in Appendix D.1.

Llama-3.1-70B shows the same accuracy-improvement pattern but a different contrast. Anchoring raises ZH-ZH from .613 to .673 and ZH-EN from .582 to .693. Because ZH-EN improves more, the anchored Chinese-input contrast becomes -.020. This reversal does not mean that anchoring fails; it means that anchoring changes the measured prompt-language contrast by improving conditions asymmetrically.

Mapping variance directly tests whether visible category names stabilize A/B/C binding. Anchored ABC reduces the largest ranges, although not uniformly across Qwen3-8B conditions. The clearest reductions are Qwen3-8B EN-ZH (.120 to .062) and Llama-3.1-70B ZH-ZH/ZH-EN (.208 to .080; .185 to .065). Thus restoring category names generally makes option mappings less consequential, while residual sensitivity remains (Table 5; Figure 3).

4.4 Semantic output is secondary

The anchored regime makes category names visible but still requires A/B/C output. The semantic-output control keeps the same anchored options but asks for the category name rather than the option

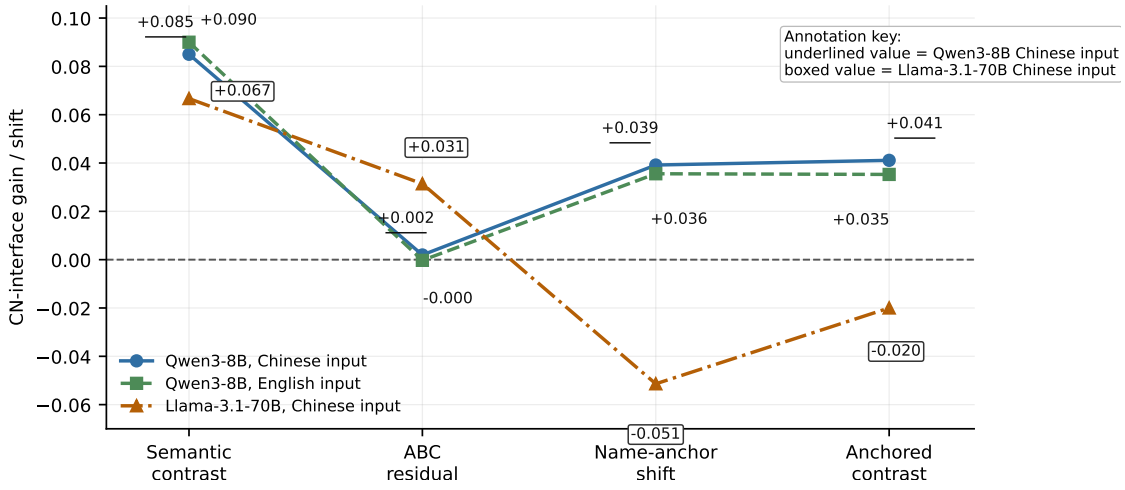

Figure 4: Gain decomposition. Semantic-label contrast is the broad contrast; ABC residual follows category-name removal; name-anchor shift follows restoration inside A/B/C; anchored contrast is the resulting contrast.

Model	Cond.	ABC	SD	Anchor	SD
Qwen3-8B	ZH-ZH	.090	.038	.050	.020
Qwen3-8B	ZH-EN	.067	.027	.075	.030
Qwen3-8B	EN-EN	.065	.023	.072	.024
Qwen3-8B	EN-ZH	.120	.049	.062	.023
Llama-70B	ZH-ZH	.208	.083	.080	.029
Llama-70B	ZH-EN	.185	.070	.065	.026

Table 5: Mapping variance across all six A/B/C permutations. Anchored ABC uses instruction-matched category names.

letter.

For Qwen3-8B, semantic-output adds about .012–.013 accuracy in the two tested Chinese-instruction conditions. This is smaller than the category-name restoration effect and remains close to the bootstrap boundary. The main recovery therefore occurs when semantic category names become visible inside the interface, before the output contract changes from A/B/C to semantic labels. Llama-3.1-70B has a similarly small positive mean gain in ZH-ZH, but the mapping-averaged interval crosses zero and mapping-level tests are mixed. Output form may be a secondary factor, but the current evidence is not decisive. Table 6 reports this condition-level output-contract control.

Aggregate accuracy also hides class-specific movement. Appendix E.2 shows that the ABC-only drop is not evenly distributed across entailment, neutral, and contradiction. For Qwen3-8B, contradiction F1 falls sharply when category names are removed in Chinese-instruction conditions and recovers under anchoring. For Llama-3.1-70B on Chinese inputs, entailment degrades strongly under ABC-only, while contradiction remains comparatively robust. These patterns support the interface interpretation at the label level, but we treat them descriptively as the present design does not identify the mechanism behind each class's sensitivity.

5 Interface Decomposition and Discussion

Table 7 and Figure 4 summarize the diagnostic ladder as a sequence of observable label-interface contrasts. The semantic-label contrast is the ordinary whole-interface effect. The ABC residual is the remaining contrast after category names are removed. The name-anchor shift is the change after names are restored inside A/B/C options, and the resulting value is the *anchored contrast*. The semantic-output control is kept separate because it is a condition-level accuracy change in selected Chinese-instruction conditions, not a full contrast-level component.

For Qwen3-8B, the ABC residual is near zero on both input sides. Once semantic category names are removed, little observable contrast remains in this model under the tested prompts. Restoring category names produces the largest contrast shift: +.039 on Chinese inputs and +.036 on English inputs. This supports that the Qwen3-8B semantic-label contrast is highly sensitive to the tested label interface design, even though anchored ABC only partly reconstructs the original magnitude.

For Llama-3.1-70B, the Chinese-input ABC residual remains +.031. An account based only on visible category names is therefore incomplete. Llama-3.1-70B prevents a label-only account: category names stabilize the symbolic interface, but residual Chinese-input sensitivity remains after la-

Model	Condition	Anchored ABC	Semantic-output	Δ Acc.	95% CI
Qwen3-8B	ZH-ZH	.659 / .664	.672 / .677	+.013	[-.001, .027]
Qwen3-8B	EN-ZH	.757 / .758	.769 / .770	+.012	[.000, .025]
Llama-70B	ZH-ZH	.673 / .664	.686 / .674	+.013	[-.009, .036]

Table 6: Semantic-output control. Cells report accuracy/macro-F1; Δ is semantic-output accuracy minus anchored-ABC accuracy.

Model	Contrast	Semantic-label contrast	ABC residual	Name-anchor shift	Anchored contrast
Qwen3-8B	Chinese input	+.085	+.002	+.039	+.041
Qwen3-8B	English input	+.090	.000	+.036	+.035
Llama-70B	Chinese input	+.067	+.031	-.051	-.020

Table 7: Gain decomposition. Values are Chinese-interface accuracy minus English-interface accuracy on the same input side.

bel removal. Category names still matter: anchoring improves accuracy and reduces the largest mapping variance. But because ZH-EN improves more strongly than ZH-ZH, the signed contrast reverses under anchored ABC. Llama-3.1-70B thus shares the anchoring and stabilization pattern while retaining a residual that the present diagnostics do not localize.

Together, these contrasts support the paper's central claim. A broad prompt-language difference can be nearly removed by suppressing category names, partly restored by reintroducing them, redirected by asymmetric condition-level improvement, or left residual when other strata continue to contribute. The measured contrast is not a property of "Chinese instruction" or "English instruction" alone. In these experiments, it is a property of a stratified task interface in which label names, symbols, and output contracts change the measurement.

This interpretation also clarifies the relation to verbalizer work. The Qwen3-8B result is indeed verbalizer-like: meaningful category names outperform abstract symbols. What the present decomposition adds is that this effect changes a multilingual contrast, not only an absolute classification score. The anchored-ABC regime shows that category names can matter even when they are not the final output strings, and the six-permutation design exposes a measurable symbol-binding layer. Appendix F further shows that changing Chinese label wording alters measured gains, and that crossing Qwen3-8B label language against instruction/definition language reverses both original CN-interface gains. Thus the novelty is not a new claim that labels matter, but an account of how label wording, label language, option symbols, and prompt language interact inside a generative multilingual classification interface.

The practical implication is documentation. Generative multilingual classification evaluations, especially NLI-style tasks, should report input/instruction language, definition/label language, semantic versus abstract category names, visible definitions, output form, option mappings, and invalid-output handling. These details are experimental variables: here, they change the measured prompt-language effect.

6 Conclusion

Chinese-English NLI prompt-language effects are sensitive to the label interface. Under ordinary semantic-label prompts, Qwen3-8B shows large CN-interface gains, larger Qwen3 models are near zero, and Llama-3.1-70B shows a smaller gain concentrated on Chinese inputs. Removing semantic category names with all six A/B/C permutations nearly eliminates the Qwen3-8B gains but leaves a Llama-3.1-70B Chinese-input residual. Restoring category names improves accuracy and reduces the largest mapping instabilities, while preserving model-specific signed contrasts. Direct semantic-label output adds only small gains beyond visible category names.

The broader claim is scoped but consequential: for generative NLI classification, prompt language is a stratified interface rather than a single instruction variable. Semantic label anchoring names one observable pattern within that interface: meaningful category names stabilize NLI prompting relative to arbitrary symbols. Labels, mappings, and output contracts should therefore be treated as experimental variables in multilingual classification prompting.

Limitations

The study focuses on Chinese-English XNLI and a targeted set of Qwen3 and Llama-3.1 models. Other language pairs, tasks, and model families may have different anchoring profiles. XNLI is useful as a controlled interface testbed, but NLI datasets contain artifacts and are not direct measurements of reasoning ability. The results should therefore be read as prompt-interface behavior on a fixed benchmark, not as a general account of inference.

The diagnostic scope is intentionally limited. The experiments directly manipulate category-name visibility, A/B/C symbol mappings, and output form. They do not fully separate directive wording from relation-definition wording, because definitions are translated with the instruction in the current prompt-language conditions. Residual effects after label removal indicate a remaining contribution from non-label strata, but not from any specific non-label layer.

The ABC-only and anchored ABC regimes use all six A/B/C permutations for the diagnostic models. This strengthens the mapping-variance estimates, but it does not fully separate symbol identity, option position, relation order, and token-level preference. The analysis is also behavioral: it reports accuracy, F1, mapping variance, and output validity, but does not use token-level probabilities, calibration tests, or representation-level analyses to distinguish output prior, lexical frequency, label tokenization, and semantic-representation accounts.

The experiments use one prompt template per regime. The claims are therefore about the tested prompt family rather than all possible NLI prompt formulations. The diagnostic models are Qwen3-8B and Llama-3.1-70B, selected because the semantic-label baseline made them the most informative cases for label-interface analysis. Larger Qwen3 models remain semantic-label boundary cases. Appendix F includes an ABC-only boundary check for Llama-3.1-8B, but not a full anchored-ABC diagnostic ladder for that model.

Interface-sensitive evaluation can also misstate multilingual capability. If a language appears weaker or stronger because of label language, option mapping, or output contract rather than task competence, deployment decisions may misrepresent support for non-English users. This risk is especially important for lower-resource languages, where evaluation artifacts may be mistaken for model capability gaps.

References

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2023. MEGA: Multilingual evaluation of generative AI. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4232–4267, Singapore. Association for Computational Linguistics.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Taisei Enomoto, Hwichan Kim, Zhousi Chen, and Mamoru Komachi. 2025. A fair comparison without translationese: English vs. target-language instructions for multilingual LLMs. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 649–670, Albuquerque, New Mexico. Association for Computational Linguistics.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3816–3830, Online. Association for Computational Linguistics.

R. Goldstone. 1994. Influences of categorization on perceptual discrimination. *J Exp Psychol Gen*, 123(2):178–200. Goldstone, R S07 RR7031N/RR/NCRR NIH HHS/United States Journal Article Research Support, U.S. Gov't, P.H.S. United States 1994/06/01 J Exp Psychol Gen. 1994 Jun;123(2):178-200. doi: 10.1037//0096-3445.123.2.178.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

M. A. K. Halliday. 1978. *Language as social semiotic : the social interpretation of language and meaning*. Edward Arnold, London.

M. A. K. Halliday and Christian M. I. M. Matthiessen. 2014. *Halliday's introduction to functional grammar*, fourth edition. edition. Routledge, Milton Park, Abingdon, Oxon. Includes bibliographical references and index.

Irene Heim and Angelika Kratzer. 1998. *Semantics in generative grammar*. Blackwell textbooks in linguistics ; 13. Blackwell, Malden, MA. Includes bibliographical references and index.

Shengding Hu, Ning Ding, Huadong Wang, Zhiyuan Liu, Jingang Wang, Juanzi Li, Wei Wu, and Maosong Sun. 2022. Knowledgeable prompt-tuning: Incorporating knowledge into prompt verbalizer for text classification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2225–2240, Dublin, Ireland. Association for Computational Linguistics.

Theo M. V. Janssen and Thomas Ede Zimmermann. 2025. Montague Semantics. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*, Spring 2025 edition. Metaphysics Research Lab, Stanford University.

Viet Dac Lai, Nghia Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Huu Nguyen. 2023. ChatGPT beyond English: Towards a comprehensive evaluation of large language models in multilingual learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13171–13189, Singapore. Association for Computational Linguistics.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, and 2 others. 2022. Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9019–9052, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Gary Lupyan. 2012. Linguistically modulated perception and cognition: The label-feedback hypothesis. *Frontiers in Psychology*, Volume 3 - 2012.

Gary Lupyan, David H. Rakison, and James L. McClelland. 2007. Language is not just for talking: Redundant labels facilitate learning of novel categories. *Psychological Science*, 18(12):1077–1083. PMID: 18031415.

Itai Mondshine, Tzuf Paz-Argaman, and Reut Tsarfaty. 2025. Beyond English: The impact of prompt translation strategies across languages and tasks in multilingual LLMs. In *Proceedings of the Eighth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2025)*, pages 81–104, Albuquerque, New Mexico, U.S.A. Association for Computational Linguistics.

Richard Montague. 1970. Universal grammar. *Theoria*, 36(3):373–398.

Pouya Pezeshkpour and Estevam Hruschka. 2024. Large language models sensitivity to the order of options in multiple-choice questions. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2006–2017, Mexico City, Mexico. Association for Computational Linguistics.

Timo Schick and Hinrich Schütze. 2021. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269, Online. Association for Computational Linguistics.

Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. 2020. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4222–4235, Online. Association for Computational Linguistics.

Sandra R. Waxman and Dana B. Markow. 1995. Words as invitations to form categories: Evidence from 12- to 13-month-old infants. *Cognitive Psychology*, 29(3):257–302.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Mengge Xue, Zhenyu Hu, Liqun Liu, Kuo Liao, Shuang Li, Honglin Han, Meng Zhao, and Chengguo Yin. 2024. Strengthened symbol binding makes large language models reliable multiple-choice selectors. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4331–4344, Bangkok, Thailand. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025a. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Zhen Yang, Ping Jian, and Chengzhi Li. 2025b. Option symbol matters: Investigating and mitigating multiple-choice option symbol bias of large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1902–1917, Albuquerque, New Mexico. Association for Computational Linguistics.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2024. Large language models are not robust multiple choice selectors. In *International Conference on Learning Representations*, volume 2024, pages 19426–19454.

A Released Artifacts

An anonymized review artifact is uploaded in the data section of the submission and will be released online after publication. The artifact is designed to support verification of the reported analyses without redistributing XNLI text. It therefore includes sample identifiers and labels, but not the original premise or hypothesis sentences.

The artifact contains the main balanced sample identifiers, the disjoint alternative-sample identifiers used in Appendix C, and per-example result rows for the experiments reported in the paper and appendices. These include the semantic-label baselines, the alternative-sample check, ABC-only diagnostics, instruction-matched anchored-ABC diagnostics, the semantic-output control, and the supplementary label-interface checks in Appendix F. The released files also include the six A/B/C mappings, prompt-version identifiers, model names, provider endpoints, decoding parameters, thinking-control settings where applicable, a manifest with row counts and checksums, and a README describing the file layout. API keys, provider response IDs, timestamps, local paths, prompts, scripts, and unreported exploratory runs are not included.

The original examples come from XNLI, and the released artifact is intended only for research verification and replication. To keep this use compatible with the original access conditions and to reduce redistribution risk, the artifact releases identifiers, labels, model outputs, and configuration metadata rather than dataset text. Any downstream use should retrieve XNLI through its original distribution channel and follow the applicable license or terms of use.

We checked the released files for fields that could expose identifying or sensitive operational information, including premise text, hypothesis text, API keys, provider response IDs, timestamps, row identifiers, request hashes, and local paths. The artifact also withholds non-label raw outputs, which may contain provider-generated refusal text or fragments related to offensive examples. The experiments use XNLI as distributed; we do not claim that the source dataset is free of offensive or personally identifying content, but the released artifact does not redistribute the underlying sentence text.

Model parameter scale and provider/model identifiers are reported in the paper and in the artifact configuration. The experiments used hosted API inference rather than local GPU training, so GPU hours and hardware allocation are not directly observable. The artifact reports endpoint, model name, decoding settings, maximum output length, and thinking-control settings where applicable. Preprocessing and evaluation used Python 3 with standard data-analysis packages, including `json`, `csv`, `random`, `pandas`, `numpy`, and `scikit-learn`. Stratified sampling used fixed random seeds reported in the artifact metadata. Parsing used exact-match label normalization against the allowed output set for each prompt regime. Accuracy was computed as exact label match, macro-F1 with class-wise unweighted averaging, bootstrap intervals with 2,000 resamples over examples, and McNemar-style paired comparisons over example-level correctness.

ChatGPT-5.5, ChatGPT-5.1-Codex-Max, and Deepseek v3.2 were used to assist coding and formatting; all code was reviewed and approved by the authors.

B Full Semantic-Label Baseline Results: Table B1

C Alternative Sample Check

We test the representativeness of the main 600-example sample by drawing a disjoint balanced sample of 600 Chinese-input and 600 English-input examples from the same XNLI split. We rerun the semantic-label baseline for Qwen3-8B and Llama-3.1-8B, the two smaller models in the semantic-label baseline. Accuracy is computed over all sampled examples, matching the main tables. Invalid-

Model	Condition	Acc.	95% CI	Macro-F1	95% CI
Qwen3-8B	ZH-ZH	.705	[.667, .740]	.712	[.674, .746]
Qwen3-8B	ZH-EN	.620	[.580, .660]	.618	[.577, .658]
Qwen3-8B	EN-EN	.722	[.683, .760]	.727	[.689, .763]
Qwen3-8B	EN-ZH	.812	[.780, .843]	.815	[.783, .846]
Qwen3-32B	ZH-ZH	.750	[.715, .783]	.754	[.721, .786]
Qwen3-32B	ZH-EN	.757	[.722, .790]	.762	[.728, .795]
Qwen3-32B	EN-EN	.840	[.810, .868]	.840	[.811, .868]
Qwen3-32B	EN-ZH	.830	[.800, .858]	.826	[.795, .854]
Qwen3-235B	ZH-ZH	.792	[.758, .825]	.791	[.757, .823]
Qwen3-235B	ZH-EN	.790	[.757, .822]	.789	[.755, .820]
Qwen3-235B	EN-EN	.872	[.845, .898]	.870	[.843, .895]
Qwen3-235B	EN-ZH	.863	[.837, .888]	.859	[.833, .885]
Llama-70B	ZH-ZH	.723	[.688, .760]	.719	[.683, .754]
Llama-70B	ZH-EN	.657	[.618, .693]	.654	[.615, .692]
Llama-70B	EN-EN	.785	[.753, .818]	.790	[.759, .822]
Llama-70B	EN-ZH	.808	[.775, .840]	.807	[.774, .837]
Llama-8B	ZH-ZH	.598	[.558, .637]	.554	[.514, .590]
Llama-8B	ZH-EN	.577	[.537, .617]	.584	[.544, .623]
Llama-8B	EN-EN	.630	[.592, .668]	.640	[.601, .676]
Llama-8B	EN-ZH	.615	[.577, .653]	.567	[.531, .602]

Table B1: Full semantic-label baseline condition results. Invalid outputs and API errors are zero in all rows.

output percentages use received model responses as the denominator, and API errors are reported separately. The bootstrap intervals for the differences from the main sample all include zero, so the alternative sample does not change the small-model conclusions used in the paper. See Table C1 and Table C2.

D Regime Transitions and Uncertainty

D.1 Main-text gain intervals: Table D1

D.2 Regime-transition intervals: Table D2

E Option Bias and Per-Label Behavior

E.1 Option-Bias Summary

High values indicate that one option symbol dominates the model's outputs for at least one mapping. Anchoring lowers the largest dominance values in the high-variance Llama conditions and in Qwen3-8B EN-ZH, while leaving some Qwen conditions still sensitive to option symbols, see Table E1.

E.2 Per-Label F1: Table E2

F Supplementary Label-Interface Checks

This appendix reports three supplementary checks on the original balanced XNLI sample (Table F1). They are not replacements for the main diagnostic ladder. They test whether the reported interface effects depend on one Chinese label translation, whether Qwen3-8B's semantic-label contrast can move when label language is crossed against instruction/definition language, and whether the Llama-3.1-8B near-zero boundary profile remains stable under ABC-only abstraction.

F.1 Alternate Chinese label translation

The main Chinese semantic-label prompt uses 蕴含, 中立, and 矛盾. We rerun the semantic-label baseline with the Chinese labels changed to 可推出, 无法确定, and 冲突. English-instruction templates are unchanged; English-instruction rows are therefore rerun controls. Table F2 summarizes the resulting CN-interface gains.

Table F3 compares the results to the default CN-interface. The alternate Chinese labels reduce Qwen3-8B's positive gains but do not erase them. For Llama-3.1-70B, the Chinese-input gain reverses. Thus the direction of the measured contrast is not a universal property of Chinese labels; it depends on model profile and label wording.

Model	Cond.	Acc.	95% CI	F1	95% CI	ΔAcc.	95% CI	Invalid/resp.	API
Qwen3-8B	ZH-ZH	.687	[.648, .722]	.693	[.655, .728]	-.018	[-.072, .032]	0/599 (0.00%)	1
Qwen3-8B	ZH-EN	.622	[.583, .662]	.617	[.577, .657]	+.002	[-.052, .057]	0/600 (0.00%)	0
Qwen3-8B	EN-EN	.713	[.677, .750]	.718	[.681, .754]	-.008	[-.060, .043]	0/600 (0.00%)	0
Qwen3-8B	EN-ZH	.815	[.783, .847]	.817	[.785, .847]	+.003	[-.042, .047]	0/600 (0.00%)	0
Llama-8B	ZH-ZH	.605	[.565, .642]	.566	[.527, .603]	+.007	[-.048, .062]	0/600 (0.00%)	0
Llama-8B	ZH-EN	.575	[.537, .617]	.586	[.547, .624]	-.002	[-.058, .055]	3/600 (0.50%)	0
Llama-8B	EN-EN	.653	[.613, .690]	.659	[.620, .694]	+.023	[-.032, .077]	0/600 (0.00%)	0
Llama-8B	EN-ZH	.642	[.603, .680]	.585	[.551, .620]	+.027	[-.027, .080]	0/600 (0.00%)	0

Table C1: Semantic-label results on the disjoint alternative sample. ΔAcc. is alternative-sample accuracy minus main-sample accuracy, with independent bootstrap intervals.

Model	Input side	Alt. gain	Δ vs main
Qwen3-8B	Chinese	+.065 [.012, .117]	-.020 [-.095, .053]
Qwen3-8B	English	+.102 [.052, .150]	+.012 [-.057, .078]
Llama-8B	Chinese	+.030 [-.025, .083]	+.008 [-.072, .088]
Llama-8B	English	-.012 [-.065, .042]	+.003 [-.075, .082]

Table C2: CN-interface gain on the alternative sample. Brackets are 95% bootstrap intervals; Δ compares the alternative-sample gain with the main-sample gain.

Main table	Model	Side	Estimate	95% CI
Table 2	Qwen3-8B	mean	+.087	[.065, .108]
Table 2	Qwen3-32B	mean	-.008	[-.023, .006]
Table 2	Qwen3-235B	mean	-.003	[-.014, .008]
Table 2	Llama-8B	mean	+.003	[-.027, .034]
Table 2	Llama-70B	mean	+.045	[.020, .069]
Table 3	Qwen3-8B	Chinese	+.002	[-.012, .016]
Table 3	Qwen3-8B	English	.000	[-.013, .013]
Table 3	Llama-70B	Chinese	+.031	[.015, .048]
Table 3	Llama-70B	English	-.007	[-.023, .009]
Table 4	Qwen3-8B	Chinese	+.041	[.022, .061]
Table 4	Qwen3-8B	English	+.035	[.012, .059]
Table 4	Llama-70B	Chinese	-.020	[-.036, -.004]

Table D1: Full 95% paired bootstrap intervals for the compact main-text gain tables.

F.2 Qwen3-8B label-language swap

The label-language swap keeps instruction and definition language fixed but crosses the visible category/output label language. For example, ZH-ZH is rerun with Chinese instruction and definitions but English labels, while ZH-EN is rerun with English instruction and definitions but Chinese labels. Table F4 reports condition-level movement, and Table F5 reports the resulting CN-interface gains.

The swap reverses both original Qwen3-8B CN-interface gains. Within this pilot scope, the broad Chinese-interface gain is therefore not simply an effect of Chinese directive wording: visible category/output label language can reshape the measured contrast while instruction and definition language are held fixed.

F.3 Llama-3.1-8B ABC-only boundary check

Llama-3.1-8B is a near-zero semantic-label boundary model in the main baseline. We run an ABC-only check with all six mappings to test whether label abstraction is neutral for this model profile. Some hosted rows failed or were invalid; raw logs retain both failed and repaired rows, and the tables

Model	Condition	Comparison	Acc. diff.	95% CI
Qwen3-8B	ZH-ZH	ABC-only minus semantic	-.140	[-.169, -.112]
Qwen3-8B	ZH-ZH	anchored minus ABC-only	+.094	[.073, .116]
Qwen3-8B	ZH-ZH	semantic-output minus anchored	+.013	[-.001, .027]
Qwen3-8B	ZH-EN	ABC-only minus semantic	-.057	[-.082, -.031]
Qwen3-8B	ZH-EN	anchored minus ABC-only	+.055	[.038, .072]
Qwen3-8B	EN-EN	ABC-only minus semantic	-.042	[-.066, -.018]
Qwen3-8B	EN-EN	anchored minus ABC-only	+.042	[.027, .056]
Qwen3-8B	EN-ZH	ABC-only minus semantic	-.132	[-.159, -.104]
Qwen3-8B	EN-ZH	anchored minus ABC-only	+.078	[.056, .101]
Qwen3-8B	EN-ZH	semantic-output minus anchored	+.012	[.000, .025]
Llama-70B	ZH-ZH	ABC-only minus semantic	-.110	[-.143, -.077]
Llama-70B	ZH-ZH	anchored minus ABC-only	+.060	[.039, .081]
Llama-70B	ZH-ZH	semantic-output minus anchored	+.013	[-.009, .036]
Llama-70B	ZH-EN	ABC-only minus semantic	-.075	[-.099, -.050]
Llama-70B	ZH-EN	anchored minus ABC-only	+.111	[.084, .139]
Llama-70B	EN-EN	ABC-only minus semantic	-.059	[-.081, -.036]
Llama-70B	EN-ZH	ABC-only minus semantic	-.090	[-.116, -.064]

Table D2: Bootstrap uncertainty for key regime differences. Differences are reported as the second regime minus the first regime named in the comparison.

Model	Condition	ABC max	Anchor max
Qwen3-8B	ZH-ZH	.760	.587
Qwen3-8B	ZH-EN	.758	.712
Qwen3-8B	EN-EN	.597	.620
Qwen3-8B	EN-ZH	.617	.457
Llama-70B	ZH-ZH	.825	.643
Llama-70B	ZH-EN	.838	.588

Table E1: Largest single-option prediction share observed across the six A/B/C permutations. High values indicate that one option symbol dominates the model's outputs for at least one mapping.

below use latest-row deduplication. All latest rows are valid after repair, with 3,600 rows per condition (600 examples $\times$ six mappings). Because this is an ABC-only boundary check rather than a full anchored ladder, it is used only as supplementary evidence. Table F6, Table F7 and Table F8 summarize the results.

For Chinese inputs, the Llama-3.1-8B ABC-only residual remains near zero. For English inputs, ABC-only creates a positive residual because EN-EN drops more than EN-ZH. This does not reproduce the Qwen3-8B label-mediated pattern; instead, it shows that A/B/C abstraction can reshape a near-zero profile through condition-specific degradation and mapping instability.

G Prompt Templates and A/B/C Mappings

G.1 Semantic-Label Prompt

English system message:

> You are a sentence relationship classifier. Use only Sentence 1 and Sentence 2. Do not translate, rewrite, or explain them. Return only one allowed label.

English user message:

> Classify the relationship between Sentence 1 and Sentence 2.
>
> Allowed labels: entailment: If Sentence 1 is true, Sentence 2 must be true. neutral: If Sentence 1 is true, Sentence 2 may be true or false. contradiction: If Sentence 1 is true, Sentence 2 must be false.
>
> Return only one label: entailment, neutral, or contradiction.

Model	Regime	Condition	Entail.	Neutral	Contrad.
Qwen3-8B	semantic	ZH-ZH	.709	.665	.761
Qwen3-8B	ABC-only	ZH-ZH	.605	.584	.453
Qwen3-8B	anchored	ZH-ZH	.704	.619	.670
Qwen3-8B	semantic-output	ZH-ZH	.699	.614	.717
Qwen3-8B	semantic	EN-ZH	.840	.755	.851
Qwen3-8B	ABC-only	EN-ZH	.780	.644	.594
Qwen3-8B	anchored	EN-ZH	.811	.672	.792
Qwen3-8B	semantic-output	EN-ZH	.810	.676	.822
Llama-70B	semantic	ZH-ZH	.659	.650	.847
Llama-70B	ABC-only	ZH-ZH	.487	.609	.689
Llama-70B	anchored	ZH-ZH	.574	.603	.814
Llama-70B	semantic-output	ZH-ZH	.667	.549	.805
Llama-70B	semantic	ZH-EN	.549	.647	.766
Llama-70B	ABC-only	ZH-EN	.409	.604	.654
Llama-70B	anchored	ZH-EN	.643	.631	.807

Table E2: Selected per-label F1 scores for diagnostic conditions.

Check	Scope	Latest rows	OK	Valid	Errors
Alternate Chinese labels	5 models × 4 conditions	12,000	12,000	12,000	0
Qwen3-8B label-language swap	4 conditions	2,400	2,400	2,400	0
Llama-8B ABC-only boundary	4 conditions × 6 mappings	14,400	14,400	14,400	0

Table F1: Completion check for supplementary label-interface experiments. Counts use latest-row deduplication; earlier failed or repaired rows remain in the raw logs.

Sentence 1: {premise}
Sentence 2: {hypothesis}
Label:

Chinese system message:

你是一个句子关系分类器。仅使用句子 1 和句子 2。不要翻译、改写或解释。仅返回一个允许的标签。

Chinese user message:

判断句子 1 和句子 2 之间的关系。

允许的标签：蕴含：如果句子 1 为真，则句子 2 必定为真。中立：如果句子 1 为真，句子 2 可能为真，也可能为假。矛盾：如果句子 1 为真，则句子 2 必定为假。

仅返回一个标签：蕴含、中立或矛盾。

句子 1：{premise}
句子 2：{hypothesis}
标签：

G.2 ABC-Only Prompt

English system message:

You are a sentence relationship classifier. Use only Sentence 1 and Sentence 2. Do not translate, rewrite, or explain them. Return only one allowed option.

English user message:

Classify the relationship between Sentence 1 and Sentence 2.

Allowed options: A: {option_a_definition}
B: {option_b_definition}
C: {option_c_definition}

Return only one option: A, B, or C.

Sentence 1: {premise}
Sentence 2: {hypothesis}
Option:

Chinese system message:

你是一个句子关系分类器。仅使用句子 1 和句子 2。不要翻译、改写或解释。仅返回一个允许的选项。

Chinese user message:

判断句子 1 和句子 2 之间的关系。

允许的选项：
A：{option_a_definition}
B：{option_b_definition}
C：{option_c_definition}

仅返回一个选项：A、B 或 C。

句子 1：{premise}
句子 2：{hypothesis}
选项：

Model	Input	Default	Alt.	95% CI	Δ	95% CI
Qwen3-8B	ZH	+.085	+.033	[+.008, +.058]	-.052	[-.075, -.028]
Qwen3-8B	EN	+.090	+.043	[+.017, +.070]	-.047	[-.072, -.025]
Qwen3-32B	ZH	-.007	-.018	[-.045, +.007]	-.012	[-.035, +.013]
Qwen3-32B	EN	-.010	-.035	[-.060, -.008]	-.025	[-.043, -.007]
Qwen3-235B	ZH	+.002	+.013	[-.007, +.033]	+.012	[-.005, +.028]
Qwen3-235B	EN	-.008	-.012	[-.028, +.005]	-.003	[-.017, +.010]
Llama-8B	ZH	+.022	+.047	[+.008, +.085]	+.025	[-.013, +.063]
Llama-8B	EN	-.015	+.052	[+.013, +.092]	+.067	[+.037, +.097]
Llama-70B	ZH	+.067	-.018	[-.045, +.008]	-.085	[-.128, -.037]
Llama-70B	EN	+.023	-.010	[-.035, +.015]	-.033	[-.070, +.003]

Table F2: Supplementary alternate-label CN-interface gains. Alt. uses 可推出, 无法确定, and 冲突 in Chinese-instruction semantic-label prompts. Δ is alternate-label gain minus the default semantic-label gain.

Model	Cond.	Default Acc.	Alt. Acc.	ΔAcc.	95% CI
Qwen3-8B	ZH-ZH	.705	.653	-.052	[-.075, -.028]
Qwen3-8B	EN-ZH	.812	.765	-.047	[-.072, -.023]
Qwen3-32B	ZH-ZH	.750	.737	-.013	[-.037, +.012]
Qwen3-32B	EN-ZH	.830	.808	-.022	[-.040, -.003]
Qwen3-235B	ZH-ZH	.792	.802	+.010	[-.005, +.027]
Qwen3-235B	EN-ZH	.863	.863	.000	[-.012, +.012]
Llama-8B	ZH-ZH	.598	.610	+.012	[-.022, +.045]
Llama-8B	EN-ZH	.615	.662	+.047	[+.020, +.075]
Llama-70B	ZH-ZH	.723	.642	-.082	[-.120, -.042]
Llama-70B	EN-ZH	.808	.777	-.032	[-.067, +.002]

Table F3: Direct alternate-label comparisons in Chinese-instruction conditions. ΔAcc. is alternate-label accuracy minus default semantic-label accuracy.

Cond.	Instr./defs	Labels	Acc.	F1	ΔAcc.	95% CI
ZH-ZH	Chinese	English	.595	.588	-.110	[-.143, -.078]
ZH-EN	English	Chinese	.678	.684	+.058	[+.033, +.083]
EN-EN	English	Chinese	.778	.784	+.057	[+.033, +.083]
EN-ZH	Chinese	English	.702	.704	-.110	[-.145, -.075]

Table F4: Qwen3-8B label-language swap. ΔAcc. is swap accuracy minus the original semantic-label baseline for the same condition.

Input	Original	Crossed	95% CI	Δ
ZH	+.085	-.083	[-.115, -.055]	-.168
EN	+.090	-.077	[-.105, -.050]	-.167

Table F5: Qwen3-8B CN-interface gain under crossed label languages. Δ is crossed-label gain minus original semantic-label gain; the corresponding 95% intervals are [-.217, -.118] for ZH input and [-.220, -.115] for EN input.

G.3 Anchored ABC Prompt

English system message:

> You are a sentence relationship classifier. Use only Sentence 1 and Sentence 2. Do not translate, rewrite, or explain them. Return only one allowed option.

English user message:

> Classify the relationship between Sentence 1 and

Cond.	Acc.	95% CI	F1	95% CI	Range	SD	Max share
ZH-ZH	.506	[.486, .525]	.451	[.428, .473]	.218	.095	.663
ZH-EN	.513	[.488, .536]	.474	[.446, .500]	.135	.053	.644
EN-EN	.514	[.494, .535]	.463	[.437, .488]	.215	.095	.693
EN-ZH	.568	[.546, .589]	.525	[.501, .548]	.207	.083	.647

Table F6: Llama-3.1-8B ABC-only boundary check. Values average over all six A/B/C mappings. Max share is the mean largest single-option prediction share across mappings.

Cond.	Semantic Acc.	ABC Acc.	ΔAcc.	95% CI	Semantic F1	ABC F1
ZH-ZH	.598	.506	-.093	[-.128, -.058]	.554	.451
ZH-EN	.577	.513	-.064	[-.095, -.033]	.584	.474
EN-EN	.630	.514	-.116	[-.148, -.084]	.640	.463
EN-ZH	.615	.568	-.047	[-.080, -.015]	.567	.525

Table F7: Llama-3.1-8B ABC-only shift from the semantic-label baseline. ABC values average over all six mappings.

Input	Semantic	ABC	95% CI	Δ
ZH	+.022	-.007	[-.024, +.011]	-.028
EN	-.015	+.054	[+.038, +.069]	+.069

Table F8: Llama-3.1-8B CN-interface gain under ABC-only prompting. Δ is ABC-only residual minus semantic-label gain; the corresponding 95% intervals are [-.076, +.021] for ZH input and [+.024, +.112] for EN input.

Sentence 2.

Allowed options: A ({option_a_anchor}): {option_a_definition}
B ({option_b_anchor}): {option_b_definition}
C ({option_c_anchor}): {option_c_definition}

Return only one option: A, B, or C.

Sentence 1: {premise}
Sentence 2: {hypothesis}
Option:

Chinese system message:

你是一个句子关系分类器。仅使用句子 1 和句子 2。不要翻译、改写或解释。仅返回一个允许的选项。

Chinese user message:

判断句子 1 和句子 2 之间的关系。

允许的选项：
A（{option_a_anchor}）：{option_a_definition}
B（{option_b_anchor}）：{option_b_definition}
C（{option_c_anchor}）：{option_c_definition}

仅返回一个选项：A、B 或 C。

句子 1：{premise}
句子 2：{hypothesis}
选项：

G.4 Semantic-Output Anchor Prompt

Chinese system message:

你是一个句子关系分类器。仅使用句子 1 和句子 2。不要翻译、改写或解释。仅返回一个允许的标签。

Chinese user message:

判断句子 1 和句子 2 之间的关系。

允许的选项：
A（{option_a_anchor}）：{option_a_definition}
B（{option_b_anchor}）：{option_b_definition}
C（{option_c_anchor}）：{option_c_definition}

仅返回一个标签：蕴含、中立或矛盾。不要返回 A、B 或 C。

句子 1：{premise}
句子 2：{hypothesis}
标签：

G.5 A/B/C Mappings: Table G1

Mapping	A	B	C
abc_m1	entailment	neutral	contradiction
abc_m2	contradiction	entailment	neutral
abc_m3	neutral	contradiction	entailment
abc_m4	entailment	contradiction	neutral
abc_m5	neutral	entailment	contradiction
abc_m6	contradiction	neutral	entailment

Table G1: A/B/C mappings used in diagnostic regimes.