

LLM-as-Annotator at 4B Scale: Strategy-Specific Instruction Tuning for Chinese Metaphor Identification

Yufeng WU

City University of Hong Kong
yufenwu2-c@my.cityu.edu.hk

Abstract

LLMs (large language models) are widely used as sentence- and document-level annotators, but protocol-driven token-level annotation, especially in non-English settings, remains underexplored. We present a systematic prompting study of Qwen3.5-4B, a small open-source Chinese-capable model, on MIPVU (Metaphor Identification Procedure Vrije Universiteit) token-level metaphor identification over the PSU Chinese Metaphor Corpus (PSU CMC). We cross four prompting strategies with five instruction variants, yielding a 20-cell development matrix. Across the four strategies, skill demonstration is largely insensitive to instruction variants (F_1 range = 0.057), while strategies with weaker task anchors gain up to +0.132 F_1 from added MIPVU guidance. We treat anchor strength as a qualitative analytical proxy rather than a causal measurement. A single instruction variant, MIPVU-explicit L1-v3, is dev-best or tied-best on the three weak-anchor strategies. The selected L1 configurations are associated with reduced dev-to-test degradation on three of four strategies. Our best test configuration reaches $F_{1\text{strict}} = 0.423$ on the PSU CMC test set, below the fine-tuned encoder result on the same corpus (0.7281; Wu, 2026) and human IAA (inter-annotator agreement; $\kappa \approx 0.75$). The findings position prompt-engineered small open-source LLMs as candidate first-pass annotation aids for Chinese MIPVU under human verification.

1 Introduction

Token-level metaphor identification under MIPVU (Metaphor Identification Procedure

Vrije Universiteit) (Steen et al., 2010; Pragglejaz Group, 2007; Lu and Wang, 2017) is central to computational metaphor research, but manual annotation is costly and scales poorly. Recent work on LLMs (large language models) as annotators (Tan et al., 2024; Li et al., 2023) raises the question of whether they can support protocol-driven token-level tasks. Existing LLM-based metaphor identification has focused on large proprietary models and English data (Liang et al., 2025), while Chinese-language work has prioritized fine-tuned encoders (Wu, 2026). We ask whether a small, open-source LLM, prompted appropriately, can serve as a candidate first-pass annotation aid for Chinese MIPVU that complements trained annotators in a human-in-the-loop workflow.

We address this question through a systematic prompting study on Qwen3.5-4B, a 4-billion-parameter open-source Chinese-capable model. We design four prompting strategies that differ in the type of in-context anchor they provide, namely zero-shot (no anchor), few-shot (example-based), dictionary-grounded (lexical knowledge), and skill demonstration (procedural), and cross each strategy with five variants of the task instruction (the baseline B0 plus four targeted L1 modifications). We evaluate the resulting $4 \times 5 = 20$ configurations on the PSU Chinese Metaphor Corpus (PSU CMC) (Lu and Wang, 2017) development set of 198 sentences, then validate the dev-best configurations on a held-out test set of 344 sentences.

This paper contributes a 4×5 controlled prompting study documenting that instruction sensitivity varies with in-context anchor type and that one MIPVU-explicit instruction vari-

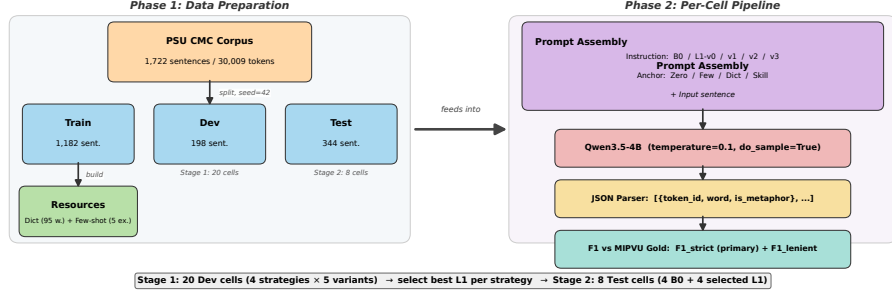


Figure 1: Prompting study pipeline. **Phase 1** splits PSU CMC and builds the 95-word dictionary and five few-shot examples. **Phase 2** combines an instruction variant with a strategy anchor and input sentence; Qwen3.5-4B outputs JSON token-level annotations.

ant performs consistently across weak-anchor strategies in our setting. Our contribution is a deployment-relevant analysis of prompting strategies and instruction sensitivity for protocol-driven metaphor annotation.

Our experiments yield three main findings:

1. **Strategies respond differently to instruction tuning.** Skill demonstration is largely insensitive (F_1 range = 0.057), while weaker anchors gain up to +0.132 F_1 on dev. Across the four evaluated strategies, the qualitative anchor-strength proxy appears to predict instruction sensitivity in this setting.
2. **A single instruction variant performs consistently across weak-anchor strategies.** L1-v3 (MIPVU-explicit), which adds a compact summary of the MIPVU criterion plus three canonical Chinese examples, is dev-best or tied-best on Zero, Few, and Dict. Skill is largely insensitive to instruction choice.
3. **Selected L1 configurations are associated with reduced dev-to-test degradation on three of four strategies.** The selected variants show smaller degradation than B0 for Zero, Few, and Skill. For few-shot, B0 drops -0.093 from dev to test while the selected configuration drops only -0.027 , a 71% reduction. Dict shows the opposite pattern and is reported explicitly in §4.4.

Our best test configuration (L1-v3 + DICT) reaches $F_{1\text{strict}} = 0.423$ on PSU CMC test 344. This remains below fine-tuned encoder performance on the same corpus (Wu, 2026) and below human inter-annotator agreement (Lu and Wang, 2017). We therefore position small open-source LLMs as candidate first-pass annotation aids under human verification, not as replacements for trained MIPVU annotators. Few-shot examples and the skill demonstration appear in Appendices A and B.

2 Related Work

Our work draws on four lines of related research. The first is the use of large language models as data annotators for NLP labeling tasks (§2.1). The second is computational identification of Chinese metaphor, including both fine-tuned and prompting approaches (§2.2). The third is the MIPVU annotation protocol that defines our task (§2.3). The fourth is prompt sensitivity in instruction-tuned LLMs, which motivates our systematic comparison of instruction variants (§2.4).

2.1 LLMs as Annotators

LLMs as data annotators have been surveyed by Tan et al. (2024), with most work focused on sentence- or document-level classification. Li et al. (2023) formalize uncertainty-based human-LLM allocation, and Gligoric et al. (2025) show that unconfident LLM annotations can support confident downstream conclusions;

LLM-based evaluation shows similar task-level variability (Bavaresco et al., 2025). Token- and span-level annotation is less studied: Kasner et al. (2025) report substantial reliability variation across English span tasks, and Jadhav et al. (2025) document degradation in low-resource settings. We contribute a token-level, protocol-driven study on Chinese using a small open-source model.

2.2 Computational Identification of Chinese Metaphor

Computational metaphor identification under MIPVU has been studied extensively in English (e.g., VU Amsterdam Metaphor Corpus), but Chinese-language work is more limited. Lu and Wang (2017) adapt MIPVU to Mandarin Chinese and construct the PSU Chinese Metaphor Corpus (PSU CMC), the primary corpus we use. Tian et al. (2024) explore scaffolding-style prompting for Chinese metaphor reasoning. Concurrent work on the same corpus reports a recent multi-architecture baseline, comparing fine-tuned encoder models (Chinese RoBERTa-wwm-large, MelBERT adapted to Chinese), and a peak token-level F1 of 0.7281 with MelBERT-MIP-only (Wu, 2026). This line of work establishes the fine-tuning reference point on PSU CMC; our contribution is complementary, exploring the prompting frontier rather than fine-tuning.

In a related but distinct English setting, Liang et al. (2025) evaluate GPT-4 zero-shot and few-shot prompting for conventional metaphor detection in English news, reporting zero-shot F1 in the range of 0.23–0.30 depending on subset. We treat this only as background on zero-shot LLM behavior for metaphor-related judgments, not as a directly comparable baseline for Chinese MIPVU.

2.3 MIPVU Protocol and Annotation Reliability

MIPVU (Steen et al., 2010) refines MIP (Pragglejaz Group, 2007) and has been adapted to Chinese (Lu and Wang, 2017). It identifies metaphor via cross-domain comparison: a lexical unit is metaphorically related if its contex-

tual meaning is understood through comparison with a more basic meaning in another conceptual domain. Lu and Wang (2017) report inter-annotator $\kappa \approx 0.75$ on Chinese MIPVU, which we use as a reference point in §5.2.

2.4 Prompt Sensitivity and Optimization

A substantial line of work documents the sensitivity of LLM outputs to prompt formulation. Webson and Pavlick (2022) show that instruction-tuned models often perform surprisingly well even with semantically irrelevant prompts, suggesting that the prompt’s surface form contributes more than its semantic content in certain settings. Min et al. (2022) demonstrate that randomly reassigned labels in few-shot demonstrations have surprisingly limited effect on in-context learning performance, indicating that demonstrations function partly as format anchors rather than semantic exemplars. Sclar et al. (2024) quantify model sensitivity to spurious format features in prompt design. Reynolds and McDonell (2021) provide an early framework for prompt programming, emphasizing the role of instruction design in eliciting target behaviors. Pecher et al. (2026) examines the consequences of prompt underspecification and proposes specification strategies for mitigating it.

We observe an asymmetric sensitivity pattern across our four strategies: stronger anchors (SKILL) absorb instruction-level variation, while weaker anchors (ZERO, DICT) are highly responsive to it. We also identify a single instruction variant (MIPVU-explicit protocol summary) that improves performance on weak-anchor strategies in our setting, providing a concrete starting point rather than a generic sensitivity observation.

3 Method

We frame Chinese metaphor identification as token-level annotation and evaluate a small open-source language model under four prompting strategies crossed with five instruction variants. Figure 1 summarizes the pipeline.

3.1 Task and Dataset

We use the PSU Chinese Metaphor Corpus (PSU CMC) (Lu and Wang, 2017), a token-level Chinese corpus annotated following the MIPVU protocol (Steen et al., 2010). The corpus contains approximately 30,009 tokens across 1,722 sentences from three registers: academic prose, fiction, and news. The overall metaphor density is 10.9%, consistent with the natural distribution of metaphorically related words (MRW) in running Chinese text.

We use a fixed file-level split to ensure that no source document overlaps across splits, yielding 1,182 training, 198 development, and 344 test sentences. The dev and test splits show comparable but non-identical genre distributions, with test containing slightly more fiction (+10.3%) and academic (+6.2%) and less news (-16.5%) than dev.

The task is to predict, for each token in a sentence, whether it is metaphorically related (**true**) or not (**false**) according to MIPVU. Punctuation, function words, and certain proper nouns are excluded from annotation following the original MIPVU exclusion list.

3.2 Model

We use the Qwen/Qwen3.5-4B checkpoint (Yang et al., 2025), a 4B-class Qwen model with strong Chinese capability. Using a small open-source model isolates prompt-strategy effects from raw scale. Inference uses low-temperature sampled decoding (`temperature=0.1`, `do_sample=True`), with one retry at `temperature=0.3` on parse failure. Each cell uses one inference run without an explicitly controlled inference random state.

3.3 Prompting Framework

We design four prompting strategies that differ in the type of in-context anchor provided to the model. We operationalize anchor strength as a qualitative proxy based on the amount and specificity of task-relevant information supplied before inference. This proxy is an analytical framework rather than a quantitative metric or causal measurement. Appendix C summarizes the information supplied by each strategy.

ZERO (zero-shot) gives only the task instruction and input sentence. FEW (few-shot) appends five worked examples from training, chosen to cover verbal, nominal, prepositional, and adjectival metaphor types at moderate sentence length; their mean metaphor density of $\sim 25\%$ is deliberately higher than the corpus mean. DICT (dictionary-grounded) appends a 95-word list of metaphor-prone tokens extracted from training (metaphor rate ≥ 0.5 , count ≥ 5), and instructs the model to attend to these while still applying MIPVU to all other tokens. SKILL (skill demonstration) appends one detailed worked example tracing the MIPVU reasoning end-to-end: exclusion list, basic-meaning lookup, contextual-meaning comparison, and cross-domain mapping judgment.

All four strategies share a common output schema: a JSON array where each element is an object `{token_id, word, is_metaphor}`, one entry per input token. This schema is the same across all conditions and variants.

3.4 Instruction Variants

To probe how the base instruction (i.e., the task description prepended to all four strategies) interacts with each strategy, we design five variants of the base instruction. The baseline (B0) is the full instruction used in pilot experiments; the four L1 variants modify B0 along distinct dimensions:

- **B0 (Baseline, 1,091 chars)** contains task definition, output schema, exclusion list, and basic protocol description.
- **L1-v0 (Minimal, 479 chars)** reduces B0 by approximately 56%, retaining only task definition, output schema, and exclusion list. It tests the hypothesis that much of the baseline instruction is redundant when anchors are strong (Webson and Pavlick, 2022).
- **L1-v1 (Type Specification, +170 chars)** adds an explicit enumeration of four metaphor types (verbal, nominal, prepositional/locative, adjectival) and probes

whether type-level specification improves the model’s coverage (Pecher et al., 2026).

- **L1-v2 (Bias Correction, +145 chars)** adds an explicit instruction to counteract conservative labeling (target density 8-15%, “when in doubt, label true”), motivated by the observation that LLMs tend to under-label minority classes (Reynolds and McDonnell, 2021).
- **L1-v3 (MIPVU Explicit, +120 chars)** adds a compact summary of the MIPVU judgment criterion (cross-domain mapping from a more basic meaning to a contextual meaning) along with three canonical Chinese examples (在, 发展, 方面). It tests whether explicit protocol guidance helps weak-anchor strategies (Steen et al., 2010).

We crossed the four strategies with the five instruction variants, yielding a $4 \times 5 = 20$ -cell design space. Due to length constraints in the few-shot and skill conditions, all variants satisfy the constraint that the rendered system prompt remains below 2,900 characters across all conditions.

3.5 Evaluation Metrics

We use two token-level F1 variants. $F1_{\text{strict}}$ is computed over all sentences in the evaluation split, with unparseable model outputs contributing as false negatives (all gold metaphor tokens in unparseable sentences count as missed); this reflects end-to-end performance. $F1_{\text{lenient}}$ is computed only over sentences with successfully parsed output, reflecting classification capability conditional on parsing success.

We additionally report `parse_rate` (proportion of sentences with valid JSON output) and `metaphor_ratio` (predicted metaphor proportion, as a sanity check against the gold density of $\sim 10\%$).

Across all 28 main cells, `parse_rate` ≥ 0.97 , so $F1_{\text{strict}}$ and $F1_{\text{lenient}}$ differ by less than 0.01. We therefore report $F1_{\text{strict}}$ as the primary metric throughout the paper.

3.6 Experimental Protocol

We adopt a two-stage evaluation protocol to separate model development from final assessment:

Stage 1: Development on Dev 198 We run all $4 \times 5 = 20$ cells (4 strategies $\times$ 5 instruction variants) on the development set of 198 sentences. For each strategy, we select the best-performing instruction variant on dev as the configuration to test. We adopt a 0.01 $F1_{\text{strict}}$ tie threshold and prefer cross-strategy consistency in selection. We acknowledge that this rule favors unified recommendations over per-strategy heterogeneity; selections without tie-breaking are reported in §4.3 for transparency.

Stage 2: Final evaluation on Test 344 We run two cells per strategy on the test set of 344 sentences: the baseline B0 prompt and the dev-selected best L1 variant. This yields $4 \times 2 = 8$ test cells, allowing us to report paired comparisons (B0 vs. selected variant) for each strategy. We do not select on test, and we do not iterate on the prompt after observing test results.

4 Results

We report baseline development performance, the full 4×5 matrix, automatic variant selection, and final paired test comparisons. The final subsection also summarizes the observed parsing and label-density patterns.

4.1 Baseline Performance on Dev

Table 1 (top row, B0) shows baseline $F1_{\text{strict}}$ of 0.372 (Skill) $>$ 0.341 (Dict) $>$ 0.277 (Few) $>$ 0.201 (Zero). This ordering is consistent with the qualitative anchor proxy defined in §3.3. All baselines parse ≥ 0.97 of sentences.

4.2 Instruction Variants on Dev

Figure 2 visualizes the pattern at a glance: instruction effects are largest for weak-anchor strategies and smallest for skill demonstration. Table 1 presents the full 4×5 matrix of strategies $\times$ instruction variants on Dev 198. Several patterns emerge:

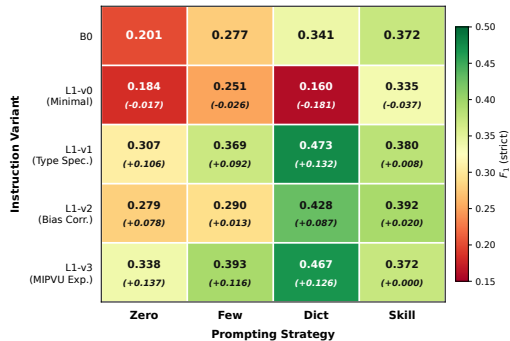


Figure 2: Dev $F1_{\text{strict}}$ across instruction variants and strategies. Skill varies least, while L1-v3 is best or tied-best on three strategies.

Dictionary-grounded benefits most from added instruction content. The three L1 variants that add content (v1, v2, v3) all substantially improve over B0 on DICT (+0.087 to +0.132 absolute F1). L1-v1 (Type Specification) is the strict dev-best dictionary-grounded variant at 0.473; L1-v3 (MIPVU explicit) is nearly tied at 0.467, 0.006 below L1-v1 and within our 0.01 tie threshold. L1-v0 (minimal instruction), in contrast, collapses on DICT (-0.181), indicating that the dictionary alone, without the surrounding protocol description, fails to drive accurate annotation.

Predicted metaphor ratio locates the collapse in under-labeling. Under L1-v0 + DICT, the model predicts metaphor for only 1.26% of tokens, below L1-v0 + ZERO on dev (2.21%). By comparison, B0 + DICT predicts 4.36% and L1-v3 + DICT predicts 8.66% on the same dev split.

Skill demonstration is largely insensitive to L1. On SKILL, F1 across the five variants spans only [0.335, 0.392] (range = 0.057). The maximum gain over B0 is +0.020 (L1-v2), the smallest range and maximum gain among the four strategies.

Few-shot and zero-shot benefit asymmetrically. For FEW, L1-v0 (minimal) yields a small *decrease* (-0.026), but adding content (L1-v1, v2, v3) improves F1 by +0.092 to +0.116, with L1-v3 best. For ZERO, L1-v0 also shows a small drop (-0.017), and v1/v2/v3 again improve, with L1-v3 best at 0.338 (+0.137

over B0). Both strategies are positively responsive to added instruction content but slightly suffer from over-aggressive reduction.

Across-strategy ranking of L1-v3. L1-v3 is the best or tied-best variant on three out of four strategies (ZERO, FEW, DICT) and matches B0 on SKILL (0.372 each). It therefore has the strongest cross-strategy rank in the evaluated matrix.

4.3 Automatic Variant Selection

After applying the cross-strategy tie-breaking rule described in §3.6, we select one L1 variant for each strategy. The resulting selections are:

- ZERO → L1-v3 (F1 = 0.338, Δ vs. B0 = +0.137)
- FEW → L1-v3 (F1 = 0.393, Δ = +0.116)
- DICT → L1-v3 (F1 = 0.467, Δ = +0.126)
- SKILL → L1-v2 (F1 = 0.392, Δ = +0.020)

On DICT, L1-v3 (0.467) and L1-v1 (0.473) differ by only 0.006, below our tie threshold of 0.01. Without tie-breaking, the strict dev-best selection on DICT would be L1-v1 rather than L1-v3. We resolve this tie in favor of L1-v3 to maintain a single recommended variant across the three weak-anchor strategies, while acknowledging that this choice slightly favors cross-strategy consistency over strict per-strategy selection. The selection does not look at test data.

4.4 Test Set Results: Paired Comparisons

Table 3 reports the paired baseline (B0) versus dev-selected L1 variant on Test 344, with 95% bootstrap CIs. All four strategies improve under L1, with effects ranging from +0.060 (SKILL) to +0.182 (FEW).

L1 effects on test are largest for weak anchors. For weak-anchor strategies (ZERO, FEW), the L1 effect on test is *larger* than on dev: FEW improves by +0.182 on test versus +0.116 on dev; ZERO improves by +0.171 on test versus +0.137 on dev. For DICT, the effect is similar across splits (+0.106 on test, +0.126

Table 1: Dev 198 $F1_{\text{strict}}$ matrix. Best per strategy in **bold**; deltas are relative to B0. For DICT, L1-v1 is the strict dev-best variant; L1-v3 is selected under the 0.01 tie threshold for cross-strategy consistency.

Variant	Zero	Few	Dict	Skill
B0	0.201	0.277	0.341	0.372
L1-v0 (Minimal)	0.184 (−0.017)	0.251 (−0.026)	0.160 (−0.181)	0.335 (−0.037)
L1-v1 (Type Spec.)	0.307 (+0.106)	0.369 (+0.092)	0.473 (+0.132)	0.380 (+0.008)
L1-v2 (Bias Corr.)	0.279 (+0.078)	0.290 (+0.013)	0.428 (+0.087)	0.392 (+0.020)
L1-v3 (MIPVU Exp.)	0.338 (+0.137)	0.393 (+0.116)	0.467 (+0.126)	0.372 (+0.000)

Table 2: Strict dev-best selections without tie-breaking. Test results are shown where the strict dev-best cell was run in Stage 2.

Strategy	Strict dev-best	Dev F_1	Test F_1
Zero	L1-v3	0.338	0.325
Few	L1-v3	0.393	0.366
Dict	L1-v1	0.473	not run
Skill	L1-v2	0.392	0.400

on dev). For SKILL, the effect is small but consistent (+0.060 on test, +0.020 on dev).

Baseline B0 usually degrades more on test than the selected variant. Comparing the test-dev gap, B0 + FEW drops by −0.093 from dev to test, while L1-v3 + FEW drops by only −0.027. Table 4 makes the full pattern explicit. The selected L1 configurations are associated with reduced dev-to-test degradation for Zero, Few, and Skill, while DICT shows the opposite pattern.

Best overall configuration on test. L1-v3 + DICT achieves the highest test $F1$ of 0.423, exceeding all other paired configurations by at least 0.023 absolute. We therefore treat it as the strongest candidate configuration in the evaluated matrix.

Effect size and paired uncertainty. We compute bootstrap 95% confidence intervals for each test cell’s $F1_{\text{strict}}$ (1000 resamples) and paired McNemar tests on token-level correctness. Bootstrap CIs in Table 3 show clear separation for ZERO and FEW, with slight overlap for DICT and SKILL. McNemar tests indicate a significant paired shift for ZERO ($\chi^2 = 5.84$, $p = .0157$) and SKILL ($\chi^2 = 8.12$, $p = .0044$), while FEW and DICT are not significant under

Table 3: Test 344 paired comparison with 95% bootstrap CIs (1000 resamples). B0 column shows baseline; L1 column shows dev-selected variant per strategy. For DICT, L1-v3 is the cross-strategy tie-selected configuration.

Strategy	B0 F_1 [CI]	L1 variant (F_1 [CI])	Δ
Zero	0.154 [.123, .188]	L1-v3 (0.325 [.291, .360])	+0.171
Few	0.184 [.150, .220]	L1-v3 (0.366 [.332, .399])	+0.182
Dict	0.317 [.275, .354]	L1-v3 (0.423 [.383, .458])	+0.106
Skill	0.340 [.304, .376]	L1-v2 (0.400 [.367, .434])	+0.060

token-level correctness.

$F1$ improvement and token-level paired correctness assess different aspects of prediction quality. $F1$ is sensitive to changes in minority-class recall, whereas McNemar’s test evaluates the balance of paired correctness disagreements across tokens and may not detect every improvement in class-distribution performance. The stability patterns from the original single-run analysis should be interpreted as empirical tendencies. Bootstrap confidence intervals and paired tests provide uncertainty estimates over evaluation samples but do not replace repeated stochastic inference across the full matrix.

Targeted decoding-variance replication.

Because the practical recommendation centers on L1-v3 + DICT, we repeated the principal DICT comparison with three explicit decoding seeds (42, 13, and 2026). With the checkpoint, prompts, parser, evaluator, and decoding settings fixed, B0 + DICT yielded $F1_{\text{strict}} = 0.314 \pm 0.006$, while L1-v3 + DICT yielded 0.427 ± 0.003 (sample standard deviations). L1-v3 exceeded B0 under all three seeds, supporting the main comparison in Table 3. Appendix D reports the per-seed results.

Table 4: Dev-to-test $F1_{\text{strict}}$ degradation. Negative values indicate lower test than dev; positive values indicate test exceeding dev.

Strategy	B0 gap	L1 gap	Reduction
Zero	−0.047	−0.013	72%
Few	−0.093	−0.027	71%
Dict	−0.024	−0.044	−83%
Skill	−0.032	+0.008	125%

Failure modes and label-density patterns.

Across the eight test cells, parse failures are rare and the most visible annotation differences concern predicted label density.

Across test cells, parse failures are rare (1–2 of 344 sentences; parse rate 0.994–0.997) and involve extra tokens beyond the JSON schema rather than formatting collapse.

B0 on weak-anchor strategies (ZERO, FEW) under-labels metaphors. B0 + FEW predicts a 4.33% metaphor ratio versus 11.85% in the gold labels on test. L1-v3 + DICT predicts 8.32%, closer to the gold ratio but still conservative.

SKILL predictions stay closest to gold density across B0 and L1-v2 (within $\pm 1\%$).

The most distinctive density shift is L1-v0 + DICT. Its predicted metaphor ratio is only 1.26% on dev, compared with 2.21% for L1-v0 + ZERO, 4.36% for B0 + DICT, and 8.66% for L1-v3 + DICT. The cell parses successfully, so the shift occurs in its label assignments rather than its output format.

5 Discussion

The matrix identifies a strategy-specific relation between prompt anchors and instruction variants. We interpret that pattern, state its practical scope, and contextualize the result.

5.1 Anchor Information Appears to Predict Instruction Sensitivity in This Setting

The qualitative anchor proxy in §3.3 provides a lens for the 20-cell matrix. SKILL, which supplies a procedural demonstration, varies only within [0.335, 0.392], while the weaker anchors show larger gains under content-adding L1 variants. Anchor strength thus appears to predict

instruction sensitivity in this setting, consistent with evidence that in-context anchors can absorb part of an instruction’s task-defining role (Webson and Pavlick, 2022; Min et al., 2022).

The selected L1 configurations are also associated with reduced dev-to-test degradation on three of four strategies. For FEW, the gap changes from −0.093 under B0 to −0.027 under L1-v3, a 71% reduction; ZERO and SKILL show the same direction. DICT is the exception (−0.024 vs. −0.044), so the association is not uniform.

5.2 L1-v3 as a First-Pass Annotation Starting Point

For Qwen3.5-4B in the evaluated setting, L1-v3 is a practical weak-anchor starting point. With a metaphor-prone word list, L1-v3 + DICT yields the highest test F1 of 0.423; without one, SKILL + L1-v2 reaches 0.400. The shared JSON schema remains consistently parseable across test cells.

The best configuration remains below human agreement ($\kappa \approx 0.75$; Lu and Wang, 2017) and the fine-tuned encoder result on the same corpus (Wu, 2026). We therefore recommend it only as a first-pass aid checked by trained annotators, not as a replacement for human MIPVU annotation. Its predicted metaphor ratio is also closer to the gold density than those of the B0 weak-anchor configurations.

5.3 Lexical Anchors Require Protocol Scaffolding

The L1-v0 + DICT collapse shows that lexical information is not automatically helpful: predicted metaphor density falls to 1.26%. One plausible interpretation is that, without protocol detail, the 95-word list functions like an eligibility whitelist. Under B0 and content-adding variants, it co-occurs with much higher density and F1. Lexical entries should therefore be framed as task-relevant evidence, not as the annotation criterion itself.

5.4 Comparison to External Reference Points

These reference points contextualize scale rather than establish a state-of-the-art comparison. Liang et al. (2025) report GPT-4 zero-shot F_1 of 0.23–0.30 in a distinct English-language setting. On PSU CMC, Wu (2026) report fine-tuned encoder F_1 of 0.7281, above our prompting result. Our contribution instead characterizes prompt-strategy and instruction interactions for a small open-source first-pass annotator.

6 Limitations

Our study has several limitations.

Single model. We evaluate only Qwen3.5-4B. Models from other families or at other scales may exhibit different instruction sensitivity and different responses to the SKILL anchor. The practical recipe is therefore established only for the evaluated model and setting.

Single dataset. We evaluate only the PSU CMC corpus (Lu and Wang, 2017). Other Chinese metaphor resources use different annotation schemas and text distributions. The evidence supports MIPVU-style token-level annotation on PSU CMC rather than a claim of cross-dataset generalization.

Single inference run per cell. The exploratory 4×5 development conditions were each evaluated once under low-temperature sampled decoding. We added a targeted three-seed replication for the principal DICT comparison, in which L1-v3 outperformed B0 under all three seeds. This provides supporting evidence for the main DICT result but does not replace a full multi-seed evaluation of the complete matrix. Bootstrap confidence intervals and paired tests quantify uncertainty over evaluation samples, while repeated stochastic inference across every condition would still be required to estimate run-level variability throughout the study.

Anchor strength as a qualitative proxy. We operationalize anchor strength as a qualitative proxy based on the amount and specificity

of task-relevant information supplied before inference. This framework supports a descriptive comparison across the four strategies but is not a quantitative metric or causal measurement. Future work could develop interpretable proxies grounded in prompt content or observable model behavior, then validate them across a finer-grained set of anchor manipulations.

Dictionary tie-breaking and unrun alternative. For DICT, L1-v1 is the strict dev-best variant (0.473), while L1-v3 is selected under the 0.01 tie threshold for cross-strategy consistency (0.467). We did not run L1-v1 on test. L1-v3 is therefore recommended as a unified weak-anchor starting point rather than as evidence that it is strictly optimal for dictionary grounding.

No human evaluation of the annotation workflow. The first-pass, human-verify workflow is motivated by the observed error structure but has not been evaluated in an annotation study with trained annotators. Direct measurement of verification time, corrected label quality, and inter-annotator agreement remains necessary to quantify workflow value.

7 Conclusion

Across a 4×5 prompting design on Qwen3.5-4B, weak-anchor strategies gain from MIPVU guidance, while skill demonstration is largely insensitive to instruction choice. Anchor strength, treated as a qualitative proxy, appears to predict instruction sensitivity in this setting. L1-v3 is a useful weak-anchor starting point and is associated with reduced dev-to-test degradation for three of four strategies.

The best configuration remains a candidate pre-annotation setting under human verification, not a replacement for trained annotators. The released prompts and evaluation artifacts support replication of this prompt-strategy study.

Ethics Statement

We use the PSU Chinese Metaphor Corpus under its original license and do not redistribute corpus text in our release package. Qwen3.5-4B

is an open-source model used under its published license; we use the released checkpoint for inference only and do not perform additional training. We position LLM output as a first-pass annotation aid intended to reduce, rather than replace, the work of trained human annotators, and we do not claim that model outputs match human annotation quality. The corpus consists of published Chinese text from academic, fiction, and news registers, and we do not process personal or identifying information. For reproducibility, we will release prompt templates and evaluation scripts, while corpus files and model weights should be obtained directly from their original distributors.

References

- Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, Andre Martins, Philipp Mondorf, Vera Neplenbroek, Sandro Pezzelle, Barbara Plank, David Schlangen, Alessandro Suglia, Aditya K. Surikuchi, Ece Takmaz, and Alberto Testoni. 2025. [LLMs instead of human judges? a large scale empirical study across 20 NLP evaluation tasks](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 238–255, Vienna, Austria. Association for Computational Linguistics.
- Kristina Gligoric, Tijana Zrnic, Cino Lee, Emmanuel J. Candès, and Dan Jurafsky. 2025. [Can unconfident LLM annotations be used for confident conclusions?](#) In *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3514–3533. Association for Computational Linguistics.
- Suramya Jadhav, Abhay Shanbhag, Amogh Thakurdesai, Ridhima Sinare, and Raviraj Joshi. 2025. [On limitations of LLM as annotator for low resource languages](#). In *Proceedings of the 8th International Conference on Natural Language and Speech Processing (ICNLSP-2025)*, pages 277–282, Southern Denmark University, Odense, Denmark. Association for Computational Linguistics.
- Zdeněk Kasner, Vilém Zouhar, Patrícia Schmidtová, Ivan Kartáč, Kristýna Onderková, Ondřej Plátek, Dimitra Gkatzia, Saad Mahamood, Ondřej Dušek, and Simone Balloccu. 2025. [Large language models as span annotators](#). *arXiv preprint arXiv:2504.08697*.
- Minzhi Li, Taiwei Shi, Caleb Ziems, Min-Yen Kan, Nancy F. Chen, Zhengyuan Liu, and Diyi Yang. 2023. [CoAnnotating: Uncertainty-guided work allocation between human and large language models for data annotation](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1487–1505, Singapore. Association for Computational Linguistics.
- Jiahui Liang, Aletta G. Dorst, Jelena Prokic, and Stephan Raaijmakers. 2025. [Using GPT-4 for conventional metaphor detection in english news texts](#). *Computational Linguistics in the Netherlands Journal*, 14:307–341.
- Xiaofei Lu and Ben Pin-Yun Wang. 2017. [Towards a metaphor-annotated corpus of mandarin chinese](#). *Language Resources and Evaluation*, 51(3):663–694.
- Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. [Rethinking the role of demonstrations: What makes in-context learning work?](#) In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Branislav Pecher, Michal Spiegel, Robert Belanec, and Jan Cegin. 2026. [Revisiting prompt sensitivity in large language models for text classification: The role of prompt underspecification](#). *arXiv preprint arXiv:2602.04297*.
- Pragglejaz Group. 2007. [MIP: A method for identifying metaphorically used words in discourse](#). *Metaphor and Symbol*, 22(1):1–39.
- Laria Reynolds and Kyle McDonell. 2021. [Prompt programming for large language models: Beyond the few-shot paradigm](#). In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 314:1–314:7, New York, NY, USA. Association for Computing Machinery.
- Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2024. [Quantifying language models’ sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting](#). In *Proceedings of the Twelfth International Conference on Learning Representations*. OpenReview.net. ICLR 2024.

Gerard J. Steen, Aletta G. Dorst, J. Berenike Herermann, Anna A. Kaal, Tina Krennmayr, and Trijntje Pasma. 2010. *A Method for Linguistic Metaphor Identification: From MIP to MIPVU*, volume 14 of *Converging Evidence in Language and Communication Research*. John Benjamins Publishing Company, Amsterdam/Philadelphia.

Zhen Tan, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Mansoor Karami, Jundong Li, Lu Cheng, and Huan Liu. 2024. *Large language models for data annotation and synthesis: A survey*. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 930–957, Miami, Florida, USA. Association for Computational Linguistics.

Yuan Tian, Nan Xu, and Wenji Mao. 2024. *A theory guided scaffolding instruction framework for LLM-enabled metaphor reasoning*. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7738–7755, Mexico City, Mexico. Association for Computational Linguistics.

Albert Webson and Ellie Pavlick. 2022. *Do prompt-based models really understand the meaning of their prompts?* In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2300–2344, Seattle, United States. Association for Computational Linguistics.

Yufeng Wu. 2026. A reproducible multi-architecture baseline for token-level chinese metaphor identification under the mipvu framework. Manuscript.

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. 2025. *Qwen3 technical report*. *arXiv preprint arXiv:2505.09388*.

Appendices

A Few-Shot Examples

The five examples below are reproduced verbatim from the stored few-shot prompt component.

示例 1(中等长度多触发)
句子: 那就不谈这件事了

```
[{"token_id":0,"word":"那","is_metaphor":false}, {"token_id":1,"word":"就","is_metaphor":false}, {"token_id":2,"word":":","is_metaphor":false}, {"token_id":3,"word":"谈","is_metaphor":false}, {"token_id":4,"word":"这","is_metaphor":true}, {"token_id":5,"word":"件","is_metaphor":true}, {"token_id":6,"word":"事","is_metaphor":false}, {"token_id":7,"word":"了","is_metaphor":false}]
```

示例 2(动词隐喻)

句子: 我怎么那么想攀附你

```
[{"token_id":0,"word":"我","is_metaphor":false}, {"token_id":1,"word":"怎么","is_metaphor":false}, {"token_id":2,"word":"那么","is_metaphor":false}, {"token_id":3,"word":"想","is_metaphor":false}, {"token_id":4,"word":"攀附","is_metaphor":true}, {"token_id":5,"word":"你","is_metaphor":false}]
```

示例 3(名词隐喻)

句子: 原来男人还有这么个世界

```
[{"token_id":0,"word":"原来","is_metaphor":false}, {"token_id":1,"word":"男人","is_metaphor":false}, {"token_id":2,"word":"还有","is_metaphor":false}, {"token_id":3,"word":"这么","is_metaphor":false}, {"token_id":4,"word":"个","is_metaphor":true}, {"token_id":5,"word":"世界","is_metaphor":true}]
```

示例 4(形容词或介词隐喻)

句子: 在屋檐下不得不低头

```
[{"token_id":0,"word":"在","is_metaphor":true}, {"token_id":1,"word":"人","is_metaphor":false}, {"token_id":2,"word":"屋檐","is_metaphor":true}, {"token_id":3,"word":"下","is_metaphor":true}, {"token_id":4,"word":"不得不","is_metaphor":false}, {"token_id":5,"word":"低头","is_metaphor":true}]
```

示例 5(全负例)

句子: 孩他爹你说呢

```
[{"token_id":0,"word":"孩","is_metaphor":false}, {"token_id":1,"word":"他","is_metaphor":false}, {"token_id":2,"word":"爹","is_metaphor":false}, {"token_id":3,"word":"你","is_metaphor":false}, {"token_id":4,"word":"说","is_metaphor":false}, {"token_id":5,"word":"呢","is_metaphor":false}]
```

B Skill Demonstration

The procedural demonstration below is reproduced verbatim from the stored SKILL prompt component.

Skill 示范: 逐 token 推理过程

示范句子

" 它们在发展国民经济和保障人民生活方面有着重要作用"

内部推理 (供你学习, 正式输出不要包含此段)

排除项处理 (直接 false):

" 它们"(代词)、" 和"(连词)、" 着"(助词) -> false

MIPVU 跨域映射 (标 true):

"在": 基本义=空间位置("书在桌上");

语境义=抽象领域内("在...方面")。

空间-> 抽象领域, 跨域映射 -> true

"发展": 基本义 = 生长、伸展; 语境义 = 社会经济进步。

生长-> 抽象过程, 跨域映射 -> true

"方面": 基本义 = 物体的某个面; 语境义 = 问题的某个维度。

空间面-> 抽象维度, 跨域映射 -> true

语境义 = 基本义 (标 false):

" 国民经济"(国家经济系统)、" 保障"(保护使其不受损害)、

" 人民"(社会成员)、" 生活"(日常状态)、" 有"(具有)、

"重要"(关键)、"作用"(功能影响) -> 各词语语境即基本义,无跨域 ->
 ↪ false

正式输出格式 (这是你应该输出的)

```
[{"token_id":0,"word":"它们","is_metaphor":false},
{"token_id":1,"word":"在","is_metaphor":true},
{"token_id":2,"word":"发展","is_metaphor":true},
{"token_id":3,"word":"国民经济","is_metaphor":false},
{"token_id":4,"word":"和","is_metaphor":false},
{"token_id":5,"word":"保障","is_metaphor":false},
{"token_id":6,"word":"人民","is_metaphor":false},
{"token_id":7,"word":"生活","is_metaphor":false},
{"token_id":8,"word":"方面","is_metaphor":true},
{"token_id":9,"word":"有","is_metaphor":false},
{"token_id":10,"word":"着","is_metaphor":false},
{"token_id":11,"word":"重要","is_metaphor":false},
{"token_id":12,"word":"作用","is_metaphor":false}]
```

关键提示

按这种思路判定测试句子,只输出最终 JSON 数组,不要输出推理过程。

按 MIPVU 协议:语法化的介词("在"/"为"/"于"等)、

↪ 方位词("方面"/"上面"/"中"等)

若有具象空间基本义,在抽象语境中仍标 true(conventional metaphor)

↪ 也算 metaphor)。

C Additional Analysis

C.1 Anchor information

Table 5 makes the qualitative operationalization in §3.3 explicit. The ordering is an analytical

framework based on the amount and specificity of task-relevant information supplied before inference, not a quantitative metric or causal measurement.

Strategy	Anchor information provided
Zero	No external anchor
Few	Five annotated examples
Dict	95 metaphor-prone lexical entries
Skill	One procedural MIPVU reasoning demonstration

Table 5: Qualitative anchor information supplied before inference.

C.2 Descriptive post-hoc genre breakdown

Configuration	Genre	Sent.	Tokens	Gold %	Pred. %	P	R	F_1
B0 + Dict	Academic	117	2,044	17.07	4.79	.643	.181	.282
B0 + Dict	Fiction	131	1,838	9.74	5.66	.577	.335	.424
B0 + Dict	News	96	1,704	7.86	2.58	.477	.157	.236
L1-v3 + Dict	Academic	117	2,044	17.07	9.93	.631	.367	.464
L1-v3 + Dict	Fiction	131	1,838	9.74	9.09	.443	.413	.428
L1-v3 + Dict	News	96	1,704	7.86	5.52	.383	.269	.316

Table 6: Descriptive genre-level performance from existing test predictions.

Table 6 pools token-level predictions within Academic, Fiction, and News for B0 + DICT and L1-v3 + DICT. It uses only the existing test predictions. The breakdown is post hoc and descriptive, includes no significance testing, and does not support a claim of genre generalization.

D Targeted Decoding Variance Validation

We repeated only the principal DICT comparison under three explicit decoding seeds, keep-

ing the Qwen3.5-4B checkpoint, Test 344 split, prompts, parser, evaluator, and decoding settings fixed. Table 7 gives the seed-level strict F_1 values. L1-v3 exceeded B0 under every seed, by 0.1058, 0.1090, and 0.1227 absolute F_1 , respectively. Table 8 summarizes the mean and sample standard deviation. This analysis is a targeted replication of the final DICT comparison and does not constitute a full multi-seed evaluation of the complete experimental matrix.

Configuration	Seed	$F_{1\text{strict}}$
B0 + Dict	42	0.3185
B0 + Dict	13	0.3160
B0 + Dict	2026	0.3075
L1-v3 + Dict	42	0.4243
L1-v3 + Dict	13	0.4250
L1-v3 + Dict	2026	0.4302

Table 7: Seed-level targeted replication results.

Configuration	Mean $F_{1\text{strict}}$	SD
B0 + Dict	0.3140	0.0058
L1-v3 + Dict	0.4265	0.0032

Table 8: Aggregate strict F_1 across the three seeds.