

Beyond Single Gold Labels: Human Variability and Conceptualisation in Chinese Classifier Use

Yongheng Peng, Guanyi Chen*, and Tingting He

Hubei Provincial Key Laboratory of Artificial Intelligence and Smart Learning,
National Language Resources Monitoring and Research Center for Network Media,
School of Computer Science, Central China Normal University
yhpeng669@mails.ccnu.edu.cn, {g.chen, tthe}@ccnu.edu.cn

Abstract

Chinese classifier prediction is often evaluated as a single-label classification problem, assuming one gold classifier for each context. We argue that this assumption overlooks the inherent variability and conceptual nature of classifier use. We propose a controlled experimental framework for studying Chinese classifiers in humans and vision-language models across three tasks: Classifier Choice, Head Noun Completion, and Referring Expression Generation. Based on a human behavioural dataset, we evaluate models against human response distributions rather than single gold labels. Our results show that classifier use is shaped by noun specificity, animacy, and task setting, and that models differ from humans not only in classifier choice but also in classifier-mediated conceptualisation. Although explicitly encoding the behavioural patterns learned from the human experiments into prompts improves model alignment with human behaviour, especially for smaller models, its effect is uneven across tasks: it helps models better match dominant human preferences, but does not enable them to reproduce the full distribution of human variability or fully close the gap in natural generation settings. The dataset collected in this study is publicly available at <https://github.com/pengyh669-oss/classifier-variability>.

1 Introduction

Mandarin Chinese obligatorily requires classifiers in quantified noun phrases. For example, while English allows *three squirrels*, the corresponding Mandarin expression is 三只松鼠(*san zhi song-shu*, “three CL squirrels”), where the classifier 只 is required between the numeral and the noun. The choice of a classifier is closely related to how speakers conceptualise referents, including their shape, animacy, function, and semantic category. As a result, Chinese classifier prediction has attracted

attention in both linguistics (Tai, 1994; Chierchia, 1998; Cheung et al., 2012) and NLP (Peinelt et al., 2017; Järnfors et al., 2021; Zhang et al., 2025), where the task is typically formulated as selecting the appropriate classifier given a head noun and its surrounding context.

Recent work has further examined whether Large Language Models (LLMs) possess knowledge of Chinese classifiers (Zhang et al., 2025). These studies suggest that, despite their strong general linguistic capabilities, LLMs still show substantial limitations in classifier prediction. However, we argue that the observed gap between LLMs and humans may be partly shaped by the evaluation paradigm itself. Existing evaluations usually assume that each context has a single gold-standard classifier and measure model performance using categorical accuracy. This assumption is problematic because classifier use is inherently variable (Zhan and Levy, 2019; Zheng and Liu, 2023). The same referent may license different classifiers, and different speakers may choose different classifiers even under the same noun condition. For example, a visual scene depicting an apple may elicit classifiers such as 个(*ge*), 颗(*ke*), or 枚(*mei*). This suggests that classifier choice should be modelled as a distribution over human responses rather than as a single categorical label.

Beyond variability, classifiers are also closely tied to conceptualisation (Lakoff, 1986; Liang and Liao, 2024). Classifier use is not merely a one-way mapping from noun to classifier. While the way a referent is named may influence classifier choice, the classifier itself may also guide how speakers conceptualise and name the referent. In other words, classifiers may participate in a bidirectional relationship between linguistic form and object conceptualisation.

A further limitation of existing evaluations is ecological validity. Most evaluations rely on fill-in-the-blank tasks, where the head noun is already

*Corresponding Author

given, and the model only needs to supply the classifier. Such settings are useful for controlled evaluation, but they may not fully reflect natural language production, where speakers typically generate the classifier and head noun jointly as part of a referring expression. It therefore remains unclear whether classifier behaviours observed in controlled elicitation tasks generalise to more natural generation settings.

To address these limitations, we propose a controlled experimental framework for studying Chinese classifier use in humans and (Vision) LLMs¹. We design three complementary tasks: (1) **Classifier Choice**, where participants select classifiers given different levels of head noun specificity; (2) **Head Noun Completion**, where participants complete a noun phrase given a classifier; and (3) **Referring Expression Generation**, where participants freely generate referring expressions for visual scenes. These tasks allow us to collect a human behavioural dataset that captures classifier variability, classifier-mediated conceptualisation, and classifier choice in more ecologically valid generation settings. We then ask multiple LLMs to complete the same tasks and compare their outputs against the human response distributions.

We are curious about the following research questions:

- RQ1** How variable is Chinese classifier use across different conceptualisations of the same referent, and to what extent do LLMs reproduce human classifier distributions?
- RQ2** To what extent do classifiers shape conceptualisation in humans and LLMs?
- RQ3** Do classifier behaviours observed in controlled elicitation tasks generalise to natural referring expression generation in humans and LLMs?

2 Experimental Design

2.1 Tasks

To investigate Chinese classifier use in a controlled yet comparable setting, we design three complementary tasks as shown in Figure 1: Classifier Choice, Head Noun Completion, and Referring Expression Generation. All tasks are grounded in the

¹Since our tasks require models to process visual inputs, we use the term LLM to refer to large vision-language models hereafter.

same set of visual scenes and are later administered to both human participants and LLMs.

Task 1: Classifier Choice. The Classifier Choice task examines how head noun specificity affects classifier selection. For each visual scene, participants are shown an image and a sentence containing a missing classifier. The target referent is described using one of three levels of head noun specificity: a superordinate noun, a basic-level noun, or a subordinate noun. For example, the same image may be paired with head nouns such as 衣服(*clothes*), 外套(*coat*), and 西装(*suit*). Participants are asked to complete the sentence by providing an appropriate classifier, as in 这是一__苹果(*There is one __ apple in the picture.*)

This manipulation allows us to test whether different ways of linguistically conceptualising the same referent lead to different classifier choices. It also allows us to capture two types of variability in classifier use: variability across noun specificity levels, and variability across speakers under the same image–noun condition.

Task 2: Head Noun Completion. Follow the task setting of Liang and Liao (2024), the Head Noun Completion task examines whether classifiers influence how speakers conceptualise and name a referent. For each visual scene, participants are shown an image and a sentence containing a given classifier but a missing head noun. The classifier is manipulated to be either a general classifier 个², or a more specific classifier that is compatible with the target referent. For example, participants may see a prompt such as 这是一个__ or 这是一颗__(*There is one CL __ in the picture.*)

This task reverses the direction of the Classifier Choice task. Instead of asking how head nouns affect classifier selection, it asks whether classifiers guide the choice of head nouns. If different classifiers lead participants to produce different nouns for the same referent, this would suggest that classifiers are not merely selected after object naming, but may also shape the level and manner of conceptualisation.

²Although some previous work has treated 只 as a general classifier within the animal domain, the present study adopts a broader cross-domain operationalisation: 个 serves as the general classifier baseline, while classifiers with stronger noun-selection restrictions, including 只, are treated as specific classifiers. This choice allows a uniform comparison across animal and furniture items. Moreover, previous findings suggest that even within the animal domain, 个 and 只 can pattern differently (Zhou, 2025), supporting the need to distinguish them in the present design.

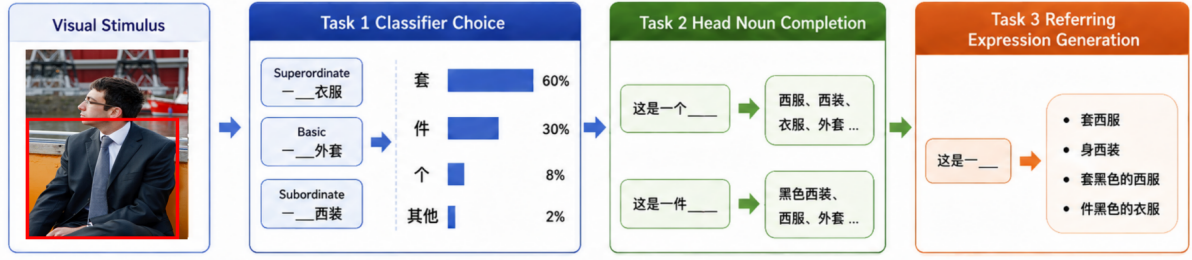


Figure 1: Overview of the three controlled tasks used in our experiment.

Task 3: Referring Expression Generation. The Referring Expression Generation (REG) task evaluates classifier use in a more natural generation setting. Following the task setting of van Deemter et al. (2017), participants are shown a visual scene together with quantity information and are asked to produce a referring expression describing the target referent freely. This task tests whether classifier patterns observed in the controlled Classifier Choice task generalise to more ecologically valid production settings.

2.2 Human Experiment

Materials. The images used in the present study were sampled from the ManyNames ZH dataset (He et al., 2023), which contains 1,319 images from seven domains: HOME, PEOPLE, ANIMAL_PLANTS, CLOTHING, VEHICLE, FOOD, and BUILDINGS. We selected a subset of 60 images as visual stimuli, including 30 images from the ANIMAL_PLANTS and PEOPLE domains depicting animate entities, and 30 images from the remaining domains depicting inanimate entities.³ Images were selected to depict real and unambiguous objects, allowing us to focus on the distinction between animate and inanimate referents without introducing finer-grained category distinctions.

For each image, we generated multiple task-specific conditions, as shown in Figure 1. In the Classifier Choice task, each image was paired with three head nouns corresponding to different levels of specificity: superordinate, basic-level, and subordinate. These nouns were manually selected following the criteria of Liang and Liao (2024). Basic-level nouns denote cognitively salient and commonly used categories in which objects share common features and can be easily recognised, such as 狗(dog) and 桌子(table). Superordinate

nouns represent more general categories encompassing multiple basic-level categories, such as 动物(animal) and 家具(furniture). Subordinate nouns denote more specific categories below the basic level, including subtypes with particular characteristics, such as 吉娃娃(Chihuahua) and 电脑桌(computer desk) for 桌子. We further ensured that the three selected nouns for each image were of the same order of magnitude in frequency based on Google N-gram.

In the Head Noun Completion task, each image was paired with two classifier conditions: the general classifier 个 and a referent-specific classifier, such as 条(tiao), 头(tou), or 朵(duo). The specific classifier was manually selected to be a conventional and highly frequent classifier associated with the target referent, based on Google N-gram. In the REG task, participants were provided only with the image and quantity information, without additional linguistic constraints.

As a result, each image yielded six experimental conditions across the three tasks, resulting in $60 \times 6 = 360$ combinations in total.

Design. The experiment used a list-based within-subjects design. The full set of 360 experimental items was divided into five lists, each containing 72 trials. Each list was balanced such that the six experimental conditions appeared 12 times each, and each image appeared at most twice within a list. Thus, each participant was exposed to all conditions while seeing only a subset of the full stimulus set, reducing repetition and fatigue effects.

To elicit natural Mandarin classifier use, the experiment was conducted as a spoken production task. Participants were instructed to read each prompt and orally produce the expression they would use to complete the blank.

Procedure. Participants were tested individually in a quiet laboratory room. The experiment was

³We balanced animacy because it is a well-established factor affecting referring expression use, although its role in Chinese classifier choice has received less direct attention.

conducted using a custom-built Python interface, as shown in Figure 2 in Appendix A. After providing informed consent, participants received instructions explaining that they would view a series of 72 images and should respond based on the object highlighted by the red window in each image. For each trial, participants pressed a “record” button to start recording their spoken response and clicked “stop” when they had finished.

Participants. Fifty native speakers of Mandarin participated in the experiment. All participants had normal or corrected-to-normal vision and hearing, and reported no history of colour vision deficiency. They received 20 RMB as compensation for their participation.

Data Processing. The final dataset contains $360 \times 10 = 3600$ responses in total. All responses were manually annotated for both classifier and head usage as well as their categories, i.e., classifier type (general and specific) and noun taxonomic level (subordinate, basic, and superordinate).

During annotation, we merged responses representing alternative expressions for the same object into a unified label. For example, responses such as 车和车子(*car*) were both treated as the basic-level noun. Semantically incorrect responses (e.g., calling a table as *a chair*) were not deleted but were manually assigned to the appropriate taxonomic level.

2.3 LLM Experiment

Material and Design. The same stimuli used in the human experiment were also employed in the LLM experiment. To fully stimulate the human experiment, we used a unified prompt (see Appendix F) to instruct LLMs to accomplish all three tasks.

Procedure and Models. We evaluated eight multimodal LLMs through their APIs, including GLM-4.6V (Glm et al., 2024), GPT-5.4 (Achiam et al., 2023), Gemini 3.1 Pro Preview (Comanici et al., 2025), Qwen3.6-Plus (Team, 2026), Claude Sonnet 4.6 (Anthropic, 2024), Llama 4 Scout (Grattafiori et al., 2024), Qwen3-VL-32B-Instruct (Bai et al., 2025a), and Qwen2.5-VL-72B-Instruct (Bai et al., 2025b).

Data Processing. The outputs of the LLMs were processed in the same manner as the human responses. Since the models occasionally produced additional content, repeated parts of the prompt, or

responses that did not strictly follow the required format, all outputs were manually cleaned and annotated by the authors. For each response, we annotated classifier usage and noun taxonomic level, following the same criteria used in the human experiment. Classifier responses were categorised as general classifier, specific classifier, or none. Noun responses were classified as basic-level, superordinate, or subordinate terms.

3 Results

In what follows, we answer the three research questions in the introduction task by task.

3.1 Task 1: Classifier Choice

We use the Classifier Choice task to examine the first research question, focusing on how classifier choice varies across noun specificity levels in humans and LLMs.

3.1.1 Human Findings

To examine how noun specificity and animacy affect classifier choice, we modelled whether participants selected a referent-specific classifier in the Classifier Choice task. Since the classifier 只(*zhi*) was overwhelmingly used for animate referents, we recoded it as a general classifier in the statistical analysis. We then fitted a binomial mixed-effects logistic regression model with classifier type (general vs. specific) as the dependent variable, noun specificity and animacy as fixed effects, and random intercepts for participants and items.

The analysis revealed a strong effect of animacy (see Appendix B.1 for more details). Relative to inanimate referents, animate referents significantly reduced the probability of selecting specialised classifiers, indicating a strong preference for the general classifier in animate contexts ($\beta = -3.81$, $p < .001$). Noun specificity also significantly affected classifier choice. For inanimate referents, both ($\beta = -0.20$, $p = .044$) and basic-level nouns ($\beta = -0.28$, $p = .010$) significantly reduce specialised classifier use compared with superordinate nouns.

Crucially, noun specificity interacted significantly with animacy. Although animate referents strongly favoured general classifiers overall, this tendency was significantly weakened under subordinate ($\beta = 1.07$, $p < .001$) and basic-level noun conditions ($\beta = 1.57$, $p < .001$). In other words, more specific noun conceptualisations increased

Model	Task 1			Task 2			Task 3					
	MRR	Acc	JSD	MRR	Acc	BLEU	CL-MRR	CL-Acc	CL-JSD	N-MRR	N-Acc	BLEU
GLM	0.836	0.917	0.295	0.817	0.667	0.938	0.906	0.950	0.250	0.767	0.617	0.902
GPT	0.816	0.911	0.310	0.843	0.717	0.959	0.710	0.833	0.392	0.842	0.733	0.889
Gemini	0.842	0.928	0.289	0.776	0.650	0.919	0.933	0.950	0.221	0.825	0.717	0.937
Qwen3.6	0.776	0.889	0.346	0.832	0.692	0.937	0.757	0.867	0.345	0.806	0.650	0.885
Claude	0.828	0.928	0.300	0.799	0.650	0.907	0.822	0.867	0.306	0.753	0.600	0.888
Llama-4	0.796	0.650	0.327	0.793	0.642	0.859	0.749	0.533	0.362	0.828	0.750	0.879
Qwen3	0.773	0.628	0.348	0.822	0.708	0.929	0.747	0.633	0.358	0.797	0.667	0.851
Qwen2.5	0.786	0.633	0.329	0.875	0.783	0.919	0.790	0.617	0.326	0.822	0.700	0.891
Oracle	1.000	1.000	0.200	-	-	-	-	-	-	-	-	-
General	0.577	0.511	0.548	-	-	-	-	-	-	-	-	-

Table 1: Overall human–LLM comparison across the three tasks. For Task 3, the prefix CL- denotes classifier-related metrics, and N- denotes noun-related metrics. Best values are shown in bold. ‘Oracle’ and ‘General’ are the oracle and general-classifier baselines for Task 1.

the likelihood of selecting specialised classifiers even for animate referents.

These findings suggest that classifier choice is jointly shaped by animacy and noun specificity, and cannot be reduced to a fixed noun–classifier mapping. Instead, classifier use varies systematically depending on how referents are linguistically conceptualised.

3.1.2 LLM-Human Comparison

Metrics. To compare LLM outputs with human classifier distributions, we employ three metrics (see Appendix C for more details): (1) **Mean Reciprocal Rank (MRR)** measures how highly an LLM-selected classifier is ranked within the human response distribution. (2) **Categorical Accuracy (Acc.)** evaluates whether the classifier category (general vs. specific) produced by the LLM matches the majority human-preferred category. (3) **Jensen–Shannon Divergence (JSD)** measures the distributional divergence between LLM-generated and human classifier distributions.

Rule-based Baselines. To contextualise LLM performance against simple non-parametric approaches, we introduce two rule-based baselines for Task 1. A natural rule-based approach is to construct a lookup dictionary from training data and, for each item, select the classifier that occurs most frequently in the corresponding training responses. Since our setting does not provide a separate training set, we instantiate an oracle baseline using the human responses themselves to construct this dictionary. The oracle therefore always selects the most frequent human-chosen classifier for each

item, providing an approximate upper-bound reference for such a dictionary-based approach. We additionally introduce a general-classifier baseline, which follows a simple animacy-based rule: $\uparrow$ (*ge*) is selected for inanimate referents and 只 (*zhi*) for animate referents.

Comparison. Table 1 reports the overall results, including the two rule-based baselines, and Appendix D provides separate results for animate and inanimate referents. All LLMs consistently outperform the general-classifier baseline across all three metrics, suggesting that their classifier choices capture item-specific preferences beyond a simple animacy-based rule. However, all models remain below the oracle dictionary baseline, indicating that they do not fully recover the dominant classifier preferences observed in human responses. Among the LLMs, GLM and Gemini show the strongest overall alignment with human classifier distributions. Models align better with humans on animate referents than on inanimate referents, while smaller models generally perform worse than larger models across the three metrics.

Behavioural Analysis. To further test whether LLMs reproduce the behavioural patterns observed in humans, we fitted a mixed-effects logistic regression model predicting specialised classifier use. The model included response source, i.e., whether the response was produced by humans or LLMs, noun specificity, animacy, and their interactions with response source. This analysis tests whether LLMs are affected by noun specificity (referred to as H_1) and animacy (H_2) in the same direction as

Model	H_1	H_2	H_3	H_4	H_5	H_6
GLM	✓	✓	✓	✓	✓	×
GPT	✓	✓	✓	✓	✓	✓
Gemini	✓	✓	✓	×	✓	×
Qwen3.6	✓	✓	✓	×	✓	×
Claude	✓	✓	✓	✓	✓	×
Llama-4	✓	✓	×	×	×	×
Qwen3	✓	✓	✓	×	✓	×
Qwen2.5	×	✓	✓	×	✓	×

Table 2: Behavioural consistency analysis across the three tasks. H_1 – H_6 correspond to the behavioural hypotheses tested in the experiments. ✓ indicates that the corresponding behavioural effect in the LLM is consistent with the direction observed in human participants, while × indicates inconsistency with human behaviour.

humans. As shown in Table 2, almost all LLMs exhibit effects in the same direction as humans for both factors. The only exception is Qwen2.5, for which the effect of basic-level nouns relative to superordinate nouns is weaker than that observed in humans.

3.2 Task 2: Head Noun Completion

We use the Head Noun Completion task to examine the second research question, focusing on whether classifier choice shapes how humans and LLMs conceptualise and name the same visual referent.

3.2.1 Human Findings

We analysed how classifier type and animacy influenced the specificity level of the completed nouns in the Head Noun Completion task. We fitted three pairwise binomial mixed-effects logistic regression models comparing basic-level vs. superordinate nouns, subordinate vs. superordinate nouns, and basic-level vs. subordinate nouns. Each model included classifier type, animacy, and their interaction as fixed effects, with random intercepts for participants and items.

The results consistently showed that specialised classifiers encouraged participants to produce more specific nouns relative to superordinate nouns (see more details in Appendix B.2). In the basic-level vs. superordinate comparison, specialised classifiers significantly increased the probability of selecting basic-level nouns ($\beta = 0.62, p < .001$). Similarly, in the subordinate vs. superordinate comparison, specialised classifiers significantly increased subordinate noun selection ($\beta = 1.10, p < .001$). These are in line with the findings in Liang and Liao (2024).

Animacy showed the opposite effect. Animate

referents significantly reduced the probability of selecting more specific nouns relative to superordinate nouns in both the basic-level vs. superordinate comparison ($\beta = -0.45, p = .003$) and the subordinate vs. superordinate comparison ($\beta = -0.50, p = .004$).

These findings suggest that classifiers influence how speakers conceptualise and linguistically categorise referents, supporting the view that classifiers participate in conceptualisation rather than merely reflecting noun semantics.

3.2.2 Human-LLM Comparison

Metrics. To compare LLM outputs with human noun completion behaviour, we employ three metrics: (1) **Categorical Mean Reciprocal Rank (MRR)** measures how highly the noun category selected by the LLM is ranked within the human response distribution. (2) **Categorical Accuracy (Acc.)** evaluates whether the noun category produced by the LLM matches the majority human-preferred noun category. (3) **BLEU** measures the surface-level similarity between the completed expressions generated by LLMs and those produced by human participants.

Comparison. Table 1 reports the overall results. Unlike Task 1, model performance in the Head Noun Completion task is relatively similar across model sizes, with the smaller model Qwen 2.5 achieving the highest MRR and categorical accuracy. Nonetheless, compared with Task 1, all models obtain substantially lower categorical accuracy, with most accuracies remaining around 70%. This suggests that, although LLMs partially capture human conceptualisation patterns, noticeable gaps remain between human and model noun conceptualisation behaviour.

Behavioural Analysis We fitted a mixed-effects logistic regression model similar to the one used for Task 1. The model tests whether LLM noun completions are affected by classifier type (H_3) and animacy (H_4) in the same direction as human responses. As shown in Table 2, the results reveal larger human–LLM differences than in the Classifier Choice task: more than half of the models show significant differences from humans in the effect of animacy, and Llama 4 also differs significantly from humans in the effect of classifier type. This suggests that LLMs are less consistent with humans in classifier-mediated conceptualisation than in classifier choice.

3.3 Task 3: Referring Expression Generation

We use the REG task to examine the last research question, focusing on whether classifier patterns observed in controlled elicitation generalise to more natural generation settings in humans and LLMs.

3.3.1 Human Findings

We first fitted a binomial mixed-effects logistic regression model on the REG task. The model predicts whether participants selected a specialised classifier from noun specificity, animacy, and their interaction.

The results revealed effects largely consistent with those observed in the controlled Classifier Choice task. Relative to superordinate nouns, both subordinate nouns ($\beta = 1.49, p < .001$) and basic-level nouns ($\beta = 1.57, p < .001$) significantly increased specialised classifier use. Animate referents again strongly reduced specialised classifier selection ($\beta = -3.53, p < .001$). Significant interactions between noun specificity and animacy further showed that the positive effect of more specific nouns was stronger for animate referents.

We further compared the REG task with the controlled Classifier Choice task using a second mixed-effects logistic regression model including task type, noun specificity, animacy, and their interactions. A significant three-way interaction between task type, basic-level nouns, and animacy ($\beta = 0.69, p < .001$) further suggests that the interaction between noun specificity and animacy differs across task settings. These findings indicate that classifier use patterns observed in controlled elicitation tasks do not fully generalise to more natural REG.

3.3.2 Human-LLM Comparison

Metrics. We evaluate LLM in Task 3 using all six metrics introduced in Tasks 1 and 2.

Comparison. Table 1 reports the human-LLM comparison for the REG task. For classifier use, the gap between smaller models and humans remains similar to that in the Classifier Choice task, whereas larger models show stronger alignment with human distributions in free generation. In particular, GLM and Gemini achieve near-perfect classifier MRR for animate referents. This again suggests that classifier choice patterns differ across task settings, highlighting the limitation of isolated Classifier Choice tasks for evaluating LLM classifier use.

For head noun use, BLEU scores decrease compared with the Head Noun Completion task, likely because LLMs generate more pre-modifiers in free referring expressions. However, the other noun-based metrics remain largely stable. This suggests that LLM conceptualisation may be determined mainly from the visual input itself and changes little during language generation.

Behavioural Analysis. We fitted a mixed-effects logistic regression model to test whether LLMs' classifier use in REG is affected by noun specificity (H_5) and animacy (H_6) in the same direction as human responses. As shown in Table 2, all models except GPT show significant differences from humans in the effect of animacy.

4 Can we encode lessons from human experiments into LLM?

Previous sections showed that although LLMs already achieve strong performance on Chinese classifier use, they still exhibit systematic gaps from human behaviour across the three tasks. A natural question is therefore whether the behavioural regularities identified in human experiments can be explicitly incorporated into prompts to further improve human-LLM alignment.

As a preliminary attempt, we first tested a simpler instruction-only prompt that asked models to complete the tasks in a human-like manner. However, this instruction produced little change in model behaviour (see Appendix E), suggesting that LLMs do not necessarily know what "human-like" classifier use means without more concrete guidance. We therefore further encoded the main findings from the human experiments as explicit behavioural rules in the prompt and re-ran all three tasks (see Appendix F for the full list of rules). Table 3 compares the original results with the rule-enhanced prompt.

Overall, adding human-derived rules leads to clear improvements in both Task 1 and Task 2, with particularly large gains for smaller models. In Task 1, most models show improved classifier MRR and accuracy together with lower JSD, indicating better alignment with human classifier distributions. Similar improvements are also observed in Task 2, suggesting that explicit behavioural regularities extracted from human experiments can help weaker models better approximate human classifier choice and classifier-mediated noun completion.

Model	Task 1			Task 2			Task 3					
	MRR	Acc	JSD	MRR	Acc	BLEU	CL-MRR	CL-Acc	CL-JSD	N-MRR	N-Acc	BLEU
glm	0.848	0.911	0.288	0.838	0.708	0.928	0.878	0.883	0.274	0.814	0.700	0.921
gpt	0.825	0.911	0.302	0.878	0.783	0.970	0.746	0.867	0.356	0.914	0.850	0.909
gemini	0.846	0.928	0.285	0.860	0.733	0.975	0.933	0.933	0.223	0.844	0.750	0.949
qwen3.6	0.830	0.928	0.302	0.863	0.733	0.897	0.819	0.883	0.306	0.861	0.750	0.910
claude	0.840	0.917	0.296	0.833	0.700	0.917	0.878	0.900	0.268	0.822	0.717	0.919
llama-4	0.801	0.900	0.328	0.817	0.675	0.935	0.782	0.833	0.337	0.864	0.800	0.896
qwen3	0.790	0.878	0.333	0.836	0.700	0.925	0.764	0.867	0.355	0.822	0.700	0.886
qwen2.5	0.814	0.911	0.307	0.864	0.758	0.912	0.774	0.900	0.342	0.889	0.783	0.891

Table 3: Overall human–LLM comparison after encoding human-derived rules into the prompt.

For Task 3, the improvements are mainly observed in noun-related metrics. Most models obtain higher noun-category MRR, accuracy, and BLEU scores, suggesting that rules containing possible naming patterns help models better name the target referent in free generation. This further supports the idea that LLM conceptualisation can be partially guided through explicit linguistic constraints.

However, the effect on classifier use in Task 3 is more mixed. Although classifier MRR and accuracy generally improve, classifier JSD often becomes worse. This indicates that the rules push models toward the majority human-preferred classifier, but do not help them reproduce the full range of human disagreement and variability. In other words, rule-based prompting improves agreement with dominant human preferences, but does not fully capture the distributional nature of human classifier use.

Interestingly, the effect of rule prompting also differs between Task 1 and Task 3. While Task 1 shows simultaneous improvements in both categorical and distributional metrics, Task 3 mainly shows improvements in majority-aligned metrics. This further suggests that controlled classifier elicitation and free generation involve different classifier production mechanisms, and that improvements in controlled settings may not fully generalise to more natural generation contexts.

5 Related Work

Classifier Use. In many languages, classifiers are obligatory when numerals modify nouns (Allan, 1977). Such classifier systems are particularly common among languages spoken in Asia, including Sinitic languages such as Mandarin and Cantonese, as well as Japanese, Korean, and Vietnamese. Classifier use in Mandarin has been extensively studied

in theoretical linguistics, focusing on its semantic and grammatical properties (Craig, 1986; Tai and Wang, 1990; Tai, 1994; Zhang, 2007; Barner et al., 2009; Cheng et al., 2008; Cheung et al., 2012; Li et al., 2009); in cognitive science, examining factors that shape classifier choice and conceptualisation (Zhan and Levy, 2018, 2019; Liang and Liao, 2024); and in computational linguistics, developing automatic classifier selection systems (Sornlertlamvanich et al., 1994; Bond and Paik, 2000; Paik and Bond, 2001; Peinelt et al., 2017; Chen et al., 2018).

Considering Variability in the Evaluation of LLMs. Recent work on evaluating the linguistic abilities of LLMs has increasingly recognised that human language use is inherently variable rather than governed by a single categorical choice. Existing studies have examined whether LLMs reproduce human variation in lexical choice and syntactic structures (Guo et al., 2025), referring expressions (Wu et al., 2025), and variation associated with social factors (Bassignana et al., 2026). However, evaluations of classifier use have largely treated classifier prediction as a single-label problem, leaving open whether LLMs capture the variability observed in human classifier choice. Our work extends variability-aware evaluation to Mandarin classifiers by comparing model outputs directly against human response distributions and examining whether models reproduce systematic patterns of variation across different conceptualisations and task settings.

6 Conclusion

In this paper, we propose a controlled experimental framework for studying Chinese classifier use in humans and LLMs across three complementary tasks: Classifier Choice, Head Noun Completion,

and Referring Expression Generation. Our human experiment shows that classifier use is inherently variable: it is shaped by noun specificity, animacy, and task setting, and cannot be adequately represented by a single gold label. We therefore argued for a variability-aware evaluation paradigm that compares LLM outputs against human response distributions rather than categorical labels alone.

Our comparison further shows that, although LLMs achieve strong performance in Chinese classifier use, they still differ systematically from humans in classifier variability, classifier-mediated conceptualisation, and natural generation settings. Rule-based prompting improves human–LLM alignment, especially for smaller models and noun-related metrics, but it mainly pushes models toward majority human preferences rather than capturing the full distribution of human disagreement. These findings suggest that Chinese classifier use provides a useful testbed for evaluating whether LLMs capture not only linguistic correctness, but also human-like variability and conceptualisation in language use.

Limitations

One limitation of the present study is that we mainly focus on visual context, while the linguistic context provided in our tasks is relatively simple. Previous studies have shown that Chinese classifier use is also strongly influenced by broader discourse and contextual information. Future work could therefore incorporate richer linguistic contexts, such as multi-sentence discourse or interactive dialogue settings, to further examine how contextual information shapes classifier use in humans and LLMs.

A further limitation concerns potential model exposure to the ManyNames ZH dataset during pretraining. Although ManyNames ZH contains images and naming descriptions rather than classifier annotations, prior exposure to the image–name pairs could still influence model behaviour. This concern is limited for Task 1, where classifier annotations are not available in the source dataset, and for Task 2, where our primary interest is whether controlled classifier input systematically shifts noun choice. It is more relevant to Task 3, where models freely generate referring expressions and prior exposure may affect head noun selection. We cannot rule out potential data contamination, particularly for noun generation in Task 3. Nev-

ertheless, head noun choice constitutes only one aspect of our Task 3 evaluation, which also includes classifier-specific and distributional measures.

Use of AI Assistants. In this work, we used GPT-5.5 to assist with refining the language of the paper, generating code for data preprocessing and result analysis, and creating workflow diagrams. All experimental design decisions, analyses, and interpretations were made by the authors.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (Project No.T2600598), the MOE (Ministry of Education in China) Project of Humanities and Social Sciences (Project No.25YJC740005), the National Language and Character Research Base (Project No.ZDI145-168), and Fundamental Research Funds for the Central Universities, Academy of Frontier Interdisciplinary Research, Central China Normal University (Project No.JC2026PT-004).

References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Keith Allan. 1977. Classifiers. *Language*, 53(2):285–311.
- AI Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku. *Claude-3 Model Card*, 1(1):4.
- Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, and 1 others. 2025a. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.
- Shuai Bai, Ke qin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025b. [Qwen2.5-vl technical report](#). *ArXiv*, abs/2502.13923.
- David Barner, Shunji Inagaki, and Peggy Li. 2009. Language, thought, and real nouns. *Cognition*, 111(3):329–344.
- Elisa Bassignana, Mike Zhang, Dirk Hovy, and Amanda Cercas Curry. 2026. [Do large language models adapt to language variation across socioeconomic status?](#)

- In *Proceedings of the 13th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 317–338, Rabat, Morocco. Association for Computational Linguistics.
- Francis Bond and Kyonghee Paik. 2000. Reusing an ontology to generate numeral classifiers. In *COLING 2000 Volume 1: The 18th International Conference on Computational Linguistics*.
- Guanyi Chen, Kees van Deemter, and Chenghua Lin. 2018. [SimpleNLG-ZH: a linguistic realisation engine for Mandarin](#). In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 57–66, Tilburg University, The Netherlands. Association for Computational Linguistics.
- Lisa LS Cheng, Jenny Doetjes, and Rint Sybesma. 2008. How universal is the universal grinder. *Linguistics in the Netherlands*, 2008:50–62.
- Pierina Cheung, Peggy Li, and David Barner. 2012. What counts in mandarin chinese: A study of individuation and quantification. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 34.
- Gennaro Chierchia. 1998. Plurality of mass nouns and the notion of “semantic parameter”. In *Events and grammar*, pages 53–103. Springer.
- Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.
- Colette G Craig. 1986. Noun classes and categorization. *Noun Classes and Categorization*, pages 1–489.
- Team Glm, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, and 1 others. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Yanzhu Guo, Guokan Shang, and Chloé Clavel. 2025. [Benchmarking linguistic diversity of large language models](#). *Transactions of the Association for Computational Linguistics*, 13:1507–1526.
- Yunke He, Xixian Liao, Jialing Liang, and Gemma Boleda. 2023. The impact of familiarity on naming variation: A study on object naming in mandarin chinese. In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 456–475.
- Jani Järnfors, Guanyi Chen, Kees van Deemter, and Rint Sybesma. 2021. [Using BERT for choosing classifiers in Mandarin](#). In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 172–176, Aberdeen, Scotland, UK. Association for Computational Linguistics.
- George Lakoff. 1986. Classifiers as a reflection of mind. *Noun classes and categorization*, 7:13–51.
- Peggy Li, Yarrow Dunham, and Susan Carey. 2009. Of substance: The nature of language effects on entity construal. *Cognitive psychology*, 58(4):487–524.
- Jialing Liang and Xixian Liao. 2024. How does grammatical category influence conceptual categorization: The case of chinese classifiers. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 46.
- Kyonghee Paik and Francis Bond. 2001. Multilingual generation of numeral classifiers using a common ontology. In *Proceedings of the 19th International Conference on the Computer Processing of Oriental Languages*, pages 141–147.
- Nicole Peinelt, Maria Liakata, and Shu-Kai Hsieh. 2017. [ClassifierGuesser: A context-based classifier prediction system for Chinese language learners](#). In *Proceedings of the IJCNLP 2017, System Demonstrations*, pages 41–44, Tapei, Taiwan. Association for Computational Linguistics.
- Virach Sornlertlamvanich, Wantanee Pantachat, and Surapant Meknavin. 1994. Classifier assignment by corpus-based approach. *arXiv preprint cmp-lg/9411027*.
- James Tai and Lianqing Wang. 1990. A semantic study of the classifier tiao. *Journal of the Chinese Language Teachers Association*, 25(1):35–56.
- James HY Tai. 1994. Chinese classifier systems and human categorization. In *honor of William S.-Y. Wang: Interdisciplinary studies on language and language change*, pages 479–494.
- Qwen Team. 2026. Qwen3. 5-omni technical report. *arXiv preprint arXiv:2604.15804*.
- Kees van Deemter, Le Sun, Rint Sybesma, Xiao Li, Bo Chen, and Muyun Yang. 2017. [Investigating the content and form of referring expressions in Mandarin: introducing the mtuna corpus](#). In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 213–217, Santiago de Compostela, Spain. Association for Computational Linguistics.
- Chengzhao Wu, Guanyi Chen, Fahime Same, and Tingting He. 2025. [Analysing reference production of large language models](#). In *Proceedings of the 18th International Natural Language Generation Conference*, pages 182–194, Hanoi, Vietnam. Association for Computational Linguistics.

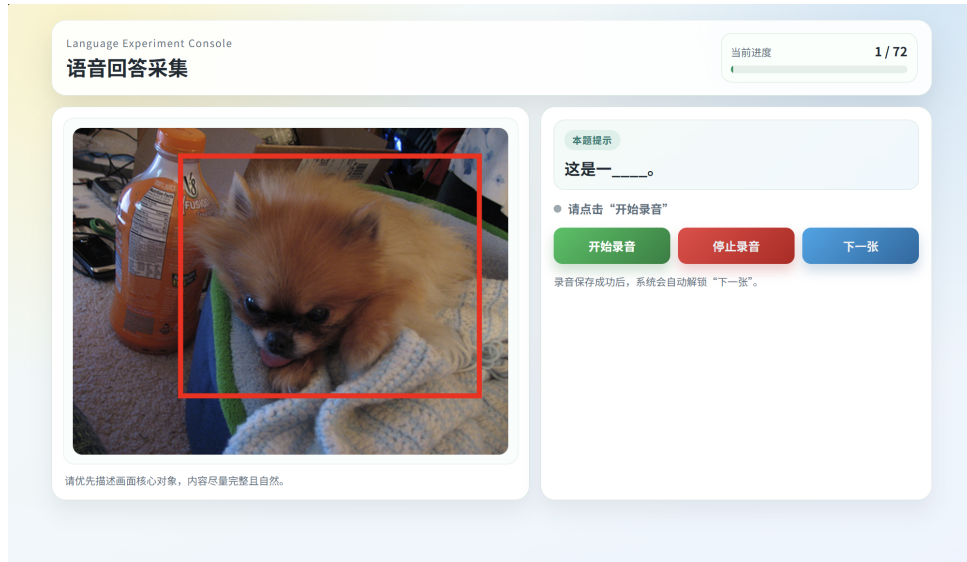


Figure 2: Example of the experimental interface used in the human experiment.

Meilin Zhan and Roger Levy. 2018. [Comparing theories of speaker choice using a model of classifier production in Mandarin Chinese](#). 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 1997–2005, New Orleans, Louisiana. Association for Computational Linguistics.

Meilin Zhan and Roger Levy. 2019. Availability-based production predicts speakers’ real-time choices of mandarin classifiers. In *Proceedings of the Annual Conference of the Cognitive Science Society*.

Hong Zhang. 2007. Numeral classifiers in mandarin chinese. *Journal of East Asian Linguistics*, 16(1):43–59.

Ziqi Zhang, Jianfei Ma, Emmanuele Chersoni, You Jieshun, and Zhaoxin Feng. 2025. [From BERT to LLMs: Comparing and understanding Chinese classifier prediction in language models](#). In *Proceedings of the 8th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 317–329, Suzhou, China. Association for Computational Linguistics.

Mingzhe Zheng and Jie Liu. 2023. The mandarin classifier is changing: How and why. *Proceedings of the Linguistic Society of America*, 8(1):5498–5498.

Wei Zhou. 2025. The general classifier gè in chinese: Evidence from mutual information computation. In *Chinese Lexical Semantics*, pages 41–53, Singapore. Springer Nature Singapore.

A Experiment Interface

Figure 2 shows an example of the experimental interface used in the human experiment.

B Statistical Analysis Details

B.1 Classifier Choice

To examine the effects of noun specificity and animacy on classifier choice, we fitted a binomial mixed-effects logistic regression model predicting whether participants selected a specialised classifier. Because the classifier 只 was overwhelmingly preferred for animate referents, we recoded 只 as a general classifier in animate conditions to better distinguish between general and specialised classifier use.

The model included noun specificity (superordinate, basic-level, subordinate), animacy (animate vs. inanimate), and their interaction as fixed effects. Random intercepts were included for participants and items. Superordinate nouns and inanimate referents were used as the reference levels. The model specification was: $\text{is_special_recode} \sim \text{noun_type} * \text{animacy} + (1|\text{subject_id}) + (1|\text{item_id})$. The testing results are reported in Table 4.

B.2 Head Noun Completion

To examine how classifier type and animacy affect noun specificity in the Head Noun Completion task, we fitted three pairwise binomial mixed-effects logistic regression models comparing: (1) basic-level vs. superordinate nouns, (2) subordinate vs. superordinate nouns, and (3) basic-level vs. subordinate nouns.

Each model included classifier type (general vs. specialised classifier), animacy, and their interac-

Predictor	β	SE	z	p	OR
Intercept	1.87	0.06	29.81	< .001	6.46
Subordinate noun	-0.20	0.10	-2.01	.044	0.81
Basic-level noun	-0.28	0.11	-2.56	.010	0.75
Animate	-3.81	0.09	-42.52	< .001	0.02
Subordinate $\times$ Animate	1.07	0.14	7.38	< .001	2.91
Basic-level $\times$ Animate	1.57	0.14	10.95	< .001	4.81

Table 4: Fixed effects from the binomial mixed-effects logistic regression model predicting specialised classifier use in the Classifier Choice task. Superordinate nouns and inanimate referents were used as reference levels.

Contrast	Predictor	β	SE	z	p	OR
Basic vs. Superordinate	Intercept	1.65	0.11	15.18	< .001	5.23
	specialised classifier	0.62	0.18	3.50	< .001	1.85
	Animate	-0.45	0.15	-2.97	.003	0.64
	specialised $\times$ Animate	1.08	0.26	4.17	< .001	2.93
Subordinate vs. Superordinate	Intercept	1.01	0.12	8.37	< .001	2.74
	specialised classifier	1.10	0.19	5.87	< .001	3.01
	Animate	-0.50	0.17	-2.91	.004	0.61
	specialised $\times$ Animate	0.54	0.28	1.91	.056	1.71
Basic vs. Subordinate	Intercept	0.56	0.07	7.50	< .001	1.75
	specialised classifier	-0.12	0.10	-1.27	.205	0.89
	Animate	0.21	0.10	2.04	.042	1.24
	specialised $\times$ Animate	0.08	0.13	0.60	.548	1.08

Table 5: Fixed effects from the pairwise mixed-effects logistic regression analyses predicting noun specificity in the Head Noun Completion task.

tion as fixed effects, with random intercepts for participants and items. General classifiers and inanimate referents were used as the reference levels. The model specification was: $\text{noun_type} \sim \text{classifier_type} * \text{animacy} + (1|\text{subject_id}) + (1|\text{item_id})$. The testing results are reported in Table 5.

In addition to findings in Section 3.2, classifier type and animacy interacted significantly in the basic-level vs. superordinate model ($\beta = 1.08$, $p < .001$), indicating that the effect of specialised classifiers on noun specificity was stronger for animate referents.

In contrast, no significant classifier effect was found in the basic-level vs. subordinate comparison ($\beta = -0.12$, $p = .205$), suggesting that specialised classifiers primarily promote noun choices that are more specific than superordinate categories, but do not reliably distinguish between basic-level and subordinate conceptualisations.

B.3 Referring Expression Generation

To examine classifier use in the Referring Expression Generation task, we fitted a binomial mixed-effects logistic regression model predicting whether participants selected a specialised classifier. The model included noun speci-

ficity, animacy, and their interaction as fixed effects, with random intercepts for participants and items: $\text{is_special} \sim \text{noun_type} * \text{animacy} + (1|\text{subject_id}) + (1|\text{item_id})$. General classifiers and inanimate referents were used as the reference levels. The testing results are reported in Table 6.

We further compared the Referring Expression Generation task with the controlled Classifier Choice task using a second mixed-effects logistic regression model: $\text{is_special} \sim \text{task_type} * \text{noun_type} * \text{animacy} + (1|\text{subject_id}) + (1|\text{item_id})$. The testing results are reported in Table 7. The analysis revealed substantial task differences. Relative to the Referring Expression Generation task, the controlled Classifier Choice task produced significantly more specialised classifiers in the superordinate–inanimate condition ($\beta = 1.46$, $SE = 0.06$, $z = 23.28$, $p < .001$, $OR = 4.33$). Moreover, the effects of subordinate nouns and basic-level nouns relative to superordinate nouns were both significantly weaker in the controlled task, indicating that noun specificity played a larger role in free generation than in controlled elicitation.

Predictor	β	SE	z	p	OR
Intercept	0.18	0.11	1.61	.107	1.19
Subordinate noun	1.49	0.20	7.32	< .001	4.46
Basic-level noun	1.57	0.14	11.46	< .001	4.80
Animate	-3.53	0.16	-22.03	< .001	0.03
Subordinate $\times$ Animate	0.76	0.33	2.28	.023	2.13
Basic-level $\times$ Animate	0.79	0.19	4.21	< .001	2.20

Table 6: Fixed effects from the mixed-effects logistic regression model predicting specialised classifier use in the Referring Expression Generation task.

Predictor	β	SE	z	p	OR
Classifier Choice task	1.46	0.06	23.28	< .001	4.33
Task $\times$ Subordinate noun	-1.41	0.10	-13.84	< .001	0.24
Task $\times$ Basic-level noun	-1.58	0.11	-14.23	< .001	0.21
Task $\times$ Animate	-0.28	0.09	-3.07	.002	0.76
Subordinate $\times$ Animate	1.02	0.13	7.67	< .001	2.78
Basic-level $\times$ Animate	1.06	0.11	9.32	< .001	2.90
Task $\times$ Subordinate $\times$ Animate	0.23	0.15	1.55	.120	1.25
Task $\times$ Basic-level $\times$ Animate	0.69	0.14	4.76	< .001	1.98

Table 7: Task comparison model examining differences between the Referring Expression Generation task and the controlled Classifier Choice task.

C Details of the Evaluation Metrics

To compare distributions, we employ Mean Reciprocal Rank (MRR), and Jensen–Shannon Divergence (JSD).

C.1 Mean Reciprocal Rank (MRR)

MRR evaluates how highly an LLM-selected classifier is ranked within the human response distribution. For each item, classifiers are ranked according to their frequency in human responses. Let r_i denote the rank of the classifier selected by the LLM for item i . The MRR is computed as:

$$\text{MRR} = \frac{1}{N} \sum_{i=1}^N \frac{1}{r_i}$$

where N is the total number of evaluation items.

C.2 Jensen–Shannon Divergence (JSD)

JSD measures the distributional divergence between the classifier distribution produced by an LLM and the corresponding human classifier distribution. Given a human distribution P and an LLM distribution Q , the JSD is defined as:

$$\text{JSD}(P \parallel Q) = \frac{1}{2}D_{\text{KL}}(P \parallel M) + \frac{1}{2}D_{\text{KL}}(Q \parallel M)$$

where

$$M = \frac{1}{2}(P + Q)$$

and D_{KL} denotes the Kullback–Leibler divergence:

$$D_{\text{KL}}(P \parallel Q) = \sum_x P(x) \log \frac{P(x)}{Q(x)}$$

Lower JSD values indicate greater similarity between human and LLM classifier distributions.

D Results for Animate and Inanimate Items

Table 8 reports the breakdown performance of the evaluated LLMs with respected to animate and inanimate stimuli.

E Results of Human-like Prompting

As a preliminary prompt intervention, we tested whether simply instructing LLMs to behave in a human-like manner would improve their alignment with human responses. Specifically, we added an instruction asking models to complete each task as a native Mandarin speaker would (see Appendix F for the whole prompt). We then re-ran all three tasks using this prompt and evaluated the outputs using the same metrics as in the main experiments.

As shown in Table 9, this instruction-only intervention produced little change compared with the original prompt. This suggests that vague instructions such as “be human-like” are insufficient for improving classifier use, likely because LLMs do

Table 8: Comprehensive evaluation results (MRR, Accuracy, BLEU, and JSD) of eight LLMs for animate and inanimate targets under Prompt-1. The best results in each column are highlighted in bold.

Target	Model	MRR ($\uparrow$)					Accuracy ($\uparrow$)					BLEU ($\uparrow$)		JSD ($\downarrow$)		
		CL	N	FCL	FN	NI-CL	CL	N	FCL	FN	NI-CL	N	F	CL	FCL	NI-CL
Animate	glm	0.850	0.814	1.000	0.744	0.898	0.944	0.650	1.000	0.600	0.922	0.941	0.914	0.247	0.157	0.219
	gpt	0.839	0.831	0.811	0.894	0.726	0.933	0.683	0.867	0.833	0.911	0.974	0.937	0.262	0.314	0.367
	gemini	0.850	0.789	1.000	0.844	0.848	0.944	0.683	1.000	0.767	0.944	0.935	0.971	0.251	0.157	0.254
	qwen3.6	0.804	0.800	0.861	0.839	0.863	0.900	0.633	0.833	0.733	0.944	0.934	0.945	0.297	0.260	0.241
	claude	0.852	0.819	0.894	0.750	0.861	0.944	0.650	0.867	0.633	0.944	0.940	0.934	0.250	0.242	0.244
	llama-4	0.822	0.847	0.811	0.861	0.835	0.400	0.717	0.267	0.800	0.944	0.873	0.896	0.277	0.299	0.266
	qwen3	0.809	0.872	0.817	0.839	0.854	0.378	0.750	0.367	0.733	0.944	0.975	0.874	0.298	0.297	0.250
	qwen2.5	0.831	0.906	0.911	0.867	0.878	0.389	0.817	0.333	0.800	0.944	0.981	0.958	0.267	0.216	0.229
Inanimate	glm	0.822	0.819	0.811	0.789	0.793	0.889	0.683	0.900	0.633	0.856	0.935	0.891	0.344	0.342	0.353
	gpt	0.793	0.856	0.608	0.789	0.769	0.889	0.750	0.800	0.633	0.878	0.943	0.840	0.358	0.471	0.370
	gemini	0.833	0.764	0.867	0.806	0.789	0.911	0.617	0.900	0.667	0.900	0.904	0.903	0.327	0.285	0.348
	qwen3.6	0.749	0.864	0.653	0.772	0.781	0.878	0.750	0.900	0.567	0.889	0.940	0.824	0.395	0.429	0.358
	claude	0.804	0.778	0.750	0.756	0.807	0.911	0.650	0.867	0.567	0.889	0.874	0.842	0.349	0.369	0.338
	llama-4	0.770	0.739	0.686	0.794	0.791	0.900	0.567	0.800	0.700	0.911	0.844	0.862	0.378	0.424	0.353
	qwen3	0.737	0.772	0.678	0.756	0.794	0.878	0.667	0.900	0.600	0.889	0.882	0.829	0.397	0.420	0.348
	qwen2.5	0.741	0.844	0.669	0.778	0.769	0.878	0.750	0.900	0.600	0.878	0.856	0.825	0.392	0.437	0.370

Model	Task 1			Task 2			Task 3					
	MRR	Acc	JSD	MRR	Acc	BLEU	CL-MRR	CL-Acc	CL-JSD	N-MRR	N-Acc	BLEU
glm	0.840	0.911	0.293	0.817	0.675	0.927	0.881	0.917	0.271	0.844	0.733	0.889
gpt	0.814	0.906	0.309	0.831	0.700	0.951	0.715	0.850	0.385	0.850	0.733	0.889
gemini	0.842	0.917	0.289	0.810	0.675	0.947	0.911	0.950	0.238	0.828	0.717	0.938
qwen3.6	0.784	0.889	0.339	0.835	0.692	0.935	0.785	0.883	0.324	0.839	0.700	0.892
claude	0.825	0.928	0.303	0.796	0.642	0.904	0.819	0.850	0.317	0.761	0.617	0.882
llama-4	0.806	0.906	0.321	0.822	0.683	0.854	0.790	0.867	0.331	0.872	0.767	0.912
qwen3	0.776	0.883	0.351	0.818	0.692	0.921	0.747	0.867	0.358	0.822	0.717	0.858
qwen2.5	0.787	0.900	0.328	0.858	0.750	0.916	0.763	0.883	0.346	0.867	0.767	0.904

Table 9: Overall human-LLM comparison after instructing LLMs to be humanlike.

not have an explicit operational understanding of what human-like classifier behaviour entails. This motivated our rule-based prompt intervention in Section 4, where we instead encoded concrete behavioural regularities extracted from the human experiments.

F Prompt

Table 10 shows the full prompt we used in the main experiment, while Table 11 includes the intended linguistic rules of classifier use.

你将看到一张图

(You will see an image and a Chinese sentence with an underline.)

请根据红框中的目标，对下划线做最小必要补全，并只输出补全后的完整句子。

要求:

- (First determine the components already present in the sentence, then decide what needs to be filled in the underline.)

- (Only fill missing components, do not repeat existing parts of the sentence.)

- (Base your judgment solely on the target inside the red box, ignoring anything outside.)

- (Do not add extra details like color, position, action, usage, unless required by the original sentence.)

- (If multiple valid answers exist, choose the shortest and most natural one without over-completion.)

(Output Requirements)

(Only output the completed sentence; no explanations, analysis, or extra content.)

待补全句子: {question.prompt}

请按以下规则补全:

- (“这是一_____。” → Include both classifier and noun in the completion.)

- (Only fill in the classifier.)

- (Only fill in the noun.)

- (If the original sentence already contains a classifier, do not add another; if it contains a noun or numeral, do not add another.)

- (Only output the completed full sentence.)

16987

[System Prompt]

...

...

要求:

...

6. 回答要与中文母语者的回答习惯一致。

7. 母语者的补全习惯按以下规则理解:

(Completion habits of native speakers are understood according to the following rules:)

7.1 名词的具体程度分为三类:

(The specificity of nouns is categorized into three levels:)

- 上位词: 动物、植物、载具、家具、衣服、建筑、食物等大类词。

(-Hyponyms: broad categories such as animals, plants, vehicles, furniture, clothing, buildings, food, etc.)

- 普通词: 猫、狗、鸟、车、椅子、花、房子、蛋糕等常用基本名词。

(-Common nouns: frequently used basic nouns like cat, dog, bird, car, chair, flower, house, cake, etc.)

- 下位词: 波斯猫、金毛、公交车、办公椅、玫瑰花、摩天楼等更具体名词。

(-Hyponyms: more specific nouns like Persian cat, Golden Retriever, bus, office chair, rose, skyscraper, etc.)

7.2 补名词时, 优先使用日常最常见、最自然的普通词; 不要无依据地使用过细的下位词。

(When filling nouns, prefer the most common and natural words; do not use overly specific hyponyms without justification.)

7.3 若图片目标可以用普通词自然命名, 优先使用普通词; 只有当目标类别非常明确, 且该下位词在中文中常用自然时, 才使用下位词。

(If the target can naturally be named with a common noun, prefer it; only use hyponyms when the category is very clear and the hyponym is commonly used naturally in Chinese.)

7.4 若无法确定具体种类, 使用普通词, 不使用下位词。

(If the specific category cannot be determined, use common nouns; do not use hyponyms.)

7.5 若原句已有特殊量词, 如辆、张、把、栋、座、匹、头、条、件, 应补与该量词自然搭配的名词, 避免补动物、植物、载具、家具、食物等过宽泛上位词。

(If the original sentence already has a specific classifier (e.g., 辆, 张, 把, 栋, 座, 匹, 头, 条, 件), complete with nouns that naturally match the classifier, avoiding overly broad hypernyms.)

7.6 若原句已有通用量词, 如个、只, 补名词时优先使用普通词, 不要无依据地补过细下位词。

(If the original sentence has a general classifier (e.g., 个, 只), prefer common nouns when filling; do not use overly specific hyponyms without justification.)

7.7 补量词时, 有生命目标, 尤其是动物, 优先使用“只”; 不要轻易使用更专门的量词, 除非该搭配非常固定自然。

(When filling classifiers for animate targets, especially animals, prefer “只”; do not use more specific classifiers unless the combination is very fixed and natural.)

7.8 动物中的固定搭配可以使用特殊量词, 例如一匹马、一头牛、一条鱼; 其他动物通常用“一只”。

(Fixed combinations for animals may use specific classifiers (e.g., 一匹马, 一头牛, 一条鱼); for other animals, generally use “一只”).

7.9 无生命目标若有稳定搭配, 优先使用特殊量词, 例如一辆车、一张桌子、一把椅子、一件衣服、一栋楼、一座建筑。

(For inanimate targets with stable classifier combinations, prefer specific classifiers (e.g., 一辆车, 一张桌子, 一把椅子, 一件衣服, 一栋楼, 一座建筑).)

7.10 无生命目标若没有稳定量词搭配, 使用通用量词“个”。

(For inanimate targets without stable classifier combinations, use the general classifier “个”).

7.11 补“量词+名词”时, 先选择最自然的名词, 再选择该名词最自然搭配的量词; 不要为了使用特殊量词而选择过细或不自然的名词。

(When filling “classifier + noun”, first choose the most natural noun, then its most natural classifier; do not select overly specific or unnatural nouns just to use a special classifier.)

7.12 若特殊量词和通用量词都可以, 选择更短、更常见、更符合日常口语的表达。

(If both special and general classifiers are acceptable, choose the shorter, more common, and colloquial expression.)

7.13 总体优先级是: 自然普通词> 明确且常用的下位词> 过宽泛的上位词; 但如果原句已经限定了上位类别, 只补缺失成分, 不改变句子结构。

(Overall priority: common natural nouns > specific and commonly used hyponyms > overly broad hypernyms; if the original sentence already specifies a hypernym, only fill missing parts without changing the sentence structure.)

输出要求:

(Output Requirements)

...

[User Content Input Rules]

...

...

Table 11: The prompt with linguistic rules encoded. The omitted part is identical to the prompt above.