Benchmarking Argument Identification and Ellipsis Recovery in Chinese Comparative Constructions

Wenjing Xu
Institute of Language Sciences
Shanghai International Studies University
xuwenjing@shisu.edu.cn

Hongzhi Xu
Institute of Language Sciences
SISU, Shanghai China
hxu@shisu.edu.cn

Abstract

Ellipsis is pervasive in comparative constructions across languages, and recovering omitted constituents is essential for accurate semantic interpretation. In this work, we introduce a new benchmark for two related tasks for Chinese: (i) comparative argument identification and (ii) recovery of elliptical constituents. We build a corpus of 2,076 Chinese comparative sentences, manually annotated for comparative argument boundaries and elliptical constituent recovery. In terms of semantics, five comparison dimensions are differentiated. We evaluate three large language models, including DeepSeek-V3.1, GPT-4o, and Gemini-2.0-Flash. Results show that all models perform well on comparative argument identification but substantially worse on ellipsis recovery. These findings reveal persistent limitations of current LLMs in recovering implicit semantic content. The resource is publicly available to support future research on structured semantic understanding in Chinese and beyond [1]

1 Introduction

In recent years, large language models (LLMs) have achieved surprisingly strong performance across a wide range of natural language processing tasks. Nevertheless, their ability to handle implicit structures and ellipsis still warrants finer-grained investigation, especially for non-English languages such as Chinese. Existing benchmarks have focused on various tasks regarding natural language semantics, including sentiment analysis, text classification, natural language inference, and question answering (Kamalloo et al., 2023; Peskine et al., 2023; Hannani et al., 2024; Yang et al., 2023).

Ellipsis is a highly pervasive phenomenon in languages including Chinese, and it is especially prominent in comparative constructions. The typical syntactic pattern of Chinese comparative constructions is 'ARG0 *bi* ARG1 V ARG2', where ARG0 and ARG1 denote the objects under comparison, i.e. comparee and standard; *bi* is a marker of comparative constructions, typically treated as a preposition but exhibiting distinct syntactic behavior; V is the predicate expressing the comparative result, which also determines the dimension of comparison; and ARG2 indicates the degree of difference and is often optional. Example (1) illustrates a typical case: *ta de shu* 'her books' (ARG0) and *wo de shu* 'my books' (ARG1) are the compared objects, *duo* 'more' expresses the comparative relation, and *liangbei* 'twice' (ARG2) specifies the degree of difference. The implicit comparison dimension in this example is therefore *quantity*.

(1) 她的书比我的书多两倍。

Ta de shu bi wo de shu duo liangbei.
She DE book BI I DE book more twice.

'Her books are twice more than mine.'

Although Chinese also employs other comparative constructions, such as *gen...yiyang* ("as...as") for equality comparison and *yue...yue...* ("the more..., the more...") for correlative comparison, the *bi*-comparative is the most productive and prototypical construction for expressing inequality comparison in Modern Chinese. It has been extensively studied in the linguistic literature and is particularly rich in argument ellipsis. We therefore selected *bi*-comparatives as a representative and well-defined testbed for investigating comparative argument identification and ellipsis recovery. For convenience, and without ambiguity, we will simply refer to these as comparative constructions.

Linguistic studies have shown that comparative constructions are structurally diverse and frequently involve the omission of comparative arguments or predicates (Liu et al., 2001; Li, 1986; Liu, 2004). The examples from (2) to (4) demon-

[1] https://github.com/xuhongzhi/chinese-comparative

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 31694–31707
October 24-29, 2026 ©2026 Association for Computational Linguistics

strate different types of ellipsis in Chinese comparative constructions, where the elliptical elements (as shown in brackets) can occur in different parts of the comparison arguments.

(2) 他的表演比我(的表演)好。

Ta de biaoyan bi wo (de biaoyan) hao.
She DE performance BI I (DE performance) good.

'Her performance is better than mine.'

(3) 她有很多书，(她的书)比我（的书）多。

Ta you henduo shu, (ta de shu) bi wo (de shu) duo.
She have many books, (she DE books) BI I (DE shu) more.

'She has many books, more than I do.'

(4) 她(现在)的状况比(她)过去(的状况)好。

Ta (xianzai) de zhuangkuang bi (ta) guoqu (de zhuangkuang) hao.
She (now) DE condition BI (she) before (DE condition) good.

'Her condition is better than before.'

The structure and ellipsis mechanisms of Chinese *bi*-comparative constructions have been extensively discussed in the linguistic literature from syntactic, typological, and semantic perspectives (Ren, 1987; Liu, 2003). However, only a limited number of NLP studies have examined comparative constructions, primarily through the development of comparative sentence corpora and models for structure identification and ellipsis recovery (Zhou et al., 2023). Existing studies on ellipsis phenomena typically focus on structurally isolated sentences and cover only a limited range of ellipsis types. As a result, they mainly employ task-specific models rather than general-purpose LLMs. Consequently, it remains unclear whether their findings generalize to contemporary LLMs.

Comparative constructions encode structured relational semantics across domains including space, time, quantity, social hierarchy and other topological relations. Identifying and recovering elliptical constituents is also essential for many downstream NLP tasks, including information extraction, machine translation, and sentiment analysis (Jindal and Liu, 2006). Ellipsis in comparative constructions is pervasive and highly flexible. Recovering elliptical constituents is therefore essential for constructing accurate semantic relations between entities. Comparative constructions therefore provide a challenging testbed for evaluating whether modern LLMs can go beyond surface-level pattern matching in semantic interpretation. Although large language models perform well on many language understanding tasks, it remains unclear whether they can reliably process the implicit semantics underlying comparative constructions. This motivates the development of a benchmark for evaluating comparative argument identification and ellipsis recovery in Chinese.

In this study, we construct an evaluation dataset of 2,076 comparative sentences drawn from authentic written Chinese and annotate them with comparative argument boundaries, elliptical constituents, and comparison dimensions. We formulate the task as three subtasks: (1) identification of the boundaries of comparative arguments; (2) recovery of elliptical constituents where applicable; and (3) classification of comparison dimensions. This decomposition enables us to distinguish surface-level structural identification from deeper semantic reconstruction. We evaluate three representative large language models, namely DeepSeek-V3.1, GPT-4o, and Gemini-2.0-Flash, under a unified evaluation protocol to assess their ability to handle ellipsis in Chinese comparative constructions.

2 Related Work

2.1 Ellipsis in Comparative Constructions

Ellipsis has been recognized as a fundamental phenomenon in Chinese linguistics since Ma (1929); Li (1924), particularly in comparative constructions where structural asymmetry frequently arises between compared arguments. Prior linguistic studies have shown that ellipsis in Chinese should be recognized under different contextual factors (Lu, 1942, 1979; Wang, 1944), and are typically constrained by recoverability, semantic plausibility, and structural parallelism (Zhu, 1982; Lu, 1979). Researchers have further found that compared arguments tend to preserve syntactic and semantic symmetry, while ellipsis functions as the primary mechanism underlying surface asymmetry (Liu et al., 2001; Ma, 1986; Li, 1986).

Beyond descriptive accounts, research on ellipsis mechanisms has focused on how ellipsis is licensed, and how omitted constituents can be recovered. Ellipsis in comparative sentences is widely regarded

as an economical strategy of expression, with common types including nominal, verbal, and clausal ellipsis, whose interpretation is highly context-dependent and whose omitted content is often semantically recoverable (Yang et al., 2023). Existing studies generally attribute comparative ellipsis to the interaction of structural symmetry, contextual information, and discourse focus: omitted material is typically supplied by an antecedent (Liu, 2020), while the suppression of given information serves to highlight contrastive elements and enhance communicative efficiency (Wu, 2020).

Different theoretical frameworks provide complementary explanations for these mechanisms. In systemic functional linguistics, ellipsis is treated as a discourse cohesion device closely related to field, tenor, and mode (Halliday and Hasan, 2013), offering criteria for when omission does not result in ambiguity. Cognitive linguistic approaches emphasize conceptual integration and cognitive salience, viewing ellipsis as a means of compressing redundant information to foreground comparison (Langacker, 1987; Wu, 2020). Building on these perspectives, more recent work has proposed finer-grained constraints on ellipsis licensing. Niu (2010) showed that ellipsis becomes unstable when comparative arguments diverge in conceptual category, revealing semantic limits on ellipsis recoverability. Corpus-based studies further suggest that ellipsis in differential comparatives exhibits patterned distributions that are, in principle, predictable and modelable (Zhou et al., 2023). These studies suggest that recovering omitted constituents depends jointly on semantic coherence and syntactic consistency between comparative arguments, which motivate our annotation principles for Chinese comparative constructions.

2.2 Computational Approaches to Ellipsis Identification and Recovery

Recent work introduced a range of annotation schemes and datasets for studying ellipsis phenomena across languages and construction types. Existing resources have focused on specific forms of ellipsis, including noun ellipsis (Khullar et al., 2020), sluicing (Anand et al., 2021), and omitted verbs (Testa et al., 2023) in English, as well as broader typological investigations of ellipsis in low-resource languages (Ćavar et al., 2024a). For Chinese, Liu (2020) constructed an AMR-based corpus for systematic analysis of ellipsis types, while Yan et al. (2022) manually annotated ellipsis recovery in a semantic dependency graph corpus.

Comparative constructions play an important role in expressing structured relations between entities and are therefore highly relevant to semantic understanding and information extraction. Early NLP studies on Chinese comparatives mainly focused on comparative sentence identification (e.g. Huang et al., 2008; Liu et al., 2013; Wang et al., 2015) and comparative argument extraction using traditional machine learning approaches based on syntactic and semantic features (Hou and Li, 2008; Wang et al., 2010). However, these studies were developed prior to the emergence of large language models and primarily targeted surface-level structural analysis.

More recent work has begun to investigate ellipsis processing with neural models and LLMs. The study most closely related to ours is Ćavar et al. (2024b), which evaluated both BERT-based models and GPT-style LLMs on the detection and recovery of six types of ellipsis across English, Spanish, and Arabic. Their results showed that both traditional NLP models and LLMs struggle to recover omitted constituents. Other recent studies have formulated ellipsis recovery as a question-answering task using generative LLMs (Qi et al., 2022).

Despite these advances, most studies focus primarily on ellipsis detection rather than ellipsis recovery. In addition, existing resources cover only a limited range of ellipsis types, with little attention paid to comparative constructions, especially in Chinese. To address this gap, we construct a corpus of Chinese comparative constructions annotated with comparative argument boundaries, recovered elliptical constituents, and comparison dimensions. We evaluate state-of-the-art LLMs on three subtasks–argument identification, ellipsis recovery, and comparison dimension classification–and provide detailed error analysis of their semantic understanding capabilities.

3 Building a Benchmark for Ellipsis Recovery in Chinese Comparative Constructions

The benchmark aims to transform implicit information in comparative constructions into explicit, scorable, and reproducible evaluation targets, providing a systematic evaluation of performance on comparative argument identification, ellipsis recovery, and comparison dimension classification.

3.1 Annotation Guideline

3.1.1 Annotation Targets

Interpretation of Chinese comparative sentences requires three interrelated steps: identifying the comparative arguments, namely comparee and standard (henceforth, ARG0 and ARG1 respectively), recovering elliptical constituents if they exist in arguments, and classifying comparison dimension. Since the comparison result argument (ARG2) is optional and typically absent (not due to ellipsis) in most comparative constructions, it is not included in the present study.

3.1.2 Identification of Ellipsis and Recovery Principle

To ensure annotation consistency, we introduce three principles based on insights from linguistic studies: syntactic parallelism, semantic coherence, and component uniqueness. Syntactic parallelism requires the comparative arguments to form structurally comparable units with respect to syntactic category, hierarchical level, and head element. When this parallelism is disrupted, ellipsis may be involved and the omitted material may need to be recovered. Semantic coherence concerns the plausibility of ellipsis recovery at the semantic level: if the surface structures are not fully parallel but the comparative arguments already establish a stable semantic correspondence, it is not considered ellipsis. Component uniqueness constrains the boundaries of comparative arguments, requiring each argument to constitute a single, complete, and independently identifiable syntactic unit. Linear spans across multiple constituents are not valid comparative arguments.

Consider the comparative sentence in (5). First, the predicate 多*duo* 'be more' requires the compared objects to be evaluated along the semantic dimension of *quantity*. Under the principle of semantic coherence, the compared arguments must satisfy this semantic constraint; therefore, the heads of ARG0 and ARG1 must denote entities that possess the property of quantity, which in this case is 书*shu* 'books' as a set.

Second, syntactic parallelism requires ARG0 and ARG1 to share the same syntactic category. Here, both are noun phrases (NPs) with the same internal structure, namely 'Pronoun + DE + Noun', i.e. 她的书*ta de shu* 'her books' and 我的书*wo de shu* 'my books'.

Finally, although the preceding clause already contains the information corresponding to ARG0, namely 她*ta* 'she' and 书*shu* 'books', the principle of constituent uniqueness requires the comparative arguments to form a single syntactic constituent and prohibits annotation across constituent boundaries. Consequently, ARG0 in this example involves complete ellipsis and must be recovered explicitly.

3.1.3 Comparison Dimensions

Every comparison involves a semantic dimension along which the compared entities are evaluated. Recognizing this comparison dimension is essential to understanding the comparative meaning as a whole. However, the comparison dimension is not always explicitly mentioned in comparative constructions. For example, the comparison *A bi B hao* "A is better than B" can refer to any property of A and B. Correctly classifying the comparison dimension therefore also requires a full semantic understanding of the construction as a whole.

The inventory of comparison dimensions constitutes an open class. To make the classification easier to evaluate, we define five comparison dimensions by consulting the semantic taxonomy WordNet: *quantity*, *measure*, *monetary_value*, *metric*, and *quality*. These upper-level categories are designed to cover the major types of comparative semantics observed in the dataset. The *quantity* dimension encompasses countable objects; the *measure* dimension covers uncountable quantities that are typically measured using instruments; the *metric* dimension refers to values derived from multiple variables through a defined calculation or formula; the *quality* dimension refers to abstract properties whose values typically require subjective evaluation; and the *monetary_value* dimension captures quantities particularly expressed in monetary terms. Detailed definitions of each comparison dimension, together with representative examples, are presented in Table 1.

3.2 Data Collection and Annotation

The data used in this study are drawn from the BCC Corpus developed by Beijing Language and Culture University [2]. We first extract candidate sentences containing the comparative marker *bi* and then manually verify them to exclude invalid cases. The final dataset contains 2,076 sentences. Following the annotation guidelines, each sentence was

[2]https://bcc.blcu.edu.cn/

(5) 她有很多书，[(她的书)]{ARG0}比[我(的书)]{ARG1}[多]{V:QUANTITY}。

Ta you henduo shu, [(ta de shu)]{ARG0} bi [wo (de shu)]{ARG1} [duo]{V:QUANTITY}.
She have many book, [(she DE book)]{ARG0} BI [I (DE book)]{ARG1} [more]{V:QUANTITY}.

'She has many books, more than I do.'
(The constituents enclosed in square brackets denote comparative arguments, while those enclosed in round brackets represent recovered elliptical constituents.)

Dimension	Semantic definition	Examples
QUANTITY	the cardinality of a set of objects	number of people, equipment count, student enrollment
MEASURE	measurable properties to which numerical values are assigned according to a defined system	length, area, volume, weight, time, temperature
MONETARY_VALUE	the property of having material worth, often expressed in monetary terms	price, cost, revenue, profit, budget, salary
QUALITY	an essential and distinguishing attribute of an entity	importance, safety, stability, severity, attractiveness
METRIC	an indicator derived from related measures and used to quantify a particular characteristic	growth rate, conversion rate, efficiency, index

Table 1: Definitions of comparison dimensions with illustrative examples.

annotated by a linguistic expert for three components: (1) comparative argument boundaries, (2) recovered arguments, and (3) comparison dimensions.

3.3 Annotation Result and Statistics

3.3.1 Ellipsis Types

Based on the syntactic position of the comparative elements and the semantic function of the elliptical constituents along with their recovery strategies, we categorize ellipsis into seven types: (1) *head*, where the head noun of a comparative argument is omitted; (2) *modifier*, where a modifier of a comparative argument is omitted; (3) *whole*, where the entire comparative argument is absent; (4) *synthetic*, where the elliptical constituents cannot be directly recovered from the context and recovery requires contextual inference and reasoning; (5) *temporal*, where a temporal expression is omitted; (6) *none*, where arguments are fully realized without ellipsis; and (7) *other*, which involves idiomatic or non-standard patterns.

As shown in Table 2, only 32.5% of ARG0 and 14.5% of ARG1 are without ellipsis, and *head* ellipsis in ARG1 and *whole* and *temporal* ellipsis in ARG0 account for the largest proportions of the data. Notably, *synthetic* takes a small proportion, indicating that most elliptical constituents are recoverable from symmetric structure or context. Table 3 shows that *head* ellipsis in ARG1 often co-occurs with *whole* or *temporal* ellipsis in ARG0.

Ellipsis	ARG0		ARG1	
	Freq.	%	Freq.	%
Head	20	1.0	1341	64.6
Modifier	43	2.1	91	4.4
Whole	586	28.2	0	0.0
Synthetic	49	2.4	99	4.8
Temporal	459	22.1	0	0.0
None	675	32.5	300	14.5
Other	244	11.8	244	11.8

Table 2: Distribution of ellipsis types for ARG0 and ARG1.

ARG0	ARG1	Freq.	%
Whole	Head	449	21.63
Temporal	Head	432	20.81
None	Head	388	18.69
Other	Other	244	11.75
None	None	199	9.59
Whole	None	71	3.42
None	Synthetic	67	3.23
Whole	Modifier	44	2.12
Modifier	Head	38	1.83
Synthetic	Head	31	1.49
Whole	Synthetic	22	1.06
None	Modifier	20	0.96

Table 3: Joint distribution of ellipsis types for ARG0 and ARG1.

3.3.2 Distribution of Comparison Dimensions

Regarding comparison dimensions, Table 4 indicates that QUANTITY and MONETARY_VALUE are the most frequent categories, followed by QUALITY, while MEASURE and METRIC occur less frequently. However, the latter two categories exhibit finer semantic boundaries and are more challenging to distinguish.

Dimension	Freq.	%
MEASURE	600	28.9
METRIC	105	5.0
MONETARY_VALUE	614	29.5
QUALITY	347	16.7
QUANTITY	410	19.7

Table 4: Distribution of comparison dimensions in the corpus.

4 Experiments

This section describes our experiments evaluating LLMs' ability to understand Mandarin Chinese *bi*-comparative constructions.

4.1 Experimental Settings

4.1.1 Models and Parameters

We evaluate three representative large language models with strong Chinese language capabilities: DeepSeek-V3.1, GPT-4o, and Gemini-2.0-Flash. All models are accessed via their official APIs and evaluated under identical conditions, including the same input data, prompt content, and evaluation pipeline. To reduce randomness and prioritize structural accuracy over generative diversity, we adopt conservative decoding settings across all models: temperature = 0.2, top_p = 0.9, and max_tokens = 512. Each model processes the full dataset of 2,076 comparative sentences independently, and all outputs are saved in JSON format for subsequent evaluation.

4.1.2 Task Formulation and Prompting Strategy

The evaluation is decomposed into two subtasks corresponding to processing of different linguistic levels. The first subtask targets comparative argument identification and ellipsis recovery. Models are instructed to identify the surface spans of the comparative elements (ARG0 and ARG1) and to recover any elliptical constituents when ellipsis is detected. The judgments follow the same three principles: syntactic parallelism, semantic coherence, and component uniqueness. The prompt provides explicit operational rules and representative examples to guide model behavior across different ellipsis patterns.

The second task focuses on comparison dimension classification. Given a comparative sentence, the model selects one label from a predefined set of five comparison dimensions: *quantity*, *measure*, *monetary_value*, *metric*, and *quality*. Dimension definitions and contrastive examples are included in the prompt to clarify category boundaries. The prompts are in English as we found Chinese and English prompts generate similar performance. See Appendix A for the full prompts of both tasks.

4.2 Results and Analysis

Model outputs are evaluated against human annotations with two different metrics for the evaluation of argument identification and recovery: exact match (EM) and Character n-gram F-score (ChrF). Comparison dimension classification is evaluated using accuracy and macro-F1 measure. We examine the overall results, performance variations across different ellipsis types, and dominant error patterns. For clarity and comparability, we report both argument-level accuracy, covering ARG0 identification, ARG1 identification, ARG0 recovery, ARG1 recovery, and comparison dimension classification, and type-wise statistics.

4.2.1 Overall Results

Table 5 reports the performance of the three models. Overall, all models perform well in boundary identification of ARG1, achieving EM accuracy of 0.96 and ChrF of 0.97. This is consistent with our understanding of the structural properties of Chinese comparatives that ARG1 is never fully omitted and always lies between the comparative marker *bi* and the comparative result. In contrast, ellipsis recovery is substantially more challenging. Ellipsis of ARG0, whether partial or complete, is very common, and its recovery is correspondingly more challenging, yielding EM accuracies of 0.55–0.63 and ChrF scores of 0.82–0.85. By contrast, ARG1 involves only partial ellipsis, and its recovery achieves higher performance, with EM accuracies of 0.65–0.73 and ChrF scores of 0.86–0.89.

Cross-model comparison shows that DeepSeek-V3.1 achieves the best performance in both argument identification and recovery, achieving an overall EM of 0.780 and a ChrF score of 0.881. Gemini-2.0-Flash achieves 0.732 of EM and 0.859 of ChrF. GPT-4o obtains 0.740 of EM and 0.833 of ChrF.

The results of comparison dimension classification are shown in Table 6. GPT-4o achieves the highest accuracy (0.858) and macro-F1 (0.831), substantially higher than DeepSeek V3.1 and Gemini-2.0-Flash. This contrasts with its performance on argument identification and ellipsis recovery, where GPT-4o performs the worst. This pattern indicates that strong semantic abstraction alone does not necessarily translate into stable syn-

Model	ARG0 id.		ARG1 id.		ARG0 rec.		ARG1 rec.		All	
	EM	ChrF	EM	ChrF	EM	ChrF	EM	ChrF	EM	ChrF
DeepSeek-V3.1	0.779	0.802	0.967	0.983	0.631	0.852	0.732	0.889	0.780	0.881
Gemini-2.0-Flash	0.740	0.780	0.960	0.977	0.548	0.818	0.654	0.861	0.732	0.859
GPT-4o	0.618	0.666	0.964	0.980	0.583	0.824	0.679	0.863	0.740	0.833

Table 5: Model performance on argument boundary identification (id.) and ellipsis recovery (rec.).

tactic boundary detection or reliable ellipsis recovery.

Across categories, *monetary_value* and quality achieve high performance, with F1 scores above 0.90. *Quantity* follows with F1 scores in the range of 0.70–0.80, while *measure* and *metric* show lower and more variable performance, with F1 scores ranging from 0.50 to 0.80. As shown in the confusion matrices in Figure 1, *measure* is frequently misclassified as *metric* or *quantity*, particularly by DeepSeek-V3.1 and Gemini-2.0-Flash. This pattern suggests that these models have difficulty clearly distinguishing between the corresponding semantic dimensions.

4.2.2 Performance by Ellipsis Types

As shown in Tables 7, head ellipsis poses the greatest challenge for recovery. Under exact match, all three models achieve ARG1 identification close to 0.98, yet exhibit significant divergence in recovering the core nominal component, with DeepSeek-V3.1 clearly leading (0.55 ARG0 recovery vs. 0.30–0.40 for others). The ChrF scores reveal a more nuanced picture: even when exact match fails, models often produce outputs with substantial character-level overlap (e.g., DeepSeek-V3.1 achieves 0.872 ChrF on ARG0 recovery despite only 0.55 EM), indicating that errors frequently involve minor variations rather than complete failures.

Models perform poorly on modifier ellipsis recovery in ARG1, with exact-match scores as low as approximately 0.10, while ChrF scores for the same task remain much higher (0.59–0.61). This pattern indicates that models struggle to determine whether modifiers function as structurally necessary elements for maintaining parallelism, often producing semantically plausible but structurally asymmetric completions.

For temporal ellipsis, models can reliably identify explicit temporal cues (ChrF 0.78–0.94) but consistently underperform in reconstructing implicit temporal anchors. The gap between ChrF (0.70–0.77) and exact match (0.22–0.39) in ARG0 recovery shows that models often approximate the correct content but fail to reproduce the exact temporal specification.

Synthetic recovery, which lacks explicit syntactic triggers and relies on contextual inference, exhibits a similar pattern: although exact match performance is low (0.12–0.35), ChrF scores are substantially higher (0.57–0.76), indicating partial inference of the required completion. This suggests that the main difficulty lies not in complete failure, but in achieving precise structural alignment with gold annotations.

5 Error Analysis and Discussion

To trace the sources of error, we conducted a retrospective analysis of samples exhibiting argument-level matching failures. Drawing on the error distributions and representative cases, we identified three recurring patterns: boundary misalignment in comparative argument identification, asymmetric recovery of temporal and restrictive components, and systematic confusion in comparison-dimension classification.

5.1 Boundary Misalignment

The first error type concerns boundary segmentation errors in identifying comparative arguments, typically manifesting as either "centripetal contraction" or "boundary expansion". Centripetal contraction refers to the model's tendency to retain only the head noun while discarding temporal, locative, or other restrictive modifiers, yielding semantically incomplete comparative units, as illustrated in Example (6). Boundary expansion, by contrast, occurs when the model incorporates sentence-initial adverbials or clausal background information into ARG0, thereby violating the uniqueness constraint, as illustrated in Example (7).

(6) "Input": "一九五七年该地区农民总收入比1956年提高13%以上。The total income of farmers in this region increased by more than 13% in 1957 compared with 1956.",
"Gold_ARG0": "一九五七年该地区农民

METRICS	MODELS	MEASURE	METRIC	MONETARY	QUALITY	QUANTITY	MACRO-AVG
Precision	DeepSeek-V3.1	0.906	0.341	0.964	0.903	0.708	0.764
	GPT-4o	0.857	0.456	0.977	0.899	0.915	0.821
	Gemini-2.0-Flash	0.920	0.339	0.974	0.883	0.615	0.746
Recall	DeepSeek-V3.1	0.499	0.954	0.960	0.923	0.822	0.832
	GPT-4o	0.812	0.972	0.911	0.932	0.777	0.881
	Gemini-2.0-Flash	0.425	0.972	0.909	0.946	0.815	0.813
F1	DeepSeek-V3.1	0.644	0.502	0.962	0.913	0.761	0.756
	GPT-4o	0.8340	0.621	0.943	0.915	0.840	0.831
	Gemini-2.0-Flash	0.582	0.502	0.940	0.914	0.701	0.728
Accuracy	DeepSeek-V3.1			-			0.789
	GPT-4o			-			0.858
	Gemini-2.0-Flash			-			0.756

Table 6: Model performance on classification of comparison dimensions.

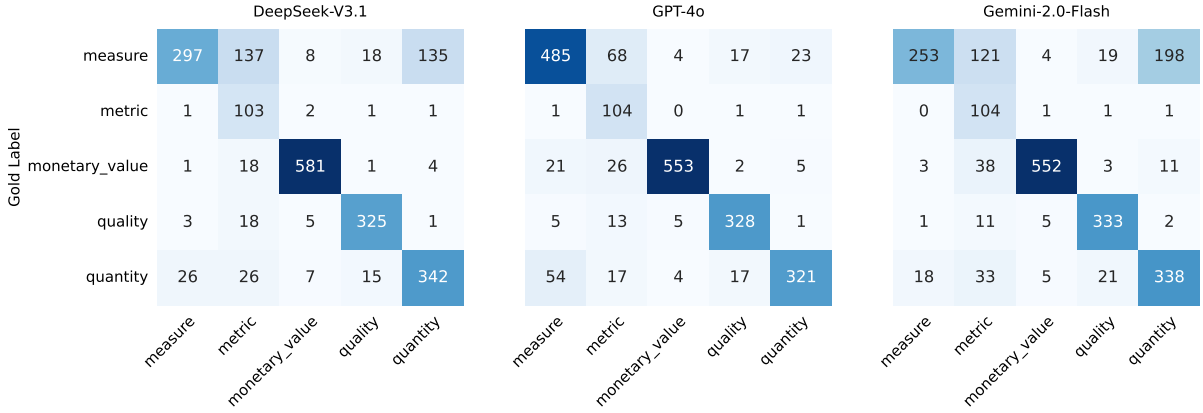


Figure 1: Confusion matrices of the three models on comparison dimension classification.

Type	Model	ARG0 id.		ARG0 rec.		ARG1 id.		ARG1 rec.	
		EM	ChrF	EM	ChrF	EM	ChrF	EM	ChrF
Head	DeepSeek-V3.1	0.850	0.895	0.550	0.872	0.985	0.994	0.822	0.927
	GPT-4o	0.800	0.920	0.400	0.798	0.987	0.994	0.758	0.896
	Gemini-2.0-Flash	0.600	0.807	0.300	0.745	0.982	0.991	0.714	0.887
Modifier	DeepSeek-V3.1	0.535	0.579	0.326	0.680	0.890	0.942	0.989	0.594
	GPT-4o	0.884	0.936	0.233	0.614	0.791	0.879	0.187	0.614
	Gemini-2.0-Flash	0.791	0.855	0.326	0.716	0.736	0.853	0.132	0.604
Temporal	DeepSeek-V3.1	0.758	0.783	0.390	0.772	–	–	–	–
	GPT-4o	0.902	0.940	0.342	0.740	–	–	–	–
	Gemini-2.0-Flash	0.869	0.906	0.224	0.704	–	–	–	–
Synthetic	DeepSeek-V3.1	0.796	0.796	0.163	0.606	0.980	0.993	0.354	0.756
	GPT-4o	0.143	0.152	0.163	0.579	0.980	0.993	0.273	0.759
	Gemini-2.0-Flash	0.408	0.437	0.122	0.583	0.980	0.988	0.293	0.769

Table 7: Model performance on identification (id.) of argument boundaries and the recovery (rec.) for different ellipsis types.

总收入total income of farmers in this region in 1957",
"DeepSeek-V3.1_ARG0": "农民总收入total income of farmers"

(7) "Input": "二月份，一些主要的汽车运输总站为各地农村运送的物资，比一月份又有增加。In February, the goods transported by major automobile transportation terminals to rural areas increased again compared with January.",
"Gold_ARG0": "一些主要的汽车运输总站为各地农村运送的物资goods transported by major terminals to rural areas",
"DeepSeek-V3.1_ARG0": "二月份，一些主要的汽车运输总站为各地农村运送的物资In February, goods transported by major terminals to rural areas"

5.2 Asymmetric Recovery

The second error type involves asymmetric completion during ellipsis recovery, particularly in aligning temporal references and multiple restrictive modifiers. Models often capture overt temporal or nominal cues in the ARG1 but fail to project temporal anchors, spatial restrictions, or scope modifiers from the ARG0 into the recovered ARG1, resulting in non-comparable argument pairs, as shown in Example (8).

(8) "Input": "采煤能力要比1952年将近增加一倍，发电设备能力将增加一倍还多。Coal mining capacity nearly doubled compared with 1952, and power generation equipment capacity more than double.",
"Gold_ARG0_Rec": "今年采煤能力coal mining capacity this year",
"DeepSeek-V3.1_ARG0_Rec": "采煤能力coal mining capacity",
"GPT-4o_ARG0_Rec": "采煤能力coal mining capacity",
"Gemini-2.0-Flash_ARG0_Rec": "采煤能力coal mining capacity"

5.3 Dimension Misclassifications

The third error type involves fine-grained boundary confusion in comparison-dimension classification, primarily among *measure*, *metric*, and *quantity*, as well as cases where *monetary_value* is subsumed under *metric* or *measure* when explicit currency units are absent. Models tend to over-rely on surface cues such as numerals, percentages, or multiples, rather than reliably identifying the measurement type encoded by the comparison result. A typical example is shown in (9), where bean production is measured by weight, i.e. *measure*, but DeepSeek-V3.1 and Gemini-2.0-Flash classify it as *quantity*, which is intended for countable objects.

(9) "Input": "生产大豆四千三百万斤，今年跃进到七千万斤，比1957年也增加了二分之一以上。Soybean production was 43 million jin; this year it leaped to 70 million jin, an increase of more than 50% compared with 1957.",
"Gold_Comp_Dim": "measure",
"DeepSeek-V3.1_Comp_Dim": "quantity",
"GPT-4o_Comp_Dim": "measure",
"Gemini-2.0-Flash_Comp_Dim": "quantity"

6 Conclusion

This paper presents a new benchmark for assessing large language models' ability to handle ellipsis in Mandarin Chinese *bi*-comparative constructions. By transforming implicit structural and semantic requirements into explicitly defined, automatically scorable subtasks, we provide a reproducible benchmark that probes model performance beyond surface-level comparison detection.

This study provides empirical evidence that ellipsis-sensitive phenomena remain a challenging frontier for large language models. The findings support more fine-grained diagnosis of structural and semantic capabilities and inform future model design and evaluation.

Limitations

This study has several limitations. First, the evaluation is restricted to Mandarin *bi*-comparative constructions, and the proposed framework does not directly generalize to other types of ellipsis or to other languages. Second, the dataset was annotated by a single linguistic expert, and its moderate size and the low frequency of certain ellipsis types limit the scope for fine-grained statistical comparisons. Nevertheless, the annotation follows explicitly formulated principles grounded in established linguistic analyses of Chinese comparative constructions, which helps ensure consistency in the recovery of elliptical constituents. Finally, the experiments use prompt-based inference without task-specific fine-tuning and therefore reflect the models' zero- or few-shot capabilities rather than their performance under supervised adaptation.

References

Pranav Anand, Daniel Hardt, and James McCloskey. 2021. The santa cruz sluicing data set. *Language*, 97(1):e68–e88.

Damir Ćavar, Ludovic Mompelat, and Muhammad Abdo. 2024a. The typology of ellipsis: a corpus for linguistic analysis and machine learning applications. In *Proceedings of the 6th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 46–54.

Damir Ćavar, Zoran Tiganj, Ludovic Veta Mompelat, and Billy Dickson. 2024b. Computing ellipsis constructions: Comparing classical nlp and llm approaches. In *Proceedings of the Society for Computation in Linguistics 2024*, pages 217–226.

M.A.K. Halliday and Ruqaiya Hasan. 2013. *Cohesion in english*. Routledge.

Mohamed Hannani, Abdelhadi Soudi, and Kristof Van Laerhoven. 2024. Assessing the performance of chatgpt-4, fine-tuned bert and traditional ml models on moroccan arabic sentiment analysis. In *Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities*, pages 489–498.

Feng Hou and Guo-hui Li. 2008. Mining chinese comparative sentences by semantic role labeling. In *2008 International Conference on Machine Learning and Cybernetics*, volume 5, pages 2563–2568. IEEE.

Xiaojiang Huang, Xiaojun Wan, Jianwu Yang, and Jianguo Xiao. 2008. Learning to identify comparative sentences in chinese text. In *Pacific Rim International Conference on Artificial Intelligence*, pages 187–198. Springer.

Nitin Jindal and Bing Liu. 2006. Mining comparative sentences and relations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 22, page 9.

Ehsan Kamalloo, Nouha Dziri, Charles Clarke, and Davood Rafiei. 2023. Evaluating open-domain question answering in the era of large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5591–5606.

Payal Khullar, Kushal Majmundar, and Manish Shrivastava. 2020. Noel: An annotated corpus for noun ellipsis in english. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 34–43.

Ronald W. Langacker. 1987. *Foundations of cognitive grammar, Vol. 1, Theoretical Prerequisites*. Stanford: Stanford University Press.

Jinxi Li. 1924. *A New Chinese Grammar*. The Commercial Press.

Linding Li. 1986. *Modern Chinese Sentence Patterns*. The Commercial Press.

Danqing Liu. 2003. A survey framework and research approach for differential comparative sentences. *Modern Linguistic Theory and Minority Language Studies in China. Beijing: Ethnic Publishing House*, pages 1–22.

Quanchao Liu, Heyan Huang, Chen Zhang, Zhenzhao Chen, and Jiajun Chen. 2013. Chinese comparative sentence identification based on the combination of rules and statistics. In *International Conference on Advanced Data Mining and Applications*, pages 300–310. Springer.

Yan Liu. 2004. *The Semantic and Cognitive Basis of Comparative Categories in Modern Chinese*. Xuelin Press.

Yihuan Liu. 2020. *A Study on Ellipsis Phenomena Based on Abstract Meaning Representation*. Nanjing Normal University.

Yuehua Liu, Wenyu Pan, and Wei Gu. 2001. *Practical modern Chinese grammar*. Beijing: The Commercial Press.

Shuxiang Lu. 1942. *Esserta of Chinese Grammar*. The Commercial Press.

Shuxiang Lu. 1979. *Problems in the Analysis of Chinese Grammar*. The Commercial Press.

Jianzhong Ma. 1929. *Ma's Grammar (Ma Shi Wen Tong)*, volume 1. The Commercial Press.

Zhen Ma. 1986. An exploratory study on the substitution patterns of comparative item y in bi-sentences. *Studies of the Chinese Language*, (2):97–105.

Fangfang Niu. 2010. A cognitive analysis of why "xiao wang's grandfather is more robust than xiao li's" cannot be said—also questioning shen jiaxuan's "metonymy and reference transfer". *Modern Chinese: Linguistic Studies*, (6):4–6.

Youri Peskine, Damir Korenčić, Ivan Grubisic, Paolo Papotti, Raphael Troncy, and Paolo Rosso. 2023. Definitions matter: Guiding gpt for multi-label classification. In *EMNLP 2023, Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Jialu Qi, Yanqiu Shao, Wei Li, and Zizhuo Shen. 2022. Mcer: A multi-domain dataset for sentence-level chinese ellipsis resolution. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 29–42. Springer.

Haibo Ren. 1987. Types of conclusion items in modern chinese bi-sentences. *Language Teaching and Linguistic Studies*, 4:91–103.

Daniele Testa, Elisa Chersoni, and Alessandro Lenci. 2023. We understand elliptical sentences, and language models should too: A new dataset for studying ellipsis and its interaction with thematic fit. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3340–3353.

Li Wang. 1944. *Collected Works of Wang Li, Volume 2: Modern Chinese Grammar*. Shandong Education Press.

Suge Wang, Hongxia Li, and Xiaolei Song. 2010. Automatic semantic role labeling for chinese comparative sentences based on hybrid patterns. In *2010 International Conference on Artificial Intelligence and Computational Intelligence*, volume 1, pages 378–382.

Wei Wang, Tiejun Zhao, Bing Xu, and Dequan Zheng. 2015. Automatic recognition of chinese comparative sentences. *Intelligent Computer and Applications*, 5(5):1–3.

Yingfei Wu. 2020. *Research and Teaching of Atypical Chinese Differential Comparative Sentences*. East China Normal University.

Han Yan, Yiran Zhao, Peipei Sun, and Yanqiu Shao. 2022. A study on the recovery of omitted constituents in chinese elliptical sentences. In *Workshop on Chinese Lexical Semantics*, pages 253–264. Springer.

Linyi Yang, Shuibai Zhang, Libo Qin, Yafu Li, Yidong Wang, Hanmeng Liu, Jindong Wang, Xing Xie, and Yue Zhang. 2023. Glue-x: Evaluating natural language understanding models from an out-of-distribution generalization perspective. In *Findings of the association for computational linguistics: ACL 2023*, pages 12731–12750.

Pengfei Zhou, Weiguang Qv, Tingxin Wei, Junsheng Zhou, Bin Li, and Yanhui Gu. 2023. A study on identification and completion of comparative sentence structures with ellipsis phenomenon. In *Proceedings of the 22nd Chinese National Conference on Computational Linguistics*, pages 388–398.

Dexi Zhu. 1982. *Lectures on Grammar*. The Commercial Press.

A Prompt for Comparative Argument Identification and Recovery

You are an expert in Chinese linguistics and syntactic analysis. Your core capability is to identify the comparative arguments and, based on context and syntactic rules, recover omitted elements so that the sentence structure is complete and symmetrical.

Core Task: Your task is to analyze a given sentence containing the marker “bi” (a *bi*-comparative sentence) and complete the following steps:

comparative arguments identification: Precisely identify the surface-level expressions in the original sentence that function as ARG0 and ARG1.

Deep structure recovery: In accordance with the core principles, recover omitted elements and produce a semantically complete and syntactically parallel recovered ARG0 and ARG1.

Core Principles: During recovery, you must strictly adhere to the following principles: Syntactic parallelism: This is the highest criterion for judging omission and performing recovery. The recovered ARG0 and ARG1 should be as symmetrical and consistent as possible in syntactic structure. Semantic consistency: The recovered elements must conform to the logic and intended meaning of the original sentence. Constituent uniqueness: Each comparative argument may have only one head. If multiple elements are involved, they must stand in a modifier–head relation; adverbials that modify the entire clause must not be conflated with noun phrases as a single constituent.

Detailed Rules:

ARG0 recovery: When the surface form before "bi" lacks an explicit comparative subject, that subject is usually the subject or topic of the preceding context or the main clause. You must identify this omitted core topic from the context and recover it.

ARG1 recovery: This is the most common omission type. The ARG1 often retains only the modifier that differs from the ARG0 (such as time or location), while the shared core noun phrase is omitted. You should copy the core noun phrase from the ARG0 and insert it into the ARG1 to achieve syntactic parallelism.

Compound omission handling: When whole ARG0 omission and ARG1 head omission occur simultaneously, you must first recover the ARG0, and then recover the ARG1 based on the recovered ARG0.

Special structure handling: Structures such as “year by year” or “one by one” are complete in themselves and require no recovery. Their surface forms and recovered forms should remain identical.

When selecting surface structures, you may only focus on the constituents directly before and after *bi*. If the clause subject preceding the *bi*-construction functions as the comparative ARG0, this counts as omission and should not be selected as a surface constituent; it may be added during

recovery.

Constituent consistency: Surface comparative arguments must be selected based on syntactic tree analysis, with only one head identified. If multiple elements appear, they must be in a modifier–head relation; adverbials modifying the entire sentence must not be grouped together with noun phrases as a single constituent.

Output Format: Please strictly follow the JSON format below when outputting your analysis results. Do not add any extra explanations or descriptive text.

{
"ARG0": "...",
"ARG1": "...",
"ARG0_Rec": "...",
"ARG1_Rec": "..."
}

Examples:

Example 1: Basic type – ARG1 head omission Input sentence: The trade volume in 1983 increased by more than double compared with 1978. Key point: In the ARG1 "1978", the shared head "trade volume" is omitted. Output:

{
"ARG0": "1983年的贸易额'trade volume in 1983'",
"ARG1": "1978年'1978'",
"ARG0_Rec": "1983年的贸易额'trade volume in 1983'",
"ARG1_Rec": "1978年的贸易额'trade volume in 1978'"
}

Example 2: Compound type – whole-ARG0 omission + ARG1 head omission Input sentence: In 2000, fixed assets reached 259 million yuan, increasing by 12.27 times compared with 1993. Key point: The ARG0 is wholly omitted and must be recovered from the subject "fixed assets in 2000" in the preceding clause. The ARG1 omits the head "fixed assets" shared with the recovered ARG0. Output:

{
"ARG0": "",
"ARG1": "1993年'1993'",
"ARG0_Rec": "2000年固定资产'fixed assets in 2000'",
"ARG1_Rec": "1993年固定资产'fixed assets in 1993'"
}

Example 3: Logical inference type – implicit current-time reference omission Input sentence: Afforestation reached 108,000 mu, increasing by three times compared with 1969. Key point: The ARG1 refers to a past year; logically, the ARG0 should be "the current year", which must be inferred and recovered as "this year". Output:

{
"ARG0": "",
"ARG1": "一九六九年'1969'",
"ARG0_Rec": "今年植树造林'afforestation this year'",
"ARG1_Rec": "一九六九年植树造林'afforestation in 1969'"
}

Example 4: Special structure type – incremental "X compared with X" structure, no recovery needed Input sentence: The political situation is stable and the economy is developing; conditions are improving year by year. Key point: "year by year" is a fixed, complete structure and requires no recovery. Output:

{
"ARG0": "一年'year'",
"ARG1": "一年'year'",
"ARG0_Rec": "一年'year'",
"ARG1_Rec": "一年'year'"
}

Example 5: Special structure type – negative comparison, no recovery needed Input sentence: Such essays generally carry substantial weight and are not comparable to ordinary light songs. Key point: The "not comparable to" construction is a complete negative comparison and requires no constituent recovery. Output:

{
"ARG0": "这样的杂文'such essays'",
"ARG1": "一般轻歌小曲'ordinary light songs'",
"ARG0_Rec": "这样的杂文'such essays'",
"ARG1_Rec": "一般轻歌小曲'ordinary light songs'"
}

Example 6: Syntactic function type – comparative construction as an attributive modifier Input sentence: We have already achieved the ambitious goal of doubling the gross national product compared with 1980. Key point: The entire comparative phrase functions as an attributive modifier of "ambitious goal". During annotation, the focus is on recovering internal parallelism. Output:

{
"ARG0": "国民生产总值'gross national product'",
"ARG1": "1980年'1980'",

"ARG0_Rec": "国民生产总值'gross national product'",
"ARG1_Rec": "1980年国民生产总值'gross national product in 1980'"
}

Example 7: Advanced inference type – nominalization of the comparison dimension Input sentence: Its distinguishing feature is that it can be deployed to depths of about 500 meters to explore the seabed, about twice as deep as ordinary underwater television equipment. Key point: What is compared is not the entities themselves but one of their properties (depth). The adjective "deep" in the comparison result must be nominalized as "deployable depth" to achieve true semantic parallelism. Output:
{
"ARG0": "",
"ARG1": "一般海底电视机'ordinary underwater television equipment'",
"ARG0_Rec": "它能够放到的水深'the depth it can be deployed to'",
"ARG1_Rec": "一般海底电视机能够放到的水深'the depth ordinary underwater television equipment can be deployed to'"
}

Example 8: Constituent consistency type – judgment of comparative arguments Input sentence: Last year, the actual freight volume completed by this county increased by 1.2 times compared with 1959. Key point: "last year" modifies the entire sentence and is neither in a modifier relation with nor in the same subtree as "the actual freight volume completed by this county", so it cannot be treated as part of the same constituent. Therefore, the identified ARG0 is only "the actual freight volume completed by this county". However, "last year" is a necessary element for semantic symmetry with the ARG1 and must be recovered as a modifier of the head in the recovered form. Output:
{
"ARG0": "这个县实际完成的运输量'the actual freight volume completed by this county'",
"ARG1": "一九五九年'1959'",
"ARG0_Rec": "去年这个县实际完成的运输量'last year's actual freight volume completed by this county'",
"ARG1_Rec": "一九五九年这个县实际完成的运输量'the actual freight volume completed by this county in 1959'"
}

Example 9: Attributive type – identification of attributive structures Input sentence: A Soviet aircraft designer said that this year a jet aircraft superior to the Tu-104 will take to the skies. Key point: The comparative structure inside the attributive clause must be separated, and omitted comparative arguments must be recovered based on the structure of the attributive clause. Output:
{
"ARG0": "",
"ARG1": ""图—104" 'Tu-104'",
"ARG0_Rec": "一种喷气式客机'a jet aircraft'",
"ARG1_Rec": ""图—104" 'the Tu-104'"
}

Start Evaluation: Now, strictly follow all the rules and formatting requirements above to analyze the following sentence: {user_input}

B Prompt for Comparison Dimension Classification

Role: You are a semantic expert in Chinese comparative sentences. Given a comparative sentence, you need to determine its comparison dimension by understanding its deep semantic structure. You may choose only one label from the following five options as the output: quantity, measure, monetary_value, metric, quality.

Dimension Descriptions and Examples:

quantity This dimension compares countable entities, typically measured using discrete counting units. Examples: "The number of participants this year is three hundred more than last year." "The number of hospitals in this region has doubled compared with five years ago."

measure This dimension compares uncountable measures or magnitudes, usually measured using conventionally defined units, such as length, time, volume, or weight. Examples: "The flight time is half an hour shorter than before." "Steel output decreased by 30% compared with last year." "The physical effort exerted by this athlete is greater than that of ordinary people."

monetary_value This dimension compares money or monetary value, usually measured in currency units. It can be regarded as a special subtype of measure, but when applicable, monetary_value should be selected instead of measure. Examples: "Per capita income increased by 0.20 compared with last year." "The project cost is much lower than the original plan."

metric This dimension compares composite indicators derived from multiple basic indicators

through formulas, calculations, or tests, such as ratios, efficiency, scores, GDP, or GPA. Examples: "The athlete's performance has improved compared with last month." "The efficiency of the newly manufactured ball-center valve is much higher than that of ordinary valves."

quality This dimension compares subjective or abstract attributes that usually lack clearly defined metrics and are based on evaluation, impression, or perception. Examples: "The current situation is more severe than in the past." "People's quality of life is improving day by day."

Decision Strategy: When determining the comparison dimension, you should first identify the comparison focus, that is, what is being compared, and then determine the typical measurement method used for that focus, which corresponds to the comparison dimension.

Output Requirements: You may provide an analysis process, but the final result must be output in JSON format as follows:

```
{
"original_sentence": "",
"comparison_dimension": ""
}
```

Start Evaluation: Now, strictly follow all the rules and formatting requirements above to analyze the following sentence: {user_input}