Can Large Language Models Understand Internet Buzzwords Through User-Generated Content

Chen Huang[♠♣] **Junkai Luo**[♠] **Xinzuo Wang**[♢] **Wenqiang Lei**[♠♣*] **Jiancheng Lv**[♠♣]
[♠]College of Computer Science, Sichuan University, China
[♣]Engineering Research Center of Machine Learning and Industry Intelligence, Ministry of Education, China [♢]JD.com, China
huangc.scu@gmail.com, luojunkai@stu.scu.edu.cn, wangxinzuo@jd.com
{wenqianglei, lvjiancheng}@scu.edu.cn

Abstract

The massive user-generated content (UGC) available in Chinese social media is giving rise to the possibility of studying internet buzzwords. In this paper, we study if large language models (LLMs) can generate accurate definitions for these buzzwords based on UGC as examples. Our work serves a threefold contribution. First, we introduce CHEER, the first dataset of Chinese internet buzzwords, each annotated with a definition and relevant UGC. Second, we propose a novel method, called RESS, to effectively steer the comprehending process of LLMs to produce more accurate buzzword definitions, mirroring the skills of human language learning. Third, with CHEER, we benchmark the strengths and weaknesses of various off-the-shelf definition generation methods and our RESS. Our benchmark demonstrates the effectiveness of RESS while revealing crucial shared challenges: over-reliance on prior exposure, underdeveloped inferential abilities, and difficulty identifying high-quality UGC to facilitate comprehension. We believe our work lays the groundwork for future advancements in LLM-based definition generation. Our dataset and code are available at https://github.com/SCUNLP/Buzzword.

1 Introduction

Internet buzzwords emerge as newly coined terms representing abstract concepts that may extend beyond their literal definitions (Malyuga and Rimmer, 2021), such as '窝囊费'(i.e., *gutlessness fee*). They often rapidly gain popularity through social media platforms and are amplified by user-generated content (UGC) such as posts and reviews (Liu et al., 2020). However, their inherent abstractness creates ambiguity, presenting significant challenges for human understanding without further explanation (Tsur and Rappoport, 2015; Cornwall, 2007; Malyuga and Rimmer, 2021; Huang et al., 2022).

Figure 1: Task Illustration: generating definitions for Chinese buzzwords using UGC.

For instance, as illustrated in Figure 1, the Chinese buzzword '窝囊费' is commonly used to reflect user dissatisfaction with inadequate compensation for their efforts, highlighting a feeling of being undervalued rather than the literal meaning of '窝囊', which implies cowardice. Given the absence of these terms in traditional dictionaries, online UGC becomes a crucial resource for understanding their meaning. Therefore, **our task, as shown in Figure 1, is to generate definitions for internet buzzwords using UGC as illustrative examples**, bridging the gap between their widespread usage and a precise understanding of their meaning.

Formally, our task falls under the category of context-aware definition generation (Li et al., 2020; Huang et al., 2021), which involves automatically generating dictionary definitions by incorporating both the target word and contextual information, such as example sentences. While existing methods effectively learn to define common words using static training data (Mickus et al., 2019; Zheng et al., 2021; Huang et al., 2021; Zhang et al., 2023), they struggle with the dynamic nature of online buzzwords, which emerge and disappear rapidly.

*Correspondence to Wenqiang Lei.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 12916–12941
July 27 - August 1, 2025

Even advanced large language models (LLMs) exhibit limitations with long-tail words (Wu et al., 2024; Rodríguez-Betancourt and Casasola-Murillo, 2023), let alone newly coined buzzwords. Therefore, **effectively comprehending buzzwords remains a significant challenge**.

To substantiate our analysis, we construct a dataset of CHinEsE internet buzzwoRds, called **CHEER**, and evaluate the performance of existing methods (cf. Section 3). The evaluation results clearly demonstrate that existing methods, including LLM-based ones, struggle to accurately define buzzwords. Therefore, leveraging UGC to imbue LLMs with an understanding of rapidly evolving buzzwords poses a significant challenge.

To tackle the challenge, we propose **RESS** to steeR the comprEhending proceSS of LLMs to produce buzzword definitions. By mimicking human language acquisition, RESS codifies key comprehension skills into distinct aspects, prompting the LLM to produce aspect-specific definition candidates. These candidates are then integrated through an ensemble process for a more nuanced and accurate final definition, enriching the LLM's understanding towards internet buzzwords.

To thoroughly evaluate the efficacy and limitations of existing methods and RESS across various LLM backbones, we establish a benchmark using our curated dataset CHEER. Our results demonstrate that RESS surpasses baseline performance, achieving an average improvement of +2.51% in semantic accuracy and +3.31% in semantic completeness[1] over the best baseline. Despite these gains, the overall performance of all methods remains suboptimal. Further analysis reveals key challenges LLMs face in interpreting buzzwords derived from UGC. Specifically, we identify an overreliance on prior exposure to buzzwords[2] and a limited capacity to infer the meaning of buzzwords, hindering their ability to handle the dynamic nature of internet buzzwords. Additionally, both the volume and quality of UGC are critical factors influencing definition accuracy. However, obtaining sufficient high-quality UGC without prior knowledge of buzzword meanings remains a significant challenge and warrants further research. We conclude our main contributions as follows.

- We call attention to the importance of generating definitions for internet buzzwords using UGC, a task of interest in socio- and psycholinguistics for understanding the dynamics of online language.
- We introduce CHEER, the first Chinese buzzword dataset of its kind, comprising over 1K entries, each including a definition and a corresponding set of UGC exemplifying contemporary usage.
- We propose a simple yet effective method, called RESS, to effectively steer the comprehending process of LLMs to produce buzzword definitions, mirroring the skills of human language learning.
- Using CHEER, we benchmark existing methods and our RESS for buzzword definition generation. Results demonstrate the effectiveness of RESS while revealing a crucial shared challenge: comprehending unseen buzzwords and leveraging sufficient, high-quality UGC to facilitate this comprehension. This benchmark underscores the need for further research in these critical areas.

[1] Semantic completeness refers to a definition encompassing all and only the relevant aspects of a word's meaning.

[2] Performance is significantly reduced on unseen buzzwords compared to seen ones.

2 Related Work

Buzzwords Understanding. While the analysis of buzzwords holds significant interest from a purely linguistic (socio/psycho-) perspective (Fiasco and Massarella, 2022; Qian et al., 2023; Mei et al., 2024), however, their inherent abstractness create their inherent ambiguity, posing substantial challenges for both Natural Language Processing (NLP) systems and human comprehension (Tsur and Rappoport, 2015; Cornwall, 2007; Malyuga and Rimmer, 2021). The rapid emergence of new buzzwords and their absence from traditional dictionaries further compound these challenges, making it difficult to interpret their meaning without additional context (Huang et al., 2022). To this end, we closely revolve around Chinese buzzwords and, for the first time, introduce a novel method for automatically understanding and generating definitions based on UGC as example sentences. Furthermore, we present dataset CHEER to benchmark existing methods and illuminate the challenges inherent in buzzword definition generation.

Context-aware Definition Generation. Definition generation aims to automatically generate dictionary definitions for words (Noraset et al., 2017; Yin and Skiena, 2023), assisting the construction of dictionaries. Unlike non-context definition generation methods that rely solely on the word itself (Zheng et al., 2021; Yang et al., 2020), context-aware definition generation methods incorporate additional context information, such as example sentences

# Buzzwords	1127.0
# UGC (Example Sentences)	34607.0
Avg. #examples per buzzword	30.7
Avg. length of description per buzzword	262.5
Avg. length of definition per buzzword	50.0
Avg. length of examples per buzzword	85.4

Table 1: Data statistics of CHEER

(Ishiwatari et al., 2019; Li et al., 2020; Huang et al., 2021; Mei et al., 2024) and definitional information (Huang et al., 2022). This contextual input assists in disambiguating word senses, leading to more accurate and nuanced definitions. However, the rapid evolution of buzzwords limits the effectiveness of existing methods, even those specifically designed for unfamiliar words and slang (Ishiwatari et al., 2019; Pei et al., 2019; Sun et al., 2022; Mei et al., 2024). This limitation stems from their reliance on static training datasets and rote memorization of definitions (cf. Section 5.2).

3 Benchmark on Definition Generation for Internet Buzzwords

We introduce a benchmark to analyze limitations of existing definition generation methods and highlight their inability to handle Chinese buzzwords.

3.1 CHEER: Chinese Buzzword Dataset

We present the first dataset of Chinese internet buzzwords, called CHEER, offering insights into their contemporary usage. Containing over 1K entries, each buzzword is meticulously profiled with its name, description, definition, and real-world UGC example sentences. Table 2 and Table 1 provide illustrative examples and data statistics, respectively. The data collection process is outlined below. For more details, refer to **Appendix** A.

- **Buzzword Collection**. We gather Chinese buzzwords from various reputable online dictionary platforms specializing in trending buzzwords (e.g., '梗百科'), and eliminated any duplicate.
- **Definition Collection**. We gather descriptions for each buzzword by scraping those platforms, typically including its origin, cultural references, and informal colloquial explanations. We then prompt an LLM to summarize a concise definition encompassing both its literal and figurative meanings (if applicable).
- **Example Collection**. We collect UGC containing buzzwords from two popular Chinese social media platforms, Xiaohongshu[3] and Weibo[4]. This exemplifies contemporary usage.
- **Quality Control**. CHEER's quality is rigorously controlled through three vetting layers: dictionary websites, internet users, and our review process: we manually remove inappropriate buzzwords, refine definitions, and purge existing definitional information from the crawled UGC.

Internet Buzzword
0帧起手 (0 frame startup)
Description from Online Source
0帧起手指零帧技能，一般指的是点击即可释放， 并且立刻判定无法打断的技能。0帧起手在网络上表示动作极快， 没有丝毫等待，绝不拖泥带水，闪电般突然出现的动作。 ('0 frame startup' generally refers to a skill that can be released by clicking and immediately determines that it cannot be interrupted. This term, often used online, signifies lightning-fast action with no delay—a sudden strike like a bolt of lightning.)
Definition
原意是指游戏中一些无需准备时间，可以瞬间释放的技能， 引申为行动迅速，毫不拖延。 (Originally referring to in-game abilities usable without any setup time, the term has broadened to describe taking swift and immediate action)
Examples (i.e., UGC)
这就不得不说我那一放歌就会0帧起手开唱的隔壁同事了 (I'm reminded of my colleague next door who '0 frame startup' into singing every time I play music.)

Table 2: Case of buzzword '0帧起手'. For clarity, we only include a single example sentence. Here, we also provide its English translation for better understanding.

3.2 Benchmark Setup

Benchmark Overview. We require existing definition generation methods to generate definitions for each buzzword. These definitions must be derived solely from the real-world UGC sentences within CHEER, which represent how users actually use these buzzwords in their online interactions.

Baselines. 1) We consider two LM-based context-aware definition generation models tailored for Chinese (Kong et al., 2022; Song et al., 2019), including MASS-zh, a pretrained language model specialized in definition generation, and SimpDefiner (SDefiner, for short), an enhanced version of MASS-zh incorporating multi-task learning. 2) Additionally, we explore three LLM-based context-aware methods: Direct Prompt (DP) (Jhirad et al., 2023), Chain-of-Thought (CoT) (Wu et al., 2024), and FOCUS (Mei et al., 2024), which is currently considered the SOTA approach. Moreover, we include $\text{DP}_{\text{w/o UGC}}$ as a context-free baseline, which generates definitions based solely on the buzzword

[3] https://www.xiaohongshu.com
[4] https://weibo.com

Methods	BLEU	R-L	BScore	SA	SC	Methods	BLEU	R-L	BScore	SA	SC
LM-based Backbone: *MASS*						LM-based Backbone: *MASS*					
MASS-zh (Song et al., 2019)	0.40	28.5	56.67	1.02	1.01	SDefiner (Kong et al., 2022)	0.67	26.94	54.78	1.01	1.00
LLM-based Backbone: *Qwen2-7B*						LLM-based Backbone: *Qwen2-72B*					
$DP_{w/o\ UGC}$	10.27	41.49	64.77	1.89	1.55	$DP_{w/o\ UGC}$	10.87	41.58	66.13	2.07	1.65
DP (Jhirad et al., 2023)	**15.35**	**43.05**	**65.38**	2.19	1.97	DP (Jhirad et al., 2023)	19.27	**44.37**	67.58	2.71	2.45
CoT (Wu et al., 2024)	15.26	40.41	65.12	2.30	2.14	CoT (Wu et al., 2024)	**19.50**	43.49	**67.67**	2.77	2.54
FOCUS (Mei et al., 2024)	12.41	33.57	63.89	**2.39**	**2.51**	FOCUS (Mei et al., 2024)	12.09	29.81	64.75	**2.88**	**3.20**
LLM-based Backbone: *GPT-4o Mini*						LLM-based Backbone: *GPT-4o*					
$DP_{w/o\ UGC}$	7.96	38.81	65.65	1.87	1.49	$DP_{w/o\ UGC}$	9.56	39.42	66.56	2.05	1.62
DP (Jhirad et al., 2023)	15.53	**44.81**	**66.67**	2.26	1.93	DP (Jhirad et al., 2023)	17.85	**67.56**	45.22	2.50	2.13
CoT (Wu et al., 2024)	**16.64**	44.79	66.55	2.32	2.04	CoT (Wu et al., 2024)	**18.33**	44.49	**67.46**	2.60	2.30
FOCUS (Mei et al., 2024)	13.37	33.42	65.05	**2.64**	**2.67**	FOCUS (Mei et al., 2024)	15.08	35.10	66.05	**2.95**	**2.92**

Table 3: Benchmark results of off-the-shelf definition generation methods using CHEER. We highlight their limitations in handling internet buzzwords, as evidenced by the low scores for both SA and SC (1-5).

itself without UGC. For all baselines, we implemented them using their official code. More implementation can be found in **Appendix** C.

Evaluation Metrics. We employ a comprehensive evaluation framework that extends beyond conventional metrics such as BLUE, ROUGE-L (R-L, for short), and BERTScore (BScore), which have been widely used in previous research (Zheng et al., 2021; Huang et al., 2021; Li et al., 2020). In addition to these metrics, we prioritize the Semantic Accuracy (SA) and Semantic Completeness (SC) of generated definitions, as emphasized in previous studies (Li et al., 2020; Segonne, 2023). To assess these aspects, we utilize both GPT4-based evaluation, assigning a score ranging from 1 to 5. Notably, our LLM-based evaluator is equipped with detailed scoring rubrics, following established practices for enhancing evaluation reliability (Gao et al., 2024; Liu et al., 2023b). To validate our evaluation, we incorporated human evaluation using win rate to assess alignment with human judgment. Details on LLM and human evaluation are in **Appendix** E.

3.3 Evaluation Findings

Table 3 shows that incorporating additional usage examples generally improves the quality of generated definitions (as seen in the comparison between DP and $DP_{w/o\ UGC}$). Furthermore, LLM-based methods substantially outperform traditional LM-based methods on all metrics when generating buzzword definitions. However, even the best-performing method, FOCUS, achieves suboptimal semantic accuracy and completeness, with overall SA and SC scores below 3 out of 5. Therefore, current definition generation methods struggle with internet buzzwords given UGC, highlighting the need for more effective approaches.

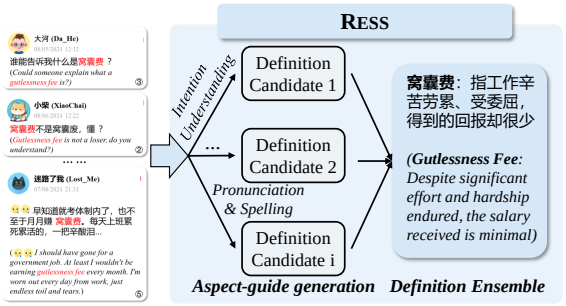


Figure 2: Illustration of RESS.

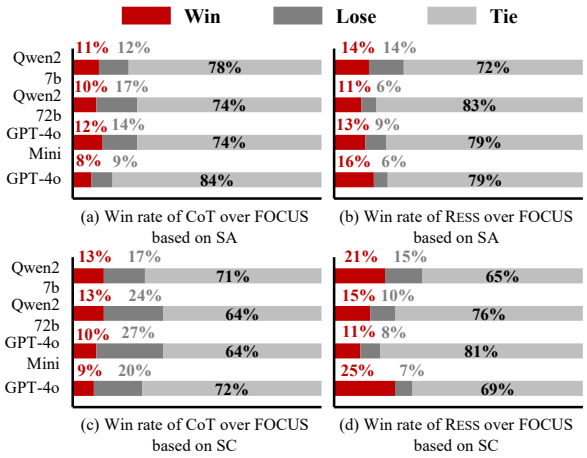
4 RESS: The Proposed Method

To enhance the performance of existing baselines, RESS leverages key skills of child learning and codifies them into illustrative aspects to guide LLM-driven buzzword definition generation. As illustrated in Figure 2, RESS generates aspect-specific definition candidates from UGC and subsequently ensembles these candidates to produce a more accurate and contextually grounded definition. See **Appendix** C.1 for implementation.

Aspect Initialization. Established skills of child language development encompass the following six aspects (Beck et al., 1982; Nagy et al., 1985; Bloom, 2000), which we incorporate into RESS: 1) *Intention Understanding (IU)*: discerning the speaker's communicative goal when using the buzzword (Bloom, 2000), such as expressing an emotion; 2) *Concept Association (CA)*: linking the buzzword to relevant concepts (e.g., associating '窝囊费' with "job") (Meylan and Bergelson, 2022; Swingley, 2010); 3) *Language Structure (LS)*: analyzing the buzzword's grammatical function (Bloom, 2000); 4) *Social Cue Interpretation (SCI)*: inferring social context from UGC such as the speaker's facial expressions, tone of voice, and gestures (Bloom, 2000); 5) *Word Context (WC)*: leveraging surrounding text for semantic disam-

Method	BLEU	R-L	Bscore	SA	SC
			Qwen2-7b		
FOCUS	**12.41**	**33.57**	**63.89**	2.39	2.51
RESS	10.87$_{\downarrow 12.41\%}$	29.09$_{\downarrow 13.35\%}$	63.33$_{\downarrow 0.86\%}$	**2.41**$_{\uparrow 0.84\%}$	**2.57**$_{\uparrow 2.39\%}$
			Qwen2-72b		
FOCUS	12.09	29.81	64.75	2.88	**3.20**
RESS	**15.74**$_{\uparrow 30.19\%}$	**35.63**$_{\uparrow 19.52\%}$	**66.41**$_{\uparrow 2.56\%}$	**2.97**$_{\uparrow 3.13\%}$	3.09$_{\downarrow 3.44\%}$
			GPT-4o Mini		
FOCUS	13.37	33.42	65.05	2.64	2.67
RESS	**14.58**$_{\uparrow 9.05\%}$	**35.43**$_{\uparrow 6.01\%}$	**65.67**$_{\uparrow 0.95\%}$	**2.72**$_{\uparrow 3.03\%}$	**2.74**$_{\uparrow 2.62\%}$
			GPT-4o		
FOCUS	15.08	35.10	66.05	2.95	2.92
RESS	**16.52**$_{\uparrow 9.55\%}$	**36.42**$_{\uparrow 3.76\%}$	**66.74**$_{\uparrow 1.04\%}$	**3.04**$_{\uparrow 3.05\%}$	**3.06**$_{\uparrow 4.79\%}$

Table 4: Overall evaluation. The performance of RESS exceeds that of FOCUS, the best baseline.



(a) Win rate of CoT over FOCUS based on SA
(b) Win rate of RESS over FOCUS based on SA
(c) Win rate of CoT over FOCUS based on SC
(d) Win rate of RESS over FOCUS based on SC

Figure 3: Human evaluation across different methods and LLM backbones via *win rate*. RESS produces better buzzword definitions from a user-centric perspective.

biguation (Ricketts et al., 2011; Horst et al., 2011); 6) *Pronunciation and Spelling (PS)*: establishing connections between orthography, phonology, and meaning (Bloom, 2000).

Aspect-guided Definition Generation & Definition Ensemble. Given these aspects, the LLM generates individual definitions based on the provided UGC for each aspect (e.g., prompt: *comprehending the meaning of buzzwords from the given aspect*). These aspect-specific definitions are then synthesized by the LLM to produce a candidate definition (e.g., prompt: *generating definition based on the given candidates from different aspects*).

5 Benchmark Evaluation

Following the setting described in Section 3, we further provide a detailed benchmark of baselines and our proposed RESS. We present the overall performance results of all methods in Section 5.1, and provide an in-depth analysis to investigate their performance characteristics in Section 5.2.

5.1 Overall Performance

We evaluate the overall performance of all methods using automatic metrics in Table 4 and Table 3[5]. Additionally, we report human evaluation in Figure 4, measuring definition quality via *win rate*. Our key observations are detailed below.

***How effective* RESS *is?* – It demonstrates superior performance, exhibiting enhanced semantic accuracy and completeness**. As illustrated in Table 4, RESS demonstrates a substantial improvement over FOCUS, the leading baseline, across various LLM backbones. We observe average gains of +9.10% for BLEU, +3.99% for R-L, +0.92% for Bscore, +2.51% for SA, and +3.31% for SC across various LLM backbones. These results underscore the advantages of our RESS.

***What is the practical utility of* RESS*?* – It produces superior buzzword definitions from a user-centric perspective**. To demonstrate the correlation between our automatic evaluation and human judgment, we conduct a human evaluation of definitions generated for 100 randomly selected buzzwords. For each buzzword, two human evaluators compared the definitions generated by different methods across various backbones, considering both SA and SC. Following Sekulić et al. (2022), the evaluators are presented with pairs of anonymized definitions for the same buzzword, without disclosure of the originating model for each definition. Independent evaluations are followed by a discussion to resolve any discrepancies. A "*Win/Lose/Tie*" label is finally assigned if consensus is reached; otherwise, the result is recorded as a "*Tie*". Due to the resource-intensive nature of human evaluation, our analysis is limited to three representative methods. As shown in Figure 3, RESS not only outperforms baselines but also maintains a performance ranking consistent with the automated metrics reported in Table 3. This confirms the reliability of our automated evaluation and its alignment with human judgment[6].

***Why is* RESS *effective?* – Leveraging multi-faceted aspects may enhance comprehension of buzzwords**. In contrast to direct prompting, our method simulates established skills of child language acquisition by incorporating explicitly codified aspects to guide definition generation. This yields aspect-specific definitions, the semantic di-

[5]See Appendix H for case studies.

[6]Refer to Appendix D for more human evaluation.

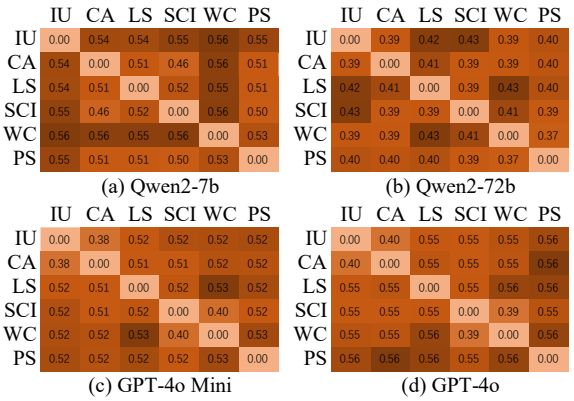


Figure 4: Semantic diversity analysis of aspect-specific definitions, measured by *1.0-Bscore*. These aspects offer a multifaceted approach to understanding buzzwords.

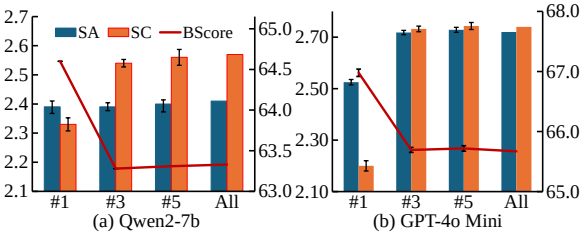


Figure 5: RESS ablation on the number of aspects. We evaluate the performance of various aspect combinations of fixed sizes (i.e., 1, 3, 5) and report their mean and standard deviation. Employing an ensemble of aspects frequently demonstrates advantages.

versity of which is explored in Figure 4. The figure demonstrates varying degrees of semantic diversity across aspect-derived definitions, exhibiting a weak correlation in terms of Bscore. This suggests that different aspects guide the LLM towards distinct perspectives on the interpretation of buzzwords, contributing to a more comprehensive understanding. Building on this observation, we investigated the impact of varying the number of aspects employed during definition generation (results shown in Figure 5). Preliminary findings indicate a positive correlation between the number of aspects and the quality of the generated definitions. While this suggests a degree of scalability inherent in RESS, further research is needed to determine whether the introduction of more aspects continues to enhance the performance. This represents a promising avenue for future investigation. Our work focuses on establishing a comprehensive benchmark and analyzing the capabilities and limitations of LLMs for generating definitions of buzzwords. Consequently, a more exhaustive exploration of the scalability of our RESS is deferred to future work.

Summary. According to the results of our benchmark, the performance of RESS exceeds that of alternative baselines[7]. Notwithstanding this advantage, and despite incorporating various aspects of word comprehension through the simulation of human learning processes, the overall performance of RESS (and baselines) still offers room for improvement. This discrepancy motivates a deeper investigation and exploration, which we outline in the following section.

5.2 In-depth Benchmark Analysis

This section presents a comparative performance analysis of leading LLM-based methods, with a particular focus on analyzing the challenges LLMs face in interpreting buzzwords derived from UGC. This analysis considers two key aspects: (1) the LLMs' capacity for inferring the meaning of buzzwords, and (2) the influence of the specific UGC employed on LLM comprehension.

5.2.1 Investigation on LLM Inference Capacity of Buzzword Meaning

In this subsection, we design a contamination-free evaluation to specifically measure the LLM capacity for buzzword meaning inference, excluding the influence of potentially memorized definitions from the training corpus. Our experimental setup is described below, with results presented in Table 5.

Contamination-free Evaluation Setup. To avoid the bias introduced by potential data leakage (Jain et al., 2024)—where LLMs may have already learned buzzwords and their meanings during training—we employ a contamination-free evaluation. This involves using *unseen* buzzwords that emerged after the LLM's release date, ensuring they were not present in the training data. However, pinpointing the precise origin date and first appearance of a buzzword is exceptionally challenging. In response, we input each buzzword into each LLM without providing UGC examples (i.e., $\text{DP}_{\text{w/o UGC}}$) to assess pre-existing knowledge of its meaning. This allows us to create LLM-specific sets of unseen buzzwords for a more practical contamination-free evaluation. In particular, for each LLM backbone, we divided our dataset into buzzwords with known definitions for that LLM and truly unseen buzzwords. While the "contaminated" buzzwords vary across LLM backbones, the contamination status for a given buzzword is consistent across all methods evaluated with the same LLM backbone. See **Appendix** B for details.

[7]See Appendix F for more characteristics of RESS.

Methods	Back-bone	Contamination Evaluation					Contamination-free Evaluation				
		BLEU	R-L	BScore	SA	SC	BLEU	R-L	BScore	SA	SC
LM-based Methods											
MASS-zh	MASS	–	–	–	–	–	0.40	28.50	56.67	1.02	1.01
SDefiner		–	–	–	–	–	0.67	26.94	54.78	1.01	1.00
LLM-based Methods											
$DP_{w/o\ UGC}$	Qwen2 7b	15.33	43.12	69.38	3.11	2.26	8.67$_{\downarrow 43.44\%}$	40.97$_{\downarrow 4.99\%}$	63.31$_{\downarrow 8.75\%}$	1.50$_{\downarrow 51.77\%}$	1.33$_{\downarrow 41.15\%}$
DP		**19.61**	**44.91**	**68.11**	2.94	2.55	14.00$_{\downarrow 28.61\%}$	**42.46**$_{\downarrow 5.46\%}$	**64.52**$_{\downarrow 5.27\%}$	1.96$_{\downarrow 33.33\%}$	1.78$_{\downarrow 30.20\%}$
CoT		18.67	42.08	67.21	3.05	2.75	**14.18**$_{\downarrow 24.05\%}$	39.88$_{\downarrow 5.23\%}$	64.46$_{\downarrow 4.09\%}$	2.06$_{\downarrow 32.46\%}$	1.95$_{\downarrow 29.09\%}$
FOCUS		13.81	34.03	65.47	3.22	3.31	11.96$_{\downarrow 13.40\%}$	33.42$_{\downarrow 1.79\%}$	63.39$_{\downarrow 3.18\%}$	**2.12**$_{\downarrow 34.16\%}$	2.26$_{\downarrow 31.72\%}$
RESS		13.33	30.89	65.76	**3.35**	**3.45**	10.10$_{\downarrow 24.23\%}$	28.52$_{\downarrow 7.67\%}$	62.56$_{\downarrow 4.87\%}$	2.11$_{\downarrow 37.01\%}$	**2.29**$_{\downarrow 33.62\%}$
$DP_{w/o\ UGC}$	Qwen2 72b	15.91	43.90	**70.08**	3.11	2.24	8.37$_{\downarrow 47.39\%}$	40.43$_{\downarrow 7.90\%}$	64.17$_{\downarrow 8.43\%}$	1.55$_{\downarrow 50.16\%}$	1.36$_{\downarrow 39.29\%}$
DP		**22.31**	**45.94**	69.65	3.27	2.88	17.77$_{\downarrow 20.35\%}$	**43.50**$_{\downarrow 5.31\%}$	66.56$_{\downarrow 4.44\%}$	2.44$_{\downarrow 25.38\%}$	2.23$_{\downarrow 29.09\%}$
CoT		22.14	44.46	69.57	3.34	2.98	**18.20**$_{\downarrow 17.80\%}$	43.02$_{\downarrow 3.24\%}$	**66.72**$_{\downarrow 4.10\%}$	2.49$_{\downarrow 25.45\%}$	2.32$_{\downarrow 22.15\%}$
FOCUS		12.90	29.79	65.98	3.50	**3.82**	11.70$_{\downarrow 9.30\%}$	29.82$_{\uparrow 0.10\%}$	64.14$_{\downarrow 2.79\%}$	2.58$_{\downarrow 26.29\%}$	**2.90**$_{\downarrow 24.08\%}$
RESS		17.24	36.07	67.86	**3.57**	3.68	15.00$_{\downarrow 12.99\%}$	35.41$_{\downarrow 1.80\%}$	65.69$_{\downarrow 3.20\%}$	**2.67**$_{\downarrow 25.21\%}$	2.79$_{\downarrow 24.18\%}$
$DP_{w/o\ UGC}$	GPT-4o Mini	13.45	41.72	**70.69**	3.10	2.18	6.36$_{\downarrow 52.71\%}$	37.96$_{\downarrow 9.01\%}$	64.19$_{\downarrow 9.20\%}$	1.51$_{\downarrow 51.29\%}$	1.29$_{\downarrow 40.83\%}$
DP		19.93	46.30	69.90	3.07	2.51	14.25$_{\downarrow 28.50\%}$	**44.38**$_{\downarrow 4.15\%}$	**65.73**$_{\downarrow 5.97\%}$	2.03$_{\downarrow 33.88\%}$	1.76$_{\downarrow 29.88\%}$
CoT		**21.00**	**46.53**	69.54	3.13	2.64	**15.38**$_{\downarrow 26.76\%}$	44.28$_{\downarrow 4.84\%}$	65.68$_{\downarrow 5.55\%}$	2.09$_{\downarrow 33.23\%}$	1.87$_{\downarrow 29.17\%}$
FOCUS		14.94	33.71	66.85	3.52	3.47	12.92$_{\downarrow 13.52\%}$	33.33$_{\downarrow 1.13\%}$	64.53$_{\downarrow 3.47\%}$	2.38$_{\downarrow 32.39\%}$	2.44$_{\downarrow 29.68\%}$
RESS		16.53	36.06	67.77	**3.60**	**3.60**	14.01$_{\downarrow 15.25\%}$	35.25$_{\downarrow 2.25\%}$	65.06$_{\downarrow 4.00\%}$	**2.46**$_{\downarrow 31.67\%}$	**2.50**$_{\downarrow 30.56\%}$
$DP_{w/o\ UGC}$	GPT-4o	15.45	43.29	**71.09**	3.11	2.21	6.87$_{\downarrow 55.53\%}$	37.65$_{\downarrow 13.03\%}$	64.49$_{\downarrow 9.28\%}$	1.55$_{\downarrow 50.16\%}$	1.36$_{\downarrow 38.46\%}$
DP		21.13	**46.72**	70.02	3.14	2.58	16.35$_{\downarrow 22.62\%}$	**44.54**$_{\downarrow 4.67\%}$	**66.44**$_{\downarrow 5.11\%}$	2.21$_{\downarrow 33.23\%}$	1.92$_{\downarrow 29.17\%}$
CoT		**21.49**	45.49	69.71	3.22	2.80	**16.88**$_{\downarrow 21.45\%}$	44.03$_{\downarrow 3.21\%}$	66.43$_{\downarrow 4.71\%}$	2.32$_{\downarrow 27.95\%}$	2.07$_{\downarrow 26.07\%}$
FOCUS		16.33	34.94	67.54	3.60	3.69	14.51$_{\downarrow 11.15\%}$	35.17$_{\uparrow 0.66\%}$	65.37$_{\downarrow 3.21\%}$	2.64$_{\downarrow 26.67\%}$	2.61$_{\downarrow 29.27\%}$
RESS		17.87	36.03	68.41	**3.75**	**3.80**	15.90$_{\downarrow 11.02\%}$	36.61$_{\uparrow 1.61\%}$	65.98$_{\downarrow 3.55\%}$	**2.71**$_{\downarrow 27.73\%}$	**2.72**$_{\downarrow 28.42\%}$

Table 5: Contamination-free evaluation for off-the-shelf definition generation methods using CHEER. The contamination evaluation for LM-based methods is empty as their backbone lacks prior knowledge of buzzword definition.

***How effectively can LLMs infer the meaning of buzzwords based on UGC?* – Their performance is limited by over-reliance on prior exposure and underdeveloped inferential abilities for unseen buzzwords**. Table 5 reveals a notable performance degradation across all methods and LLM backbones when evaluated on unseen buzzwords (cf., *contamination-free evaluation* columns). This suggests that existing methods may rely on prior exposure to these buzzwords during training, rather than possessing a strong inference capability for unseen buzzwords. Furthermore, the results of contamination-free evaluation indicate that LLM-based methods do outperform LM-based methods. A positive correlation between model size and inferential ability is also observed within the same LLM family (e.g., Qwen and GPT). These observations align with research on child language acquisition, which links vocabulary size and reading comprehension skills with the ability to infer the meaning of unseen words from context (Ricketts et al., 2011; Swanborn and de Glopper, 2002; Cain et al., 2004): 1) A larger vocabulary equips children with a richer contextual understanding, facilitating the interpretation of new words within that context. This partially explains the superior performance of LLM-based methods compared to LM-based ones. 2) However, reading comprehension can be a stronger predictor. Within the same LLM family (with the same vocabulary size), a larger model size correlates with a stronger ability to infer the meaning of unseen buzzwords. However, the low evaluation scores highlight limitations of current LLMs: their overreliance on prior exposure to buzzwords and limited capacity to infer definitions from UGC context hinder their ability to handle the dynamic and evolving nature of buzzwords.

5.2.2 Investigation on UGC Impact

While sufficient high-quality examples can facilitate word understanding (Kilgarriff et al., 2008; Benedetti et al., 2024), UGC informativeness varies considerably. For example, as shown in Figure 1 ④, the post '*What is the gutlessness fee?*' provides no insight into the term's meaning. This contrasts with the assumption of existing methods, which often rely on carefully curated example sentences (Jhirad et al., 2023; Mei et al., 2024), potentially resulting in inferior performance when applied to raw UGC. Therefore, we perform a comprehensive examination to investigate the impact of UGC on buzzword definition generation in terms of both the UGC size and quality. To achieve this, we consider

Method	BLEU	R-L	Bscore	SA	SC	Method	BLEU	R-L	Bscore	SA	SC
DP (All)	**15.53**	**44.81**	**66.67**	**2.26**	**1.93**	FOCUS (All)	**13.37**	33.42	**65.05**	**2.64**	**2.67**
-w Random	14.62	42.89	65.32	2.13	1.88	-w Random	12.97	**35.61**	63.94	2.28	2.30
-w GDEX	11.49	43.14	64.74	1.81	1.57	-w GDEX	12.97	33.54	64.63	2.47	2.49
-w WAUS	14.57	44.29	66.33	2.19	1.86	-w WAUS	12.98	33.22	64.83	2.58	2.63
CoT (All)	**16.64**	**44.79**	**66.55**	**2.32**	2.04	RESS (All)	14.58	35.43	**65.67**	**2.72**	**2.74**
-w Random	15.38	41.58	65.23	2.23	**2.05**	-w Random	14.56	36.14	65.52	2.64	2.62
-w GDEX	13.70	44.19	65.33	1.90	1.65	-w GDEX	14.05	**37.00**	64.63	2.21	2.13
-w WAUS	15.57	44.55	66.31	2.25	1.94	-w WAUS	**14.76**	36.00	65.58	2.66	2.66

Table 6: Analysis on impact of UGC quality using GPT-4o Mini as the backbone. Each sentence selection method identifies ten UGC instances, used as input for definition generation. While utilizing higher-quality UGC generally improves definition quality (cf. WAUS), accurately identifying such instances without prior knowledge of the buzzword's meaning presents a significant challenge.

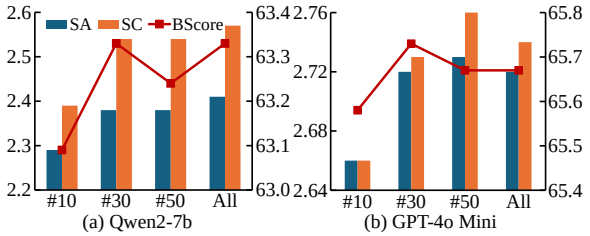


Figure 6: Analysis on the volume of UGC (x-axis). We testify RESS -w WAUS with GPT-4o Mini as backbone. Increasing the amount of UGC often shows beneficial.

the following UGC selection methods[8]:

- **All**. It uses all available UGC for each buzzword.
- **Random**. It utilizes a fixed number of randomly sampled UGC instances per buzzword as input for definition generation methods.
- **GDEX** (Kilgarriff et al., 2008). It is a well-established rule-based method for dictionary example selection to select high-quality UGC, for example, prioritizing sentences of appropriate length and avoiding the use of uncommon words.
- **WAUS**. To conduct more comprehensive evaluation, we propose a Word-meaning Agnostic UGC Selector based on BERT. It employs a masked training strategy to identify high-quality UGC. Critically, given the initially unknown semantics of target buzzwords, masking these words from training data forces WAUS to prioritize contextual and syntactic information, thereby learning patterns independent of specific word meanings and bypassing handcrafted rules used in GDEX. Refer to Appendix C.2 for implementation.

What is the impact of UGC on the accuracy of LLM-generated definitions for buzzwords? **– Both the volume and quality of UGC are crucial factors**. 1) As shown in Table 6, when controlling for the *volume* of UGC, both evaluated sentence selection methods (WAUS and GDEX) often can outperform random selection. WAUS tends to select higher-quality UGC than GDEX, leading to improved LLM-generated definitions. However, the inherent difficulty of identifying truly high-quality UGC without prior knowledge of the buzzword's meaning is evident, as even WAUS does not consistently outperform random selection. This difficulty may stem from the challenge of assessing a sentence's relevance and disambiguating its meaning when the target word's definition is unknown. Relying solely on contextual and syntactic information is insufficient to determine whether a sentence exemplifies the intended meaning or distinguishes between different senses of the buzzword. Future work could explore a self-training approach where definition candidates are generated from an initial set of selected UGC, and these candidate definitions are then used to refine the UGC selection process in an iterative manner. 2) When controlling for the *quality* selection method, as illustrated in Figure 6, increasing UGC volume generally improves performance. Interestingly, with high-quality UGC, a subset of 50 instances can potentially outperform the entire dataset (Figure 6(b)). These findings underscore the importance of both UGC volume and quality for accurate definition generation.

6 Conclusion & Discussion

This paper investigates if LLMs can effectively learn to understand Chinese buzzwords through UGC. Our work stands out as a valuable resource with threefold contributions: First, we introduce CHEER, the first publicly available dataset of Chinese internet buzzwords. Second, we propose RESS, a novel method for guiding LLMs to gen-

[8]Since buzzword definitions are initially unknown, precluding most example selection methods, we consider two simpler approaches to evaluate UGC impact. See Appendix G for details on example selection methods.

erate accurate buzzword definitions. Third, using CHEER, we benchmark existing methods and RESS to identify their strengths and weaknesses. From a broader perspective, our work bridges the fields of linguistics and artificial intelligence, fostering a deeper understanding of online language dynamics and informing the development of more robust language comprehension models (Miao et al., 2024; Saba, 2024; Cai et al., 2025; Huang et al., 2024).

Future research could prioritize the development of methods that enhance LLMs' capacity to infer the meaning of novel buzzwords, rather than relying solely on memorization of training data. A promising way is the fine-tuning of LLMs with high-quality, CoT data tailored for buzzword comprehension, mirroring the developmental strategies employed in models such as DeepSeek R1 (DeepSeek-AI et al., 2025). Furthermore, attention should be directed towards devising improved methods for selecting high-quality UGC (i.e., dictionary example) that offers insightful definitions of buzzwords. Current approaches that depend on pre-existing definitions are inadequate for this task. A potentially effective strategy is a self-training paradigm, wherein initial definitions are generated from a preliminary set of UGC, and these definitions are subsequently utilized to iteratively refine the UGC selection process.

Limitations

Multimodal user-generated content. UGC, such as pictures, reviews, and posts created by customers on social media, provides a rich source to understand internet buzzwords. By combining textual, visual, and audio elements, one can gain a more nuanced grasp of a buzzword's meaning. For the present study, we focus solely on textual information, leaving the exploration of multimodal data for future research.

More effective definition generation methods. Our benchmark analysis reveals the limitations of current buzzword definition generation methods. While our novel approach, RESS, demonstrates improvement over existing methods, overall performance remains below optimal. Future research should prioritize enhancing the selection of high-quality UGC without prior knowledge of the target buzzword and improving LLMs' capacity for semantic inference to facilitate the development of more effective definition generation methods.

Language Studied. In this paper, we limit our focus to Chinese. The reason for this is that the topic and problem studied in this paper come directly from a Chinese Internet company (i.e., JD.com). We're open to exploring how our work could be applied to other languages in the future.

Ethics Statement

The Chinese buzzword dataset presented in this work is derived from anonymized, publicly available internet content. Specifically, open-source tools were utilized to collect UGC containing buzzwords, exclusively retrieving textual data related to the target buzzwords. No personally identifiable information, including user images, IDs, or website sources, was collected. This rigorous anonymization process ensures the privacy of internet users and the ethical use of online data.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (No. 62272330 and No.U24A20328); in part by the Fundamental Research Funds for the Central Universities (No. YJ202219); in part by the Science Fund for Creative Research Groups of Sichuan Province Natural Science Foundation (No. 2024NSFTD0035); in part by the National Major Scientific Instruments and Equipments Development Project of Natural Science Foundation of China under Grant (No. 62427820); in part by the Natural Science Foundation of Sichuan (No. 2024YFHZ0233).

References

Isabel L Beck, Charles A Perfetti, and Margaret G McKeown. 1982. Effects of long-term vocabulary instruction on lexical access and reading comprehension. *Journal of educational psychology*, 74(4):506.

Enrico Benedetti, Akiko Aizawa, and Florian Boudin. 2024. Automatically suggesting diverse example sentences for l2 japanese learners using pre-trained language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 114–131.

Paul Bloom. 2000. *How Children Learn the Meanings of Words*. The MIT Press.

Ruichu Cai, Shengyin Yu, Jiahao Zhang, Wei Chen, Boyan Xu, and Keli Zhang. 2025. Dr.ECI: Infusing large language models with causal knowledge for

decomposed reasoning in event causality identification. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9346–9375, Abu Dhabi, UAE. Association for Computational Linguistics.

Kate Cain, Jane Oakhill, and Kate Lemmon. 2004. Individual differences in the inference of word meanings from context: The influence of reading comprehension, vocabulary knowledge, and memory capacity. *Journal of educational psychology*, 96(4):671.

Andrea Cornwall. 2007. Buzzwords and fuzzwords: deconstructing development discourse. *Development in practice*, 17(4-5):471–484.

Gerard De Melo and Gerhard Weikum. 2009. Extracting sense-disambiguated example sentences from parallel corpora. In *Proceedings of the 1st Workshop on Definition Extraction*, pages 40–46.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Jörg Didakowski, Lothar Lemnitzer, and Alexander Geyken. 2012. Automatic example sentence extraction for a contemporary german dictionary. In *Proceedings EURALEX*, pages 343–349.

Duanyu Feng, Bowen Qin, Chen Huang, Youcheng Huang, Zheng Zhang, and Wenqiang Lei. 2025. Legend: Leveraging representation engineering to annotate safety margin for preference datasets. *Proceedings of the AAAI Conference on Artificial Intelligence (AI Alignment Track)*.

Valentina Fiasco and Kate Massarella. 2022. Human-wildlife coexistence: Business as usual conservation or an opportunity for transformative change? *Conservation and Society*, 20(2):167–178.

Mingqi Gao, Xinyu Hu, Jie Ruan, Xiao Pu, and Xiaojun Wan. 2024. Llm-based nlg evaluation: Current status and challenges. *arXiv preprint arXiv:2402.01383*.

Eva Hartell and Jeffrey Buckley. 2021. Comparative judgment: An overview. *Handbook for Online Learning Contexts: Digital, Mobile and Open: Policy and Practice*, pages 289–307.

Jessica S Horst, Kelly L Parsons, and Natasha M Bryan. 2011. Get the story straight: Contextual repetition promotes word learning from storybooks. *Frontiers in Psychology*, 2:17.

Chen Huang, Peixin Qin, Wenqiang Lei, and Jiancheng Lv. 2024. Towards equipping transformer with the ability of systematic compositionality. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):18289–18297.

Han Huang, Tomoyuki Kajiwara, and Yuki Arase. 2021. Definition modelling for appropriate specificity. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2499–2509.

Jie Huang, Hanyin Shao, Kevin Chen-Chuan Chang, Jinjun Xiong, and Wen-mei Hwu. 2022. Understanding jargon: Combining extraction and generation for definition modeling. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3994–4004, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Shonosuke Ishiwatari, Hiroaki Hayashi, Naoki Yoshinaga, Graham Neubig, Shoetsu Sato, Masashi Toyoda, and Masaru Kitsuregawa. 2019. Learning to

describe unknown phrases with local and global contexts. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3467–3476.

Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. 2024. Livecodebench: Holistic and contamination free evaluation of large language models for code. *arXiv preprint arXiv:2403.07974*.

James Jhirad, Edison Marrese-Taylor, and Yutaka Matsuo. 2023. Evaluating large language models' understanding of financial terminology via definition modeling. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 93–100, Nusa Dua, Bali. Association for Computational Linguistics.

Ian Jones and Matthew Inglis. 2015. The problem of assessing problem solving: Can comparative judgement help? *Educational Studies in Mathematics*, 89:337–355.

Pulkit Kathuria and Kiyoaki Shirai. 2012. Word sense disambiguation based on example sentences in dictionary and automatically acquired from parallel corpus. In *Advances in Natural Language Processing: 8th International Conference on NLP, JapTAL 2012, Kanazawa, Japan, October 22-24, 2012. Proceedings*, pages 210–221. Springer.

Adam Kilgarriff, Milos Husák, Katy McAdam, Michael Rundell, and Pavel Rychlỳ. 2008. Gdex: Automatically finding good dictionary examples in a corpus. In *Proceedings of the XIII EURALEX international congress*, volume 1, pages 425–432. Institut Universitari de Linguistica Aplicada, Universitat Pompeu Fabra ....

Cunliang Kong, Yun Chen, Hengyuan Zhang, Liner Yang, and Erhong Yang. 2022. Multitasking framework for unsupervised simple definition generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5934–5943, Dublin, Ireland. Association for Computational Linguistics.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, Kai Shu, Lu Cheng, and Huan Liu. 2025. From generation to judgment: Opportunities and challenges of llm-as-a-judge. *Preprint*, arXiv:2411.16594.

Jiahuan Li, Yu Bao, Shujian Huang, Xinyu Dai, and Jiajun Chen. 2020. Explicit semantic decomposition for definition generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 708–717.

Zhen Li, Xiaohan Xu, Tao Shen, Can Xu, Jia-Chen Gu, Yuxuan Lai, Chongyang Tao, and Shuai Ma. 2024. Leveraging large language models for NLG evaluation: Advances and challenges. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16028–16045, Miami, Florida, USA. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Baoxi Liu, Peng Zhang, Tun Lu, and Ning Gu. 2020. A reliable cross-site user generated content modeling method based on topic model. *Knowledge-Based Systems*, 209:106435.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023a. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Yuxuan Liu, Tianchi Yang, Shaohan Huang, Zihan Zhang, Haizhen Huang, Furu Wei, Weiwei Deng, Feng Sun, and Qi Zhang. 2023b. Calibrating llm-based evaluator. *arXiv preprint arXiv:2309.13308*.

Elena N Malyuga and Wayne Rimmer. 2021. Making sense of "buzzword" as a term through co-occurrences analysis. *Heliyon*, 7(6).

Lingrui Mei, Shenghua Liu, Yiwei Wang, Baolong Bi, and Xueqi Cheng. 2024. Slang: New concept comprehension of large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Stephan C Meylan and Elika Bergelson. 2022. Learning through processing: Toward an integrated approach to early word learning. *Annual review of linguistics*, 8(1):77–99.

Yisong Miao, Hongfu Liu, Wenqiang Lei, Nancy Chen, and Min-Yen Kan. 2024. Discursive socratic questioning: Evaluating the faithfulness of language models' understanding of discourse relations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6277–6295, Bangkok, Thailand. Association for Computational Linguistics.

Timothee Mickus, Denis Paperno, and Mathieu Constant. 2019. Mark my word: A sequence-to-sequence approach to definition modeling. In *Proceedings of the First NLPL Workshop on Deep Learning for Natural Language Processing*.

William E Nagy, Patricia A Herman, and Richard C Anderson. 1985. Learning words from context. *Reading research quarterly*, pages 233–253.

Thanapon Noraset, Chen Liang, Larry Birnbaum, and Doug Downey. 2017. Definition modeling: Learning to define word embeddings in natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Zhengqi Pei, Zhewei Sun, and Yang Xu. 2019. Slang detection and identification. In *Proceedings of the 23rd conference on computational natural language learning (CoNLL)*, pages 881–889.

Ildikó Pilán, Elena Volodina, and Lars Borin. 2016. Candidate sentence selection for language learning exercises: from a comprehensive framework to an empirical evaluation. *Traitement Automatique des Langues*, 57(3):67–91.

Shenbin Qian, Constantin Orăsan, Félix Do Carmo, and Diptesh Kanojia. 2023. Challenges of human vs machine translation of emotion-loaded chinese microblog texts. In *Proceedings of Machine Translation Summit XIX, Vol. 2: Users Track*, pages 217–236.

Peixin Qin, Chen Huang, Yang Deng, Wenqiang Lei, and Tat-Seng Chua. 2024. Beyond persuasion: Towards conversational recommender system with credible explanations. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4264–4282, Miami, Florida, USA. Association for Computational Linguistics.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. Direct preference optimization: your language model is secretly a reward model. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Jessie Ricketts, Dorothy VM Bishop, Hannah Pimperton, and Kate Nation. 2011. The role of self-teaching in learning orthographic and semantic aspects of new words. *Scientific Studies of Reading*, 15(1):47–70.

Esteban Rodríguez-Betancourt and Edgar Casasola-Murillo. 2023. Exploring the limits of large language models for word definition generation: A comparative analysis. In *2023 XLIX Latin American Computer Conference (CLEI)*, pages 1–7. IEEE.

Walid S Saba. 2024. Llms' understanding of natural language revealed. *arXiv preprint arXiv:2407.19630*.

Vincent Segonne. 2023. " definition modeling: To model definitions." generating definitions with little to no semantics. In *Proceedings of the 15th International Conference on Computational Semantics pages*, volume 258, page 266.

Ivan Sekulić, Mohammad Aliannejadi, and Fabio Crestani. 2022. Evaluating mixed-initiative conversational search systems via user simulation. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, WSDM '22, page 888–896, New York, NY, USA. Association for Computing Machinery.

Hiroyuki Shinnou and Minoru Sasaki. 2008. Division of example sentences based on the meaning of a target word using semi-supervised clustering. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019. MASS: Masked sequence to sequence pre-training for language generation. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5926–5936. PMLR.

Ranka Stanković, Branislava Šandrih, Rada Stijović, Cvetana Krstev, Duško Vitas, and Aleksandra Marković. 2019. Sasa dictionary as the gold standard for good dictionary examples for serbian. *Electronic lexicography in the 21st century: Smart lexicography*, pages 248–269.

Zhewei Sun, Richard Zemel, and Yang Xu. 2022. Semantically informed slang interpretation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5213–5231.

Machteld SL Swanborn and Kees de Glopper. 2002. Impact of reading purpose on incidental word learning from context. *Language learning*, 52(1):95–117.

Daniel Swingley. 2010. Fast mapping and slow mapping in children's word learning. *Language learning and Development*, 6(3):179–183.

Arseny Tolmachev, Sadao Kurohashi, and Daisuke Kawahara. 2022. Automatic japanese example extraction for flashcard-based foreign language learning. *Journal of information processing*, 30:315–330.

Oren Tsur and Ari Rappoport. 2015. Don't let me be# misunderstood: Linguistically motivated algorithm for predicting the popularity of textual memes. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 9, pages 426–435.

Peiyi Wang, Lei Li, Liang Chen, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. 2023a. Large language models are not fair evaluators. *arXiv preprint arXiv:2305.17926*.

Xiaolei Wang, Xinyu Tang, Xin Zhao, Jingyuan Wang, and Ji-Rong Wen. 2023b. Rethinking the evaluation for conversational recommendation in the era of large language models. In *Proceedings of the 2023*

Conference on Empirical Methods in Natural Language Processing, pages 10052–10065, Singapore. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Jinyang Wu, Feihu Che, Xinxin Zheng, Shuai Zhang, Ruihan Jin, Shuai Nie, Pengpeng Shao, and Jianhua Tao. 2024. Can large language models understand uncommon meanings of common words? *arXiv preprint arXiv:2405.05741*.

Liner Yang, Cunliang Kong, Yun Chen, Yang Liu, Qinan Fan, and Erhong Yang. 2020. Incorporating sememes into chinese definition modeling. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:1669–1677.

Seonghyeon Ye, Doyoung Kim, Sungdong Kim, Hyeonbin Hwang, Seungone Kim, Yongrae Jo, James Thorne, Juho Kim, and Minjoon Seo. 2023. Flask: Fine-grained language model evaluation based on alignment skill sets. *arXiv preprint arXiv:2307.10928*.

Yunting Yin and Steven Skiena. 2023. Word definitions from large language models. *arXiv preprint arXiv:2311.06362*.

Hengyuan Zhang, Dawei Li, Yanran Li, Chenming Shang, Chufan Shi, and Yong Jiang. 2023. Assisting language learners: Automated trans-lingual definition generation via contrastive prompt learning. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 260–274.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Hua Zheng, Damai Dai, Lei Li, Tianyu Liu, Zhifang Sui, Baobao Chang, and Yang Liu. 2021. Decompose, fuse and generate: A formation-informed method for chinese definition generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5524–5531.

A Chinese Buzzword Dataset Collection

We collected a comprehensive dataset of Chinese buzzwords and their definitions, primarily sourced from online encyclopedias and similar platforms known for their extensive coverage of contemporary language. To analyze the usage of these buzzwords, we gathered user-generated content (UGC) from two social media platforms, focusing on posts containing corresponding buzzwords. This includes 1) *Xiaohongshu*, a leading social media platform popular among young Chinese users, and 2) *Weibo*, one of the biggest social media platforms in China with over 582 million monthly active users. Table 1 summarizes the data statistics, while Table 2 provides a specific case from our dataset. The complete dataset and code are available at: https://github.com/SCUNLP/Buzzword.

Buzzword Collection and Quality Control. To gather a comprehensive list of current Chinese internet buzzwords, we first utilized two online dictionary resources[9], both known for their up-to-date and extensive collections of trending internet vocabulary. In particular, we gathered a list of buzzwords from these platforms and eliminated any duplicates. To ensure safety and ethical standards, we manually remove potentially harmful buzzwords, including those related to sexually suggestive, offensive, or violent content.

Definition Collection and Quality Control. We crawled each online dictionary resource for the buzzword's description, which typically included details about its origin, cultural references, informal explanations, and sometimes example sentences, importantly reflecting how definitions evolved over time (e.g., the initial meaning vs. later usage). While those information are valuable for understanding the buzzword, it's not ideal for a concise definition. Therefore, we utilized *GPT-4o* to analyze each buzzword's description and summarize a succinct definition encompassing both its literal and figurative meanings (if any). Taking Table 2 in our paper for example, Descriptions of buzzword '0帧起手' from Online Source are '0帧起手指零帧技能，一般指的是点击即可释放，并且立刻判定无法打断的技能。0帧起手在网络上表示动作极快，没有丝毫等待，绝不拖泥带水，闪电般突然出现的动作。' (*Translation: '0 frame startup' generally refers to a skill that can be released by clicking and immediately determines that it cannot be interrupted. This term, often used online, signifies lightning-fast action with no delay—a sudden strike like a bolt of lightning*). LLM-summarized (i.e., generated) definition in our dataset is '原意是指游戏中一些无需准备时间，可以瞬间释放的技能，引申为行动迅速，毫不拖延。' (*Translation: Originally referring to in-game abilities usable without any setup time, the term has broadened to describe taking swift and immediate action*). Finally, each gen-

[9] https://ttseed.cn and https://gengbaike.cn

erated definition underwent a manual review and refinement to ensure its semantic accuracy, conciseness, and language fluency. Specifically, our definition quality check and refinement process involved two volunteers, both university students majoring in NLP with extensive experience using the internet and social media platforms. For each buzzword, these volunteers independently compared the GPT-4o-generated definition with the original definition and description obtained from the online dictionary websites, ensuring no interference or communication between them. Particular attention was given to semantic accuracy, conciseness, and language fluency. Regarding semantic accuracy, volunteers were instructed not only to compare the overall semantic similarity of the two definitions, but also to verify whether the GPT-4o-generated definition included both the original and figurative definitions (if provided in the source). Any content beyond these definitions was removed, and definitions were manually modified as needed. During the review process, we recorded which buzzwords and definitions had been modified. For modified definitions, the two volunteers were required to discuss and reach a consensus on the final definition. Ultimately, fewer than 5% of the buzzword definitions were modified.

Example Collection and Quality Control. To gather real-world examples of each buzzword in use, we searched for user posts on Xiaohongshu[10] and Weibo[11]. We used the buzzwords as keywords and collected the post titles and descriptions using a web crawler from the GitHub repository[12]. Since the search engines of Weibo and Xiaohongshu sometimes split keywords and returns posts containing only parts of the buzzword, we carefully filtered the results to ensure each selected example used the complete buzzword. After that, we employ a LLM to eliminate any sentences that simply describe the corresponding buzzword (i.e., the definitional information). To achieve this, we use a two-step process to remove definitional sentences. First, we leverage a LLM (Qwen-max) to automatically exclude sentences from our UGC corpus that contain definitional information. We also prompts the LLM to identify common keywords and patterns that users used to describe the definitional information (e.g., '[BUZZWORD]意味着...' (*[BUZZWORD] means that ...*) and '盘点近期网络热梗：...' (*overview of trending buzzwords online: ...*)). Second, we manually review the remaining sentences, paying particular attention to those follow the identified keywords or patterns, to guarantee the exclusion of any remaining definitional content. The refined UGC then served as representative examples for our experiment. Note that buzzwords with no corresponding example are excluded from the final corpus.

High quality of our CHEER. First, we want to emphasize that the ground truth definitions in our dataset are not directly generated by GPT-4o. Rather, the original definitions come from reputable online buzzword dictionary websites. GPT-4o was only used to summarize these original descriptions, which were already of high quality, having been verified by the websites themselves and accepted by internet users. This significantly facilitated our subsequent processing of the definitions using GPT-4o and the human verification process, and we have confidence in the reliability of our data (Please refer to the case studies in Table 2 of our paper). Thus, the quality of our dataset has been rigorously controlled, having passed through three layers of vetting: the dictionary websites, internet users, and our own review process.

B Annotation for Contamination-free Evaluation

To conduct contamination-free evaluation, we determine for each LLM whether it possesses knowledge of specific buzzwords. This involved, for each LLM backbone (e.g., Qwen2-7b, Qwen2-72b, GPT-4o mini, GPT-4o, and MASS), dividing our dataset into two distinct parts: one containing buzzwords with known definitions for that specific LLM, and the other containing truly unseen buzzwords. Consequently, the specific buzzwords considered contaminated may differ across various LLM backbones. However, the contamination status for a given buzzword remains consistent across all methods when evaluated using the same LLM backbone.

Speficically, given a LLM, we prompt the LLM to generate definitions based solely on the buzzword itself, without any contextual examples (detailed prompts are provided in Table 16). Subsequently, we conduct a multi-faceted evaluation process. Initial assessments are performed using LLM-based scoring, where *GPT-4o* is utilized to evaluate the semantic accuracy and completeness of generated definitions based on specific scoring

[10] https://www.xiaohongshu.com
[11] https://weibo.com
[12] https://github.com/NanmiCoder/MediaCrawler

Aspects for buzzword understanding
1. 意图理解 (Intention Understanding, IU): 理解说话者使用该词语的意图和目的，例如说话者是想描述一个物体，还是表达一种情感 (Discerning the speaker's communicative goal when using the buzzword, such as describing an object or expressing an emotion)
2. 概念形成 (Concept Association, CA): 将词语与特定的概念联系起来，例如将"狗"这个词与具有特定特征的动物类别联系起来 (Linking the buzzword to relevant concepts. For example, linking the word 'dog' to animal categories with specific characteristics)
3. 语法理解 (Language Structure, LS): 理解词语在句子中的语法角色和功能，例如词语是名词、动词还是形容词，以及它与其他词语之间的关系 (Analyzing the buzzword's grammatical function. For example, whether a word is a noun, verb, or adjective, and its relationship with other words)
4. 基本学习和记忆 (Social Cue Interpretation, SCI): 从该词语的发音和拼写发出，建立它与相关概念之间的联系 (Establishing connections between orthography, phonology, and meaning)
5. 社会线索 (Word Context, WC): 利用说话者的表情、语气、姿势等社会线索来理解词语的含义 (Inferring social context from UGC such as the speaker's facial expressions, tone of voice, and gestures.)
6. 上下文 (Pronunciation and Spelling, PS): 词语出现的具体语境，包括前后文和对话背景等 (Leveraging surrounding text for semantic disambiguation)

Table 7: Aspect description used in RESS.

rubrics (outlined in Table 17). Definitions scoring below a threshold of 3 are considered indicative of the LLM not understanding the buzzword. Finally, a human review process is implemented to ensure accuracy. Three independent evaluators examine the LLM-generated labels, indicating whether the LLM "knows" the buzzword (1) or not (0). A majority vote among three human evaluators determines the final classification. This comprehensive approach allows us to confidently identify which buzzwords are within the current knowledge base of existing LLMs.

C Implementation Details

We conduct all our experiments using a single Nvidia RTX A6000 GPU for the Qwen2 7b model and 4 A6000 GPUs for the Qwen2 72b model, and we implement our codes in PyTorch. We use the Huggingface Evaluator package and bert_score package to calculate the BLUE and Rouge-L, and Bertscore. Finally, for Qwen LLM deployment, we utilize the vLLM framework.

C.1 Implementation of RESS

We provide detailed prompts in Table 14 and Table 15. We incorporate six aspects, shown in Table 7, drawing inspiration from child language acquisition skills.

C.2 Implementation of Word-meaning Agnostic UGC Selector (WAUS)

Lacking prior knowledge of the target buzzword, WAUS is trained using a masked strategy, where the target buzzword within the UGC is masked. This helps prioritize contextual and syntactic information to identify high-quality UGC examples. We provide an overview of WAUS in Figure 7 and detail as follows:

Masked Training Strategy. Unlike existing example selection methods that need meticulously constructed rules, our WAUS employs a data-driven approach. We train a UGC selector by fine-tuning a BERT model with an MLP adapter on a dataset of high- and low-quality examples (details provided in the following paragraph). Crucially, we mask the target buzzword within each example, forcing the model to rely on contextual and syntactic cues rather than the buzzword's semantics to predict sentence quality. This masked training strategy implicitly learns the selection criteria without explicit rule definition.

Training Dataset Contruction. To maintain a contamination-free evaluation, our own dataset CHEER is excluded from the WAUS training pro-

cess. Instead, a new dataset devoid of buzzwords is created specifically for WAUS training. This dataset is constructed in two stages. First, Chinese buzzwords and their corresponding dictionary examples are collected from online resources, regarding as positive (i.e., high-quality) examples. Second, negative (low-quality) examples are generated: initially, the Qwen is prompted to create sentence examples with broad and vague meanings related to given buzzwords; these generated sentences are then manually reviewed. To minimize manual effort, an iterative review process assisted by WAUS is employed. The WAUS model, trained on the currently reviewed portion of the data, predicted the quality of the remaining negative examples. Human review is then prioritized for negative examples incorrectly classified as positive by WAUS. This approach allowed for efficient and cost-effective quality control of the generated negative examples. Finally, we report several training data samples in Table 8.

Buzzwords	Examples
卧薪尝胆 (endure hardships)	**Positive:** 为了一雪前耻，且让我们卧薪尝胆，埋头苦干，以图东山再起。(In order to wipe away past humiliations, let us endure hardships and work diligently with the aim of making a comeback.) **Negative:** 他讲了一个关于卧薪尝胆的故事。(He told a story about endure hardships to achieve one's goals.)
耀武扬威 (flaunt one's power)	**Positive:** 他仗著家中有财有势就耀武扬威，令人十分厌恶。(He flaunts his power and wealth from his family, which makes him very unpleasant.) **Negative:** 那本书里的主角给人一种耀武扬威的感觉。(The protagonist in that book gives off an impression of flaunts his power.)

Table 8: Training data samples

Training details of WAUS. A two-layer Multilayer Perceptron (MLP) adapter, with hidden layer dimensions of 512 and 256 (ReLU activation, 0.5 dropout), is used for classification. This adapter receives the 768-dimensional final layer output from a *BERT-base-Chinese* encoder [13], which processes our crafted training dataset with masked target words. The model is trained for 2 epochs using AdamW (learning rate = 5^{-3}, weight decay = 10^{-5}) with a batch size of 128.

[13] https://huggingface.co/google-bert/bert-base-chinese

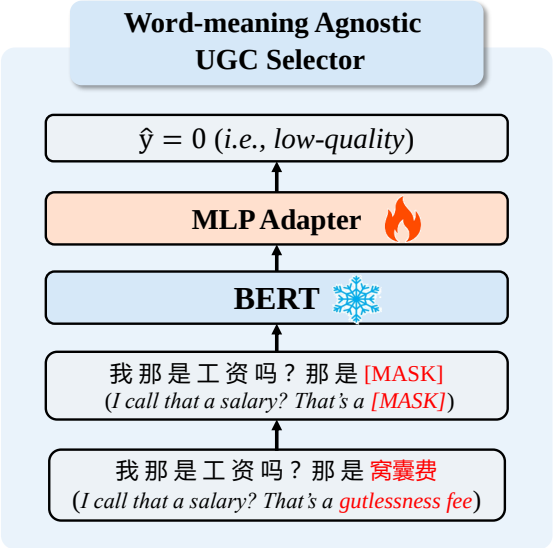


Figure 7: Overview of WAUS

C.3 Implementation of GDEX

GDEX (Kilgarriff et al., 2008) is a well-established rule-based method for dictionary example selection. We implement it using commonly used rules (Pilán et al., 2016; Stanković et al., 2019) that assign a score to each sentence based on the following three criteria.

- **Length Check**. Sentences must be between 10 and 25 characters long (inclusive). Shorter or longer sentences are deemed lower quality and receive a lower score.

- **Pronoun Check**: Sentences containing specific pronouns (e.g., 'it', 'that', 'these') are considered lower quality. These pronouns likely indicate less descriptive or informative sentences. Importantly, sentences starting or ending with numbers or punctuation marks are penalized. This aims to select sentences that are grammatically well-formed and avoid abrupt or incomplete sentences.

- **Common Word Check**. We calculate the ratio of common words to the total number of words in the sentence. A lower ratio suggests higher quality, indicating that the sentence uses more specific and frequently used vocabulary.

C.4 Implementation of Baselines

C.4.1 LM-based Baseline Implementation

For LM-based methods, we implemented them using their official code available on Github.

- **MASS-zh** (Kong et al., 2022). MASS-zh is a language model pre-trained from scratch using the *Chinese Gigaword Fifth Edition* corpus and the MASS backbone (Song et al., 2019). Here, we leverage the publicly available MASS-zh checkpoint[14] for experiments.

- **SimpDefiner** (Kong et al., 2022). This method aims to generate simple definitions to help language learners and low literacy readers. To achieve this, it forms a multi-task learning paradigm, together with the MASS-zh as the backbone. SimpDefiner jointly trains three subtasks: 1) Definition Generation: Learns to generate complex definitions from a standard dictionary. 2) Text Reconstruction: Learns to reconstruct simple sentences from corrupted versions of those sentences from a simple text corpus. 3) Language Modeling: Learns to generate coherent and simple sentences. In our experiments, we followed the author's instruction on GitHub[15] for the model training and implementation.

C.4.2 LLM-based Baseline Implementation

We took our prompts from the corresponding code repositories and papers to implement all LLM-based baselines. More details can be found at https://github.com/Meirtz/FocusOnSlang-Toolbox

- **Direct Prompt & Chain-of-Thought (CoT)**. We extended the original prompts from Mei et al. (2024) by translating them into Chinese. This modification ensures that the generated definitions align with the specific requirements of our study. Moreover, for each buzzword, we integrated all corresponding examples into the prompt. However, due to potential length constraints, we implemented a truncation mechanism to ensure prompt length remains within acceptable limits.

- **FOCUS** (Mei et al., 2024). This method is based on causal inference for enhancing comprehension of new words like buzzwords, achieving SOTA recently. It enables LLMs to analyze phrases according to usage examples and provide counterfactual interpretations, thereby understanding the evolving semantics of language. Specifically, FOCUS employs Structural Causal Models (SCMs) to map out the relationships between different factors that influence how an LLM interprets a phrase. This enables LLMs to grasp the meaning of new phrases and their colloquial context, improving models' adaptability and effectiveness in applications requiring deep understanding of language use. In essence, it helps LLMs better understand new slang, memes, and other emerging language phenomena on the internet. In this paper, we mainly followed the author's instruction on GitHub[16] for implementation. We translate the authors' prompts into Chinese and integrate multiple UGC examples into our prompts.

C.4.3 Setups of LLM Backbones

For both our baseline models and proposed methods, we explored multiple Large Language Model (LLM) backbones. Additionally, GPT-4 is employed as our LLM-based evaluator. To ensure reproducibility of our research, LLM backbones share the same temperature (i.e., 0.7) and random seed (i.e., 10086) (if any)

D Additional Human Evaluation

Win rate evaluation is a well-established and widely adopted method, as supported by prior research (Sekulić et al., 2022; Feng et al., 2025; Wang et al., 2023b; Qin et al., 2024). This choice is based on the established cognitive findings that human annotators exhibit greater proficiency in comparative assessments than in assigning absolute ratings (Rafailov et al., 2023; Jones and Inglis, 2015; Hartell and Buckley, 2021). Notwithstanding this, we undertook supplementary human evaluation to further validate our findings in our main experiments.

In particular, we conducted a targeted human evaluation using the identical dataset and evaluation criteria employed for the win rate evaluation. The annotators involved in this process were also the same individuals, ensuring consistency. During the evaluation, each annotator independently scored the samples across two dimensions: Semantic Accuracy (SA) and Semantic Completeness

[14] https://stublcuedu-my.sharepoint.com/:u:/g/personal/201921296062_stu_blcu_edu_cn/EZpcGUWQanxAt0XZNWb6QqsBauh4dqaR0JdF5u8ia5zJIQ?e=X2tV8r

[15] https://github.com/blcuicall/SimpDefiner

[16] https://github.com/Meirtz/FocusOnSlang-Toolbox

(SC). The final SA and SC scores for each buzzword were derived from the aggregated mean of the human annotations. Finally, the inter-annotator reliability among humans, measured by Krippendorff's alpha, is strong, with coefficients of 67.92% for SA and 68.17% for SC.

The main results are as follows:

- **LLM-Human Evaluation Alignment**. The Krippendorff's alpha coefficients for the agreement between human and LLM scores were 76.05% for SA and 69.32% for SC, indicating a notable level of agreement between the two evaluation methods. We further provide the Krippendorff's alpha coefficients of each definition generation method, detailed in the Table 9.

- **Quality of Generated Definitions**. As shown in the Table 9, our RESS generally demonstrates superior performance to FOCUS, which in turn outperforms CoT, across both evaluation methods. It is also noteworthy that human annotators exhibited a tendency to assign comparatively higher absolute scores for both SA and SC than the LLM evaluator. However, even with these relatively elevated human scores, the overall task performance scores remained significantly below a threshold of 4, indicating that the overall effectiveness, despite these nuances in scoring tendencies, is still considered suboptimal.

E Evaluation Details

E.1 LLM-based Evaluation

Building upon prior research on LLM evaluation (Ye et al., 2023; Wang et al., 2023a,b), we employ a GPT-4o-based instance-wise evaluator for detailed assessment. To minimize scoring bias, consistent with previous work (Ye et al., 2023; Wang et al., 2023b; Liu et al., 2023b), the evaluator uses fine-grained scoring rubrics (1-5), each rubric accompanied by a descriptive explanation. Further enhancing evaluation rigor, the evaluator is prompted to provide a rationale for each score, informed by the benefits of Chain-of-Thought (CoT) prompting (Wei et al., 2022; Ye et al., 2023). Detailed prompts are shown in Table 17.

E.2 Human Evaluation

To demonstrate the correlation between our automatic evaluation and human judgment, we conduct a human evaluation (i.e., the *win rate evaluation*) of definitions generated for 100 randomly selected buzzwords in Section 5.1. Due to the resource-intensive nature of human evaluation, our analysis is limited to three representative methods. For each buzzword, two human evaluators compared the definitions generated by different methods across various backbones, considering both SA and SC. Following Sekulić et al. (2022), the evaluators are presented with pairs of anonymized definitions for the same buzzword, without disclosure of the originating model for each definition. Independent evaluations are followed by a discussion to resolve any discrepancies. A "*Win/Lose/Tie*" label is finally assigned if consensus is reached; otherwise, the result is recorded as a "*Tie*". Our experiments reveal inter-annotator agreement rates of 71.50% and 61.88% for semantic accuracy (SA) and semantic completeness (SC), respectively. Any remaining discrepancies in evaluation results are classified as Ties.

Additionally, we conduct more human evaluation in Appendix D to further validate our findings. Basically, we conduct a targeted human evaluation using the identical dataset and evaluation criteria employed for the *win rate evaluation*. The annotators involved in this process are also the same individuals, ensuring consistency. During the evaluation, each annotator independently score the samples across two dimensions: Semantic Accuracy (SA) and Semantic Completeness (SC). The final SA and SC scores for each buzzword are derived from the aggregated mean of the human annotations. Finally, The inter-annotator reliability among humans, measured by Krippendorff's alpha, is strong, with coefficients of 67.92% for SA and 68.17% for SC.

E.3 Details on Evaluation Metrics

In this paper, we employ a comprehensive evaluation framework that extends beyond conventional metrics such as BLEU, ROUGE-L (R-L, for short), and BERTScore (BScore). We also prioritize the Semantic Accuracy (SA) and Semantic Completeness (SC) of generated definitions, as emphasized in previous studies (Li et al., 2020; Segonne, 2023). Importantly, conventional metrics are standard evaluation metrics in the definition generation domain (Zheng et al., 2021; Huang et al., 2021; Li et al., 2020). These metrics aim to provide an indication of similarity to a reference definition from the perspectives of word matches and/or semantic embedding similarity. Notably, conventional

Method	LLM Backbone	Krippendorff's Alpha		Human Evaluation		LLM Evaluation	
		SA	SC	SA	SC	SA	SC
CoT	Qwen2-7b	80.46	71.46	2.62	2.54	2.24	2.19
FOCUS		74.71	74.83	2.63	2.55	2.28	2.41
RESS		86.89	81.13	2.55	2.48	2.31	2.54
CoT	Qwen2-72b	70.94	61.29	3.2	3.09	2.70	2.62
FOCUS		72.23	68.97	3.26	3.18	2.82	3.20
RESS		78.05	73.82	3.31	3.2	2.98	3.06
CoT	GPT-4o Mini	65.82	66.30	2.79	2.65	2.27	2.07
FOCUS		74.64	67.81	2.84	2.72	2.59	2.7
RESS		73.17	63.16	2.97	2.81	2.68	2.75
CoT	GPT-4o	63.71	64.03	3.18	3.02	2.68	2.34
FOCUS		72.78	66.14	3.32	3.22	2.92	3.03
RESS		76.46	72.35	3.58	3.42	3.18	3.31

Table 9: Human and LLM evaluation results on sampled data.

metrics continue to be widely utilized in the LLM era (Mei et al., 2024). However, conventional metrics often fall short in judging subtle attributes and delivering satisfactory results. These metrics are easily misled by surface-level similarities and do not explicitly assess the validity of the generated definition's meaning. Therefore, consistent with recent studies (Li et al., 2025; Gao et al., 2024; Li et al., 2024; Liu et al., 2023a), we incorporate LLM evaluation into our experiments. By this means, we aim not only to obtain more reliable evaluation results but also to encourage the introduction of enhanced evaluation metrics within the definition generation community.

In this section, we detail all metrics used in our experiments in the following.

- **BLEU** (Papineni et al., 2002) is a common metric used to assess the quality of generated text by comparing it to a reference, or ground truth. It is a widely used metric for automatically evaluating machine-translated text. It works by comparing the generated translation to one or more human-produced reference translations. The core idea is to count matching n-grams (sequences of n words) between the generated text and the references, giving credit for matches. The closer a machine translation is to a human reference translation, the higher its BLEU score will be. In our case, we use BLEU to evaluate how well the model-generated word definition matches the ground truth.

- **ROUGE-L** (Lin, 2004) is a popular metric for evaluating text generation tasks, particularly summarization and definition generation, with the aim to capture the meaning and key information from a reference text. Unlike BLEU which focuses on precision (how much of the generated text is relevant), ROUGE-L emphasizes recall (how much of the reference text is captured by the generated text). Specifically, ROUGE-L measures the length of the longest common subsequence (LCS) between the generated text and the reference text. The LCS is the longest sequence of words that appear in the same order in both texts, but not necessarily consecutively. By focusing on the LCS, ROUGE-L can capture sentence-level structure and meaning, even if the word order is slightly different. It's generally considered good at assessing how well the generated text covers the important content from the reference.

- **BERTScore** (Zhang et al., 2019) is a metric that leverages pre-trained language models like BERT to evaluate text generation tasks. Unlike traditional metrics like BLEU and ROUGE, which rely on exact word matches, BERTScore assesses the semantic similarity between the generated text and the reference text. Usually, it gives a fine-grained measure of semantic similarity based on cosine similarity

- **Semantic Accuracy (SA)** (Li et al., 2020; Segonne, 2023) is a measure of how faithfully the generated definition reflects the accepted understanding of the word's meaning, judged against a reference definition. BLEU, ROUGE-L, and

BERTScore can provide some indication of similarity to a reference definition, and if the reference definition is accurate, a high score might suggest some level of semantic accuracy. However, they are easily fooled by surface-level similarities and don't explicitly assess the validity of the generated definition's meaning. Therefore, they are insufficient and unreliable as direct measures of semantic accuracy. Therefore, we utilize LLM-based evaluation, detailed in Appendix E.1, which is better suited for evaluating semantic accuracy.

- **Semantic Completeness (SC)** (Li et al., 2020) (or called Factuality (Segonne, 2023)) refers to a definition encompassing all and only the relevant aspects of a word's meaning. As illustrated in (Li et al., 2020), accurately defining "captain" in the context "The captain gave the order to abandon the ship" requires knowing that (1) a captain is a person, (2) a captain works on a ship, and (3) a captain is typically responsible for the ship. To assess SC, we employ a LLM-based evaluation, the specifics of which are detailed in Appendix E.1.

F Additional Analysis on RESS

Our primary objective is not to introduce a universally superior method that achieves state-of-the-art performance across all metrics and LLM backbones. Instead, this paper presents a benchmark study to investigate the fundamental question of whether LLMs can effectively understand internet buzzwords. While our proposed method RESS demonstrates some performance improvements over existing approaches (cf. Section 5.1), we include it within our benchmark analysis to critically assess the limitations of current methods, including our own (cf. Section 5.2).

Beyond the benchmark analysis, this section aims to further explore the unique characteristics of our method RESS, offering readers a new perspective to understand its underlying principles. In particular, we have the following observations.

***Why does* RESS *exhibit inferior performance under smaller LLMs?* – This stems from LLM's difficulty in understanding and applying the illustrative aspects that guide buzzword definition generation**. As evidenced by Tables 3 and 4, no single method consistently achieves the best performance across all metrics and LLM backbones. Regarding our proposed method RESS, it demonstrates superior performance compared to FOCUS when utilizing larger LLM backbones, while achieving comparable performance to FOCUS with smaller LLMs. Given that RESS leverages key skills of child learning and codifies them into illustrative aspects to guide LLM-driven buzzword definition generation, we hypothesize that its less-than-optimal performance on smaller LLMs may stem from their difficulty in understanding these illustrative aspects and, consequently, in generating accurate definitions. To verify this hypothesis, we manually checked 100 buzzwords and the corresponding definitions generated by both Qwen2-7b-based RESS and GPT-4o-based RESS. For each LLM and each buzzword, two human evaluators independently assessed whether each aspect-guided definition conformed to the specific aspect's requirements. As shown in Table 11, we found that, on average, Qwen2-7b generates a definition consistent with the specific aspect requirements in 41.8% of cases, while GPT-4o achieved a success rate of 57.2%. Furthermore, we found that the aspects, PS and CA, are particularly challenging for both LLM.

Aspects	GPT-4o-based RESS	Qwen2-7b-based RESS
Intention Understanding, IU	22.97%	22.58%
Concept Association, CA	18.92%	22.58%
Language Structure, LS	18.92%	12.90%
Social Cue Interpretation, SCI	10.81%	11.29%
Word Context, WC	18.92%	17.74%
Pronunciation and Spelling, PS	9.46%	12.90%

Table 10: The average selection rate for each of the six aspects that most effectively guide the final definition.

***Why does adding more than three aspects result in minimal performance gains?* – Understanding a word may only require a subset of key aspects**. While Figure 5 demonstrates that the performance of our method does improve with an increasing number of aspects, the rate of improvement diminishes as more aspects are added. We hypothesize that this trend arises because, for both LLMs and humans, understanding a word may not require the consideration of all possible aspects; perhaps a subset of key aspects is sufficient. Furthermore, this subset may vary depending on the word. This would explain why providing more aspects to the LLM leads to performance gains—the newly added aspects may offer enhanced understanding for a subset of words, even if not for all. To validate this hypothesis, we conducted further experiments, manually checking 100 buzzwords and their corresponding definitions generated by both Qwen2-

Aspects	GPT-4o-based RESS	Qwen2-7b-based RESS
Intention Understanding, IU	98.5	96
Concept Association, CA	28	29
Language Structure, LS	64	7.5
Social Cue Interpretation, SCI	46	38
Word Context, WC	99	78.5
Pronunciation and Spelling, PS	7.5	1.5
Avg.	57.2	41.8

Table 11: Accuracy (%) of human evaluation in judging whether aspect-guided definitions conform to specific aspect requirements.

7b-based RESS and GPT-4o-based RESS. For each buzzword and LLM, two individuals independently selected up to three of the six aspect-guided definitions that were most helpful in guiding the final definition generated by each model (e.g., definitions that were semantically closest, provided usage scenarios, revealed original meanings, or presented additional information like figurative meanings). Table 10 records the averaged selection percentages for each of the six aspects. As these results illustrate, not all aspects contribute equally to the generation of a final definition. Some aspects appear to be more beneficial for understanding specific buzzwords, leading to less significant overall performance improvements on a dataset-wide scale. This finding supports our previous hypothesis, as it indicates that: 1) using more aspects can be beneficial and 2) some aspects may be more helpful for understanding a specific subset of buzzwords, thereby contributing to overall dataset performance without all aspects being equally important.

***Why do* RESS *and FOCUS perform poorly under conventional metrics compared to other methods?* – RESS and FOCUS generate free-form, lengthy definitions with elaborations that are penalized by the n-gram matching and similarity calculations inherent in these metrics.**. The results presented in Table 3 and Table 4 indicate that neither FOCUS nor our proposed RESS method consistently outperforms the DP and CoT baselines according to conventional metrics. This discrepancy arises primarily because FOCUS and RESS tend to generate free-form and long buzzword definitions, often including elaborations on the figurative meaning and connotations of the target buzzword. Given that conventional metrics such as BLEU and R-L rely on n-gram word matching, they inherently penalize free-form, lengthy responses that do not precisely replicate the vocabulary of the gold reference (i.e., the ground-truth definitions). Consequently, generating longer definitions without using the exact words from the ground truth can lead to lower BLEU and R-L scores. Furthermore, the lower BScore observed for FOCUS and RESS compared to DP and CoT can be attributed to the additional explanatory content regarding the extended figurative meanings. This supplementary information may introduce semantic noise during BERT's similarity calculations. These inherent limitations of conventional metrics serve as a key motivation for incorporating LLM-based evaluation. As demonstrated in the LLM evaluation results of Table 3 and Table 4, FOCUS and RESS are indeed shown to be superior to DP and CoT, a finding that aligns with our human evaluation results. To illustrate this point further, we provide two case studies in Appendix H, offering a more intuitive demonstration of how the definitions generated by FOCUS and RESS include elaborations on the extended meanings, resulting in significantly longer outputs compared to other methods.

G Methods for Dictionary Example Selection

While high-quality examples effectively illustrate a word's meaning and typical usage, identifying or creating such examples can be a laborious and costly endeavor (Stanković et al., 2019; De Melo and Weikum, 2009). While methods for automatic example selection have been proposed, they frequently rely on pre-existing word definitions as supervised signals to train a model and subsequently locate suitable examples (Kathuria and Shirai, 2012; Shinnou and Sasaki, 2008; Tolmachev et al., 2022; Benedetti et al., 2024), a strategy that is not applicable in our case since the definition of the buzzword is initially unknown (See Section 5.2.2). Instead, a limited number of studies have proposed rule-based methods without relying on word definitions. These methods, for example, GDEX (Kilgarriff et al., 2008), prioritize readability and informativeness, measured by, for example, sentence length, word frequencies, and syntactic information (Pilán et al., 2016; Didakowski et al., 2012; Stanković et al., 2019). In this paper, we also present a novel example selection method that bypasses the need for meticulously constructed rules.

H Case Studies

This section provides case studies for better understanding the performance of different methods. Table 13 showcases the case of buzzword '窝囊费'.

According to the case study, RESS generates more nuanced definitions, capturing not only the concept of hard work itself but also the frustration and resignation stemming from the contrast between hard work and meager compensation. Moreover, Table 12 shows another example of buzzword '梁静茹给的勇气'. In this case, both the DP and CoT fail to deliver accurate definition. Additionally, the definitions generated by FOCUS erroneously include the sense of "willingness to proceed".

Internet Buzzword
梁静茹给的勇气 (The courage given by Fish Leong)
Description from Online Source
梁静茹给的勇气，多是疑问句， 是梁静茹给你的勇气吗? 意思是你哪来的勇气说出这句话或者干出这种事，是在嘲讽。 (The phrase "The courage given by Fish Leong" is often used sarcastically as a rhetorical question, implying, "Where did you get the nerve to say/do that?" Usually, it's a way of mocking someone's audacity.)
Definition
嘲讽某人哪来的勇气说出这句话或做出这件事 (To sarcastically mock someone for their audacious behavior)
Examples (i.e., UGC). We only show one example here to save space
oppo用脚做pro，直接一整个退订，梁静茹给的勇气吧， 短焦砍主摄还敢定价这么高，我的小米继续坚挺 (Oppo's Pro model is a complete letdown; I've canceled my pre-order. They must have gotten their courage from Fish Leong to cut the main camera on a short-focus lens and still charge such a high price. My Xiaomi phone will continue to serve me well.)
Definitions of DP
一种源自梁静茹音乐作品影响下的勇气， 鼓励人们在面对挑战或困难时勇敢做出决策和行动的心态。 (A courage inspired by the music of Fish Leong, empowering individuals to make bold decisions and take action in the face of challenges and adversity.)
Definitions of CoT
指因受到梁静茹的音乐或精神影响而产生的勇气与决心， 常用于形容在面对困难或挑战时， 勇敢做出决定或采取行动的状态。 (It refers to the courage and resolve inspired by the music or spirit of singer Fish Leong, often used to describe the state of bravely making decisions or taking action in the face of difficulty or challenge.)
Definitions of FOCUS
梁静茹给的勇气表示一种因冲动或勇敢而做出的决定或行动， 常带有调侃和自嘲的意味。 它经常被用于描述在面对困难、挑战或不确定性时， 做出大胆选择或尝试的情景。这个短语通常反映出一种无畏的态度， 可能隐含着对所做决定的无奈或讽刺， 表示虽然心中有一种不安，但仍然选择去尝试。 (It refers to a decision or action driven by impulsiveness or bravery, often with a self-deprecating or ironic undertone. It's frequently used to describe bold choices or attempts in the face of difficulty, challenges, or uncertainty. The phrase typically reflects a fearless attitude, possibly tinged with resignation or sarcasm regarding the decision, suggesting a sense of apprehension alongside the willingness to proceed.)
Definitions of RESS
梁静茹给的勇气是一种调侃的表达，用来形容某人做出大胆或不理智的行为， 通常带有讽刺或自嘲的意味，暗示该行为的勇气似乎是凭空而来， 不切实际。这种表达源自歌手梁静茹的歌曲《勇气》， 常用于描述一些看似勇敢但并未经过深思熟虑的决定或行动 (It is an ironic expression used to describe bold or reckless behavior, often with a sarcastic or self-deprecating tone. It implies that the courage seems unfounded and unrealistic, referencing the singer Fish Leong's song of the same name ("Courage"). This expression is commonly used to describe decisions or actions that appear brave but lack careful consideration.)

Table 12: Case of buzzword '梁静茹给的勇气' and generated definitions by different methods. Here, we also provide its English translation for better understanding.

Internet Buzzword
窝囊费 (gutlessness fee)
Description from Online Source
指那些在职场上辛苦耕耘、忍气吞声，最终换来的微薄薪水。 (Refers to those who work hard and endure hardships in the workplace, ultimately earning meager salaries.)
Definition
指工作辛苦劳累、受委屈，得到的回报却很少。 (Despite significant effort and hardship endured, the salary received is minimal)
Examples (i.e., UGC). We only show one example here to save space
早知道就考体制内了，也不至于月月赚窝囊费 (I should have gone for a government job. At least I wouldn't be earning gutlessness fee every month.)
Definitions of DP
形容微薄、无足轻重的收入或工资，通常带有不满或无奈的情绪 (Describing meager and insignificant income or salary, usually accompanied by dissatisfaction or helplessness)
Definitions of CoT
指微薄的工资或收入，通常带有无奈、无力感，反映出工作的辛苦 与经济压力，常用于表达对现状的不满或无奈 (Refers to meager wages or income, usually with a sense of helplessness and powerlessness, reflecting the hardship and economic pressure of work, often used to express dissatisfaction or helplessness towards the current situation.)
Definitions of FOCUS
窝囊费表示一种微薄且让人感到无奈的工资或报酬。 它经常被用于形容工作带来的失望和对生活状态的自嘲。 这个短语常常隐含着对工作的无奈与对生活现状的不满， 表达出人们在日常工作中所感受到的疲惫与无力感。 (Gutlessness fee refers to meager and disheartening pay or compensation. It's often used to express the disappointment stemming from one's work and a self-deprecating attitude towards one's life circumstances. The term subtly conveys feelings of helplessness about work and dissatisfaction with one's current life situation, reflecting the weariness and powerlessness many experience in their daily work lives.)
Definitions of RESS
窝囊费是指一种微薄或低质量的收入，通常被用来形容工资、 报酬或补贴，带有无奈、轻蔑和自嘲的情感色彩， 反映出劳动者对工作状况和生活质量的不满与失望。 这个词语常常在语境中体现出对收入微薄与辛苦工作的无奈， 表达人们对自身价值的反思和对现实的不满 与经济压力，常用于表达对现状的不满或无奈 (Gutlessness fee refers to meager or low-quality income, typically used to describe wages, remuneration, or subsidies. It carries a strong emotional undertone of resignation, disdain, and self-deprecation, reflecting the worker's dissatisfaction and disappointment with their work conditions and quality of life. The term often highlights the irony of low pay in the face of arduous work, expressing a sense of self-reflection on one's value and discontent with the current circumstances.)

Table 13: Case of buzzword '窝囊费' and generated definitions by different methods. Here, we also provide its English translation for better understanding.

Prompt for Aspect-specific definition generation
根据以下所有[例句]，分析词语[BUZZWORD]的含义，将其总结成一句通顺且易理解的定义，并简要解释原因。 注意： 1. 用中文回答 2. 你需要从[INPUT_ASPECT]角度一步一步地思考这个词语的定义，这意味着去理解[INPUT_ASPECT_EXPLANATION] 3. 在观察用法示例时，要彻底解释上下文，以推断短语的微妙含义。将你的推理分解为循序渐进的逻辑，以达成全面的理解 4. 你不能过度解读这个词 5. 以Json形式返回结果：{"词语": "[BUZZWORD]", "定义": STRING, "原因"：STRING} [生成示例]: [EXAMPLES] ========== [例句]: [UGC_SENTENCES] Based on all the following [Example Sentences], analyze the meaning of the word [BUZZWORD], summarize it into a coherent and easy-to-understand definition, and briefly explain the reason. be careful: 1. Answer in Chinese 2. You need to think step by step about the definition of this word from the perspective of [INPUT_ASPECT], which means understanding [INPUT_ASPECT_EXPLANATION] 3. When observing usage examples, thoroughly explain the context to infer the subtle meaning of the phrase. Break down your reasoning into progressive logic to achieve a comprehensive understanding 4. You cannot overinterpret this word 5. Return the result in JSON format: {"Word": "[BUZZWORD]", "Definition": STRING, "Reason": STRING} [Example of Generation]: [EXAMPLES] ========== Example Sentences: [UGC_SENTENCES]

Table 14: Prompt for RESS and its corresponding translation: Part I.

Prompt for ensembling aspect-specific definition candidates

根据以下所有[例句]，分析词语[BUZZWORD]的含义，总结该词的[参考定义]成通顺且易理解的定义，包括但不限于本义、引申义和用法等等，并简要解释原因。
注意：
1. 用中文回答
2. 你需要根据[例句]一步一步分析该词[参考定义]的重要性，不是所有的[参考定义]都是有价值的。
3. 在分析时，要结合[例句]和[参考定义]，以推断[参考定义]的微妙含义，以达成全面的理解。
4. 以Json形式返回结果：{"词语": "[BUZZWORD]", "定义": STRING, "原因"：STRING}
[生成示例]: [EXAMPLES]
==========
[参考定义]: [CANDIDATE_DEFINITION]
[例句]: [UGC_SENTENCES]

Based on all the following [Example Sentences], analyze the meaning of the word [BUZZWORD], summarize its [Reference Definitions] into a coherent and easy-to-understand definition, including but not limited to its original meaning, extended meaning, usage, etc., and briefly explain the reasons.
be careful:
1. Answer in Chinese
2. You need to analyze the importance of the word [reference definition] step by step based on [Example Sentences], not all [Reference Definitions] are valuable.
3. When analyzing, it is necessary to combine [example sentence] and [reference definition] to infer the subtle meaning of [reference definition] in order to achieve a comprehensive understanding.
4. Return the result in JSON format: {"Word": "[BUZZWORD]", "Definition": STRING, "Reason": STRING}
Example of Generation: [EXAMPLES]
==========
Reference Definitions: [CANDIDATE_DEFINITION]
Example Sentences: [UGC_SENTENCES]

Table 15: Prompt for RESS and its corresponding translation: Part II.

Prompt for DP$_{\text{-w/o UGC}}$

给出以下互联网流行词或短语的定义。
注意，
1. 你给出的定义需要是简洁易懂的一句或多句话
2. 以json形式返回结果：{'word': STRING, 'definition': STRING}
词语： [BUZZWORD]

Return definitions of the following Internet buzzwords or phrases.
be careful,
1. The definition you provide needs to be concise and easy to understand
2. Return the result in JSON format: {'word ': STRING,' definition ': STRING}
Words: [BUZZWORD]

Table 16: Prompt for DP$_{\text{-w/o UGC}}$, which is also used in contamination-free evaluation experiments.

Prompt for Aspect-specific definition generation

给定一个词语的【定义】和专家给出的【参考定义】，你需要从以下【评估角度和打分标准】，为这个【定义】的质量高低评分。
使用Json格式返回结果：{"准确性": [INT, WHY], "细节完整性": [INT, WHY]}

【定义】：[PREDICTED_DEFINITION]
【参考定义】：[GROUND_TRUTH_DEFINITION]
【评估角度和打分标准】：
==========
准确性：
1分：该定义与【参考定义】相比，严重偏离了词语的真实意义，或者包含大量错误信息。
2分：该定义与【参考定义】相比，有一定的偏差，但至少部分正确。
3分：该定义与【参考定义】相比，基本准确，但可能存在一些小错误或不完整的描述。
4分：该定义与【参考定义】相比，准确，能够清晰地传达词语的核心意义。
5分：该定义与【参考定义】相比，非常准确，全面反映了词语的意义，没有遗漏重要细节。

细节完整性：
1分：该定义与【参考定义】相比，遗漏了许多重要的细节。
2分：该定义与【参考定义】相比，遗漏了一些重要的细节，但整体还算完整。
3分：该定义与【参考定义】相比，包含大部分必要细节，但仍有改进空间。
4分：该定义与【参考定义】相比，包含了几乎所有必要的细节。
5分：该定义与【参考定义】相比，包含了所有必要的细节，没有遗漏。

==========

Given a definition of a word and its Reference Definition provided by experts, you need to rate the quality of the definition from the following evaluation perspectives and scoring criteria.
Return result in JSON format: {'SA': [INT, WHY], 'SC': [INT, WHY]}

Definition: [PREDICTED_DEFINITION]
Reference Definition: [GROUND_TRUTH_DEFINITION]
Evaluation perspective and scoring criteria:
==========
Semantic Accuracy (SA):
1 point: Compared with the Reference Definition, this definition deviates significantly from the true meaning of the word or contains a large amount of erroneous information.
2 points: This definition has some deviation compared to the Reference Definition, but at least partially correct.
3 points: Compared with the Reference Definition, this definition is generally accurate, but there may be some minor errors or incomplete descriptions.
4 points: Compared with the Reference Definition, this definition is accurate and can clearly convey the core meaning of the word.
5 points: Compared with the Reference Definition, this definition is very accurate and fully reflects the meaning of the words, without missing any important details.

Semantic Completeness (SC):
1 point: Compared with the Reference Definition, this definition misses many important details.
2 points: Compared with the Reference Definition, this definition has omitted some important details, but overall it is relatively complete.
3 points: Compared with the Reference Definition, this definition contains most of the necessary details, but there is still room for improvement.
4 points: Compared to the Reference Definition, this definition contains almost all necessary details.
5 points: Compared with the Reference Definition, this definition includes all necessary details without omission.
==========

Table 17: Prompt for GPT-4o evaluator.