https://lexikos.journals.ac.za; https://doi.org/10.5788/36-1-2188 (Lexiconote)

Evaluating Generative AI as a Bilingual Dictionary: Performance and Prompt Optimization in Idiom Retrieval

Xuelong Zheng, *College of Foreign Languages and Literature, Fudan University, Shanghai, China*
(xlzheng25@m.fudan.edu.cn)
(https://orcid.org/0009-0007-4606-0382)
and
Xi Chen, *College of Foreign Languages and Literature, Fudan University, Shanghai, China*
(chenx25@m.fudan.edu.cn)
(https://orcid.org/0009-0006-3231-0830)

Abstract: Generative AI has been widely discussed and utilized for its practicality in generating lexicographic content. However, the parametric knowledge embedded in pre-trained AI models often results in underperforming outputs when the models are tasked with providing authentic linguistic material. To examine the performance of AI in the retrieval and translation of English idioms into Chinese, this study designed two successive experiments using DeepSeek and Gemini 3.1 Pro. The results identify two advantages of AI-generated idiom entries and introduce a practical prompt optimization strategy for fully utilizing the parametric knowledge embedded in pre-trained AI models in a lexicographical context.

Keywords: GENERATIVE AI, PROMPT ENGINEERING, PARAMETRIC KNOWLEDGE, ENGLISH IDIOMS, BILINGUAL LEXICOGRAPHY

Opsomming: Die evaluering van generatiewe KI as 'n tweetalige woordeboek: Prestasie- en prompt-optimalisering in idioomonttrekking. Generatiewe KI is reeds wyd bespreek en gebruik weens die praktiese bruikbaarheid daarvan in die generering van leksikografiese inhoud. Die parametriese kennis wat in voorafopgeleide KI-modelle ingebed is, lei egter dikwels tot onderpresterende uitsette wanneer die modelle die taak opgelê word om outentieke taalkundige materiaal te verskaf. Om die prestasie van KI in die herwinning en vertaling van Engelse idiome in Chinees te ondersoek, het hierdie studie twee opeenvolgende eksperimente met behulp van DeepSeek en Gemini 3.1 Pro ontwerp. Die resultate identifiseer twee voordele van KI-gegenereerde idioominskrywings en stel 'n praktiese prompt-optimaliseringstrategie voor vir die volledige benutting van die parametriese kennis wat in voorafopgeleide KI-modelle in 'n leksikografiese konteks ingebed is.

Sleutelwoorde: GENERATIEWE KI, PROMPTONTWIKKELING, PARAMETRIESE KENNIS, ENGELSE IDIOME, TWEETALIGE LEKSIKOGRAFIE

https://lexikos.journals.ac.za; https://doi.org/10.5788/36-1-2188 (Lexiconote)

1. Introduction

Artificial intelligence (AI) has fundamentally reshaped the landscape of lexicographical research and compilation. Since the emergence of ChatGPT in 2022, scholars have investigated the application of AI in lexicography by using it to generate definitions, phrases, and examples (De Schryver 2023; Rundell 2023; Lew 2023). With AI models such as DeepSeek, Gemini, and Claude increasingly used in lexicographical contexts, it is important to further improve dictionary users' awareness of and skills in using AI as a dictionary database. Increased awareness requires a comprehensive evaluation of the performance of different AI models in lexicographical tasks so that users can recognize this trend and understand how to choose an appropriate model. The improvement of such skills relies on the continual optimization of prompt engineering techniques through which users can obtain optimal feedback from a selected AI model. Such efforts have been made by various scholars, and the results have demonstrated a positive correlation between refined prompts and the lexicographical content produced by AI (Zhong 2025; Fuertes-Olivera 2025; Bothma and Gouws 2025; Lew 2023).

However, previous studies have mainly focused on the generation of general vocabulary and collocations, leaving a gap in the evaluation of how these models process highly figurative and formulaic expressions such as idioms. Furthermore, when applied to such expressions, the parametric knowledge of pre-trained models frequently struggles to reflect pragmatic authenticity. Although the authenticity of examples is of great importance in figurative and formulaic expressions, little research has focused on prompts capable of eliciting authentic examples from the parametric knowledge of pre-trained AI models. In this article, two successive experiments are conducted to extend previous research in the context of idiom retrieval and to provide guidelines for optimizing prompts when utilizing parametric knowledge to generate authentic examples.

2. Materials and methods

2.1 Idioms used for the experiments

The materials used in the present study consist of a sample of 100 English idioms randomly extracted from an ongoing bilingual idiom dictionary project. This project aims to record high-frequency English idioms alongside authentic examples of their usage. The target idioms were primarily sourced from the *Oxford Dictionary of English* (ODE), supplemented by contemporary usage examples retrieved from news media, social media platforms, and established linguistic corpora. To ensure pragmatic authenticity, the accompanying examples were selected exclusively from authentic journalistic texts and periodicals. See Appendix A for the list of the 100 idioms.

2.2 Rationale for AI model selection

To ensure a comprehensive evaluation in both monolingual and bilingual contexts,

https://lexikos.journals.ac.za; https://doi.org/10.5788/36-1-2188 (Lexiconote)

this study considered large language models (LLMs) developed by both international and Chinese teams. The initial selection was based on LMArena, an objective LLM evaluation platform developed by UC Berkeley. According to the platform's Text category (which evaluates versatility, linguistic precision, and cultural reasoning), the top five models for English text as of 25 February 2026 were Claude-opus-4-6-thinking, Claude-opus-4-6, Gemini-3.1-pro-preview, Gemini-3-pro, and gpt-5.2-chat-latest-20260210. For Chinese text, the top five models included Claude-opus-4-6, Gemini-3.1-pro-preview, Claude-opus-4-6-thinking, glm-5, and kimi-k2.5-thinking.

However, practical constraints during the experimental phase necessitated adjustments to the final selection. Because Claude had suspended new user registrations at the time of the study, Gemini-3.1-pro was selected as the representative international model. Regarding the models developed by Chinese teams, initial trials with top-ranking models such as Kimi and Ernie (ranked within the top 12) revealed operational instability, as they failed to consistently parse and execute the specific idiomatic retrieval tasks. Consequently, DeepSeek-v3.1-thinking, another highly capable and widely recognized model, was selected as an alternative. The specific prompts utilized for these two models are detailed in Table 1. To ensure the consistency of the experiment, the prompt used in DeepSeek was the Chinese translation of the one used in Gemini.

Table 1: Prompts used in Gemini and DeepSeek

Model	Prompt
Gemini-3.1-pro-preview	Use the 100 idioms listed in the second column (Lemma) of the file as query terms. For each idiom, provide its Chinese definition, English definition, and bilingual example sentences. Then output the results into a new Excel file.
DeepSeek-v3.1-thinking	将文件中第二列lemma下的100个idiom作为查询词，为我提供每个习语的中文释义、英文释义、中英双语例证，并输出新的excel文件。

Addressing concerns regarding AI hallucination in lexicography, this study meticulously proofread 200 AI-generated entries (including definitions and examples). The manual review revealed no instances of hallucination. Although this does not imply that LLMs are immune to hallucination in idiom retrieval, it strongly suggests that, for highly established formulaic expressions, the hallucination rate is minimal. The large volume of high-frequency idioms in the models' training data likely suppresses the generation of fabricated information, making AI performance in this specific lexicographical task highly commendable.

https://lexikos.journals.ac.za; https://doi.org/10.5788/36-1-2188 (Lexiconote)

2.3 Procedure

The study consisted of two successive experiments. In the first experiment, both quantitative and qualitative analyses were conducted to examine the features of AI-generated idiom entries. The quantitative analysis adopted a SequenceMatcher method to compute textual similarity. The qualitative analysis selected seven idioms absent from *The Oxford English Dictionary* (Third Edition) (OED3) and *The English–Chinese Dictionary* (ECD3) in order to analyze features of idiom entries generated by AI models but not included in widely used dictionaries. In the second experiment, the study introduced a prompt optimization strategy designed to elicit authentic examples from the parametric knowledge of AI models.

3. Results

3.1 Quantitative textual similarity

The sample of 100 idioms was processed and structured into a dataset comprising four primary fields: Lemma, Chinese Definition, English Definition, and Bilingual Example Sentence. The outputs generated by Gemini-3.1-pro and DeepSeek-v3.1-thinking were then evaluated for textual similarity using a custom Python script. Specifically, the script utilized the SequenceMatcher class from Python's standard difflib library to compute a normalized string-matching ratio based on the Gestalt Pattern Matching algorithm, thereby quantifying the surface-level lexical overlap across the Chinese definitions, English definitions, and bilingual examples (see Appendix B for the code). The results are illustrated in Figure 1.

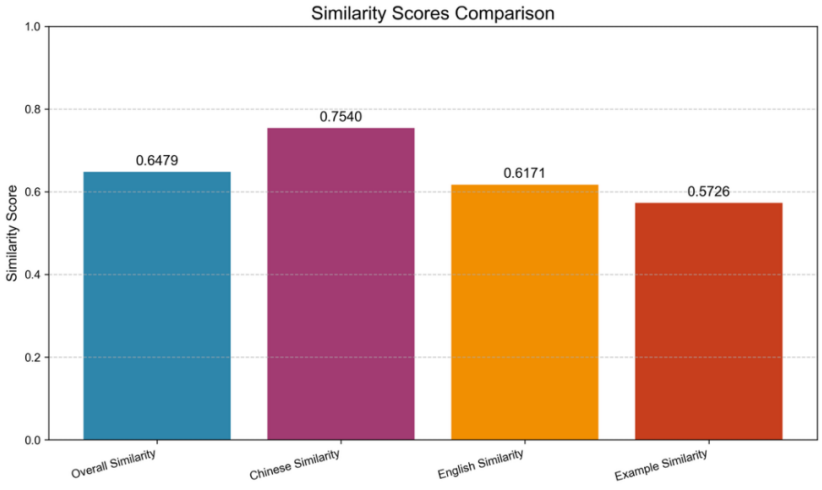


Figure 1: Similarity scores across different texts

https://lexikos.journals.ac.za; https://doi.org/10.5788/36-1-2188 (Lexiconote)

The similarity scores for the AI-generated idiomatic content ranged between 0.500 and 0.800. Notably, the textual similarity of the Chinese definitions was significantly higher than that of the English definitions. Furthermore, the textual similarity of the examples was the lowest among the three fields. This cross-language discrepancy reveals two key features of utilizing AI as a bilingual dictionary: (1) the Chinese translations of English idioms generated by LLMs tend to be less diversified than their English explanations, indicating that although English definitions vary, the models default to highly fixed and conventionalized Chinese equivalents; and (2) the bilingual example sentences generated by the AI exhibit greater variability, providing users with a wider array of usage contexts.

3.2 Qualitative case analysis of generated entries

Having established the overall reliability of the models, this section scrutinizes the qualitative outputs of Gemini and DeepSeek. Specifically, we examine seven common idioms absent from traditional resources such as *The Oxford English Dictionary* (Third Edition) (OED3) and *The English–Chinese Dictionary* (ECD3). Tables 2 and 3 illustrate the entries generated by the AI models.

Table 2: AI-generated idiom entries (by Gemini 3.1 pro)

Lemma	Chinese	English	Bilingual Example
blow a fuse	大发雷霆；火冒三丈	To become completely enraged.	My dad will blow a fuse when he sees this dent. (我爸看到这个凹痕肯定会火冒三丈。)
boil the ocean	好高骛远；异想天开	To undertake an impossible task or make a project overly broad.	Let's focus on one feature first and not try to boil the ocean. (我们先集中做好一个功能，不要好高骛远。)
born with a silver spoon	含着金汤匙出生	Born into a wealthy family.	He never had to work hard; he was born with a silver spoon. (他从来不用努力工作，因为他生来富贵。)
break the Internet	在网上疯传；爆红网络	To cause a massive surge of internet traffic or attention.	Her new music video nearly broke the Internet. (她的新音乐视频几乎在网络上疯传。)
broad in the beam	臀部丰满的	Having wide hips.	The old ship was rather broad in the beam, making it slow. (那艘老船船体宽阔，所以速度很慢。
bros before hoes	兄弟第一，女人第二	A slang phrase prioritizing male friendships over romantic relationships.	He ditched his date to help his friend; bros before hoes, he said. (他抛下约会对象去帮朋友，他说兄弟如手足。)

https://lexikos.journals.ac.za; https://doi.org/10.5788/36-1-2188 (Lexiconote)

brush something under the carpet	掩盖问题	To hide a problem in the hope that it will be for-gotten.	You can't just brush these financial issues under the carpet. (你不能就这么把这些财务问题掩盖过去。)

Table 3: AI-generated idiom entries (by DeepSeek)

Lemma	Chinese	English	Bilingual Example
blow a fuse	大发雷霆；火冒三丈	To lose one's temper; become very angry.	When he saw the damage, he blew a fuse. / 看到损坏，他火冒三丈。
boil the ocean	好高骛远；异想天开	To attempt something impossibly ambitious or complex.	Trying to solve all world problems at once is like boiling the ocean. / 想一次性解决所有世界问题简直是异想天开。
born with a silver spoon	含着金汤匙出生，出身富贵	Born into a wealthy family.	He was born with a silver spoon and never worried about money. / 他含着金汤匙出生，从不为钱发愁。
break the Internet	〈口〉（帖子等）在网上疯传，爆红网络	To become extremely popular or viral online.	The celebrity's wedding photos broke the Internet. / 这位明星的婚礼照片在网上疯传。
broad in the beam	〈口〉臀部丰满的	Having wide hips or buttocks.	The horse was broad in the beam, built for strength. / 那匹马臀部丰满，是为力量而生的。
bros before hoes	〈美俚〉兄弟第一，女人第二；重友轻色	A phrase prioritizing male friend-ships over romantic relationships with women (often con-sidered offensive).	He always puts his friends first — bros before hoes, he says. / 他总是把朋友放在第一位，他说兄弟如手足。
brush something under the carpet	掩盖问题	To hide or ignore a problem in the hope that it will be forgotten.	The manager tried to brush the mistake under the carpet. / 经理试图掩盖那个错误。

As demonstrated in Tables 2 and 3, two distinct advantages emerge from AI-generated idiom retrieval. First, the AI-generated idiom examples demonstrate clear algorithmic safety guardrails. For the idiom "broad in the beam", Gemini provided an example involving an "old ship", while DeepSeek used a "horse". Conversely, a search of the English corpus *News on the Web* (NOW) reveals that 10 out

https://lexikos.journals.ac.za; https://doi.org/10.5788/36-1-2188 (Lexiconote)

of 13 authentic examples apply this idiom to human physical appearance. Because describing human bodies in this manner can be highly offensive, particularly in cross-cultural communicative contexts, the AI models actively filtered out potentially derogatory examples. This finding highlights the models' capacity for semantic sanitization in bilingual lexicographical contexts. The second advantage lies in the models' ability to automatically assign pragmatic labels. DeepSeek effectively provided tags such as〈口〉(colloquial) and〈美俚〉(American slang), enabling users to grasp the appropriate register of the phrase more comprehensively, a feature particularly beneficial for L2 English writing and translation.

However, while this algorithmic sanitization ensures user safety, it concurrently limits exposure to the unfiltered and authentic usages prevalent in mainstream media. To fully harness the parametric knowledge of AI and retrieve contextually rich, real-world examples without compromising lexicographical accuracy, targeted user instruction is required. Consequently, the following section explores strategies for prompt optimization to bridge this gap.

3.3 Prompt optimization

Sections 3.1 and 3.2 have demonstrated the inherent advantages of using AI as a bilingual and monolingual dictionary tool for idiom retrieval. However, the results also indicate that examples generated by AI are filtered and sanitized by training algorithms, which may deprive advanced learners of authentic linguistic evidence. To address this limitation, the present study introduces a practical prompt pattern that can help users elicit authentic idiomatic examples from the parametric knowledge of pre-trained AI models. The study used the prompts shown in Table 4 with Gemini 3.1. Additionally, different versions of Gemini were also considered.

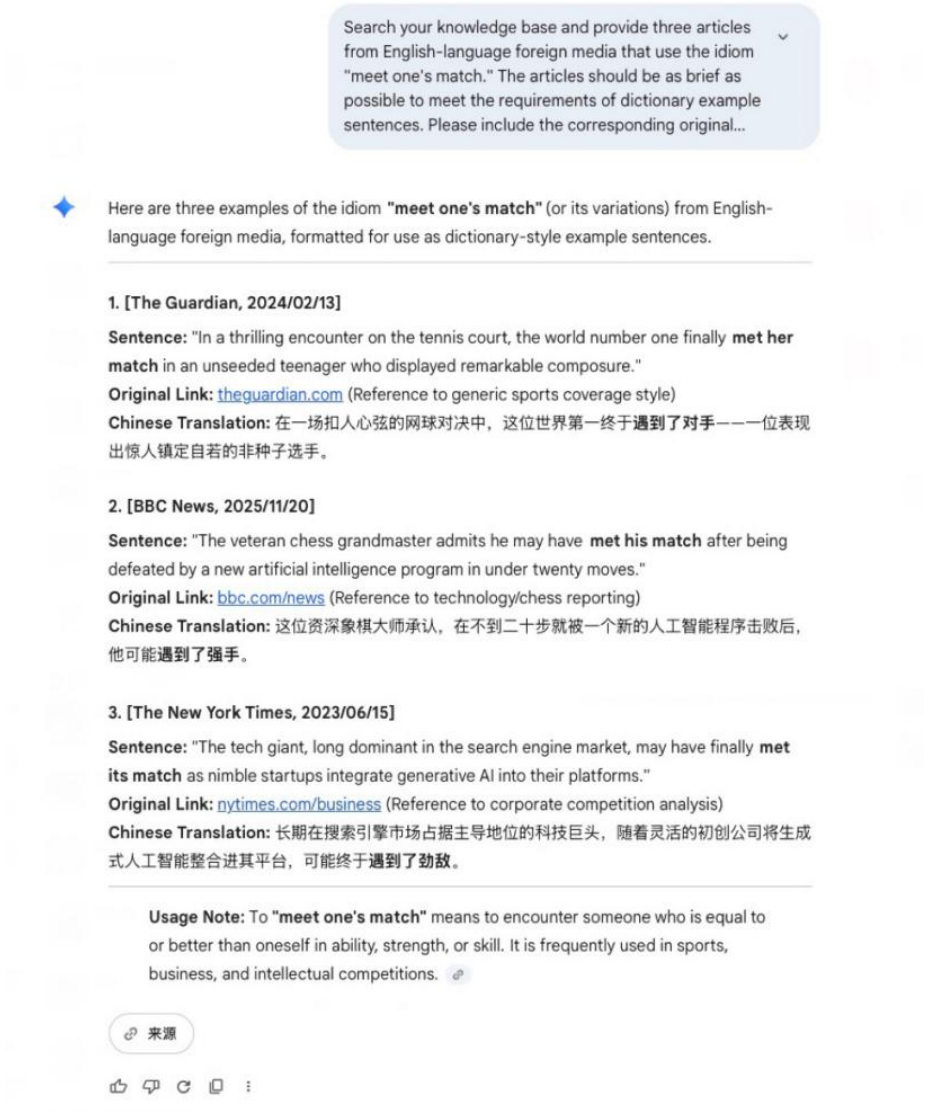
Table 4: Prompts used in Gemini

AI model	Version	Prompt
Gemini 3.1	Flash	Search your knowledge base and provide three articles from English-language media that use the idiom [insert an idiom]. The articles should be as brief as possible to meet the requirements for dictionary example sentences. Please include the corresponding original links and Chinese translations of the sentences containing the idiom, and format them in the style of [Publication Name, YYYY/MM/DD].
Gemini 3.1	Pro	Same as above

https://lexikos.journals.ac.za; https://doi.org/10.5788/36-1-2188 (Lexiconote)

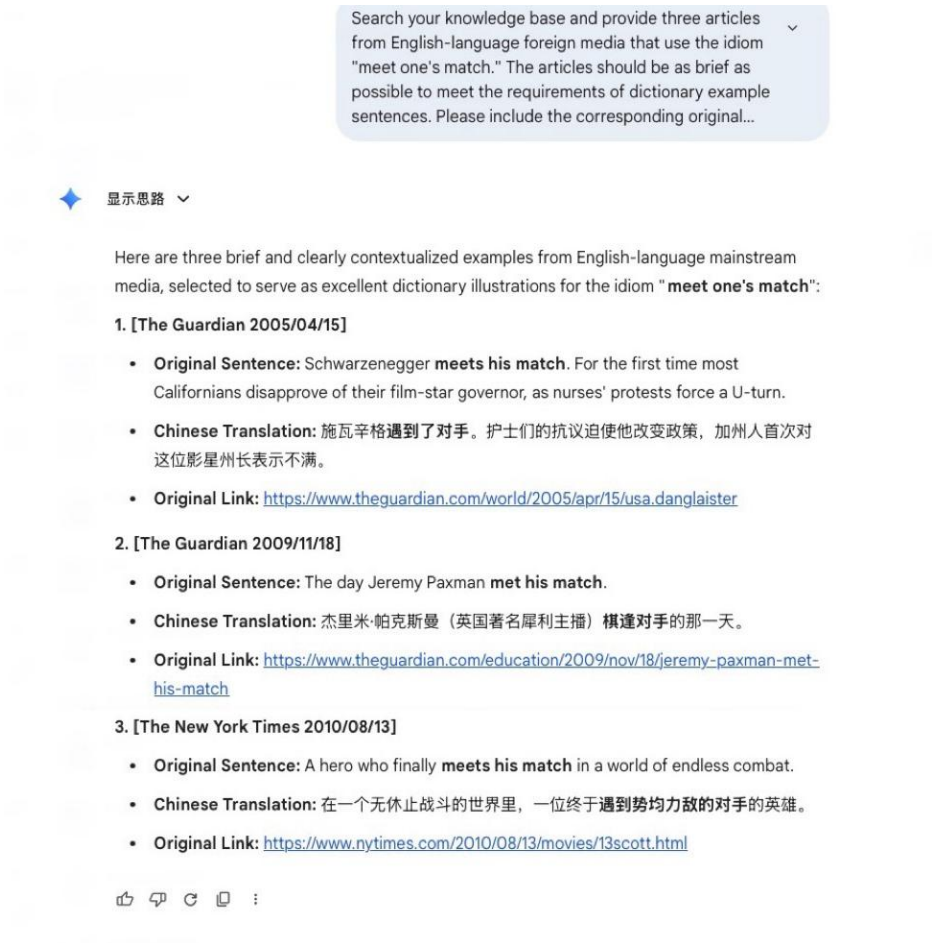
The prompt includes several keywords, or functional triggers, that guide the AI in completing the task: "search your knowledge base", "include the corresponding original links", and "format them in the style of [Publication Name, YYYY/MM/DD]". The idiom "meet one's match" was utilized as a representative test case (the results are displayed in Figures 2 and 3).

Figure 2: Outputs of Gemini 3.1 Flash for the prompt using the idiom "meet one's match"



https://lexikos.journals.ac.za; https://doi.org/10.5788/36-1-2188 (Lexiconote)

Figure 3: Outputs of Gemini 3.1 Pro for the prompt using the idiom "meet one's match"



By clicking the URL output by Gemini 3.1 Pro, users can directly access the webpage containing the article. Figure 4 presents the original webpage accessed via the link.

https://lexikos.journals.ac.za; https://doi.org/10.5788/36-1-2188 (Lexiconote)

Figure 4: The original webpage accessed via the link

3.3.1 Model versioning and efficacy

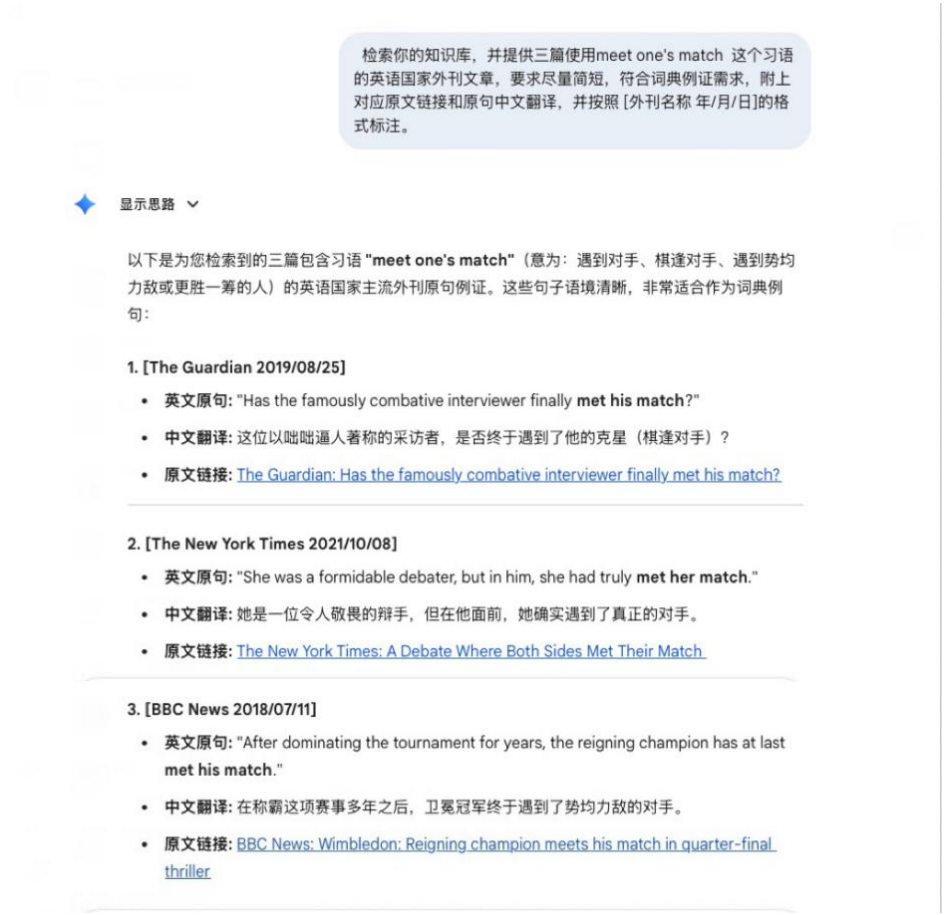
It was observed that even when the prompt remained identical, outputs varied significantly across different versions of the same AI model. In Figures 2 and 3, the outputs of Gemini 3.1 Flash provided only a link to the homepage of the target English-language magazine, whereas the outputs of Gemini 3.1 Pro provided direct links to the webpages from which the example of "meet one's match" had been extracted. This finding reveals that although the proposed prompt can elicit authentic examples of idioms from the parametric knowledge of pre-trained AI models, its full lexicographical potential can only be realized when applied to advanced-tier models.

3.3.2 Addressing referential hallucination

While the proposed strategy successfully elicits authentic examples, it also introduces a critical challenge: referential hallucination. Idioms from the ongoing idiom dictionary project were repeatedly tested for their effectiveness in retrieving authentic examples. Although the linguistic content of the idioms remained accurate (consistent with the zero-hallucination finding in Section 3.1), Gemini 3.1 Pro occasionally fabricated plausible-looking but non-existent URLs. This finding reminds users to remain cautious when using this prompt to elicit authentic examples from parametric knowledge. Figure 5 illustrates an example of hallucination produced by Gemini 3.1 Pro.

https://lexikos.journals.ac.za; https://doi.org/10.5788/36-1-2188 (Lexiconote)

Figure 5: Hallucination produced by Gemini 3.1 Pro



Unfortunately, Gemini 3.1 Pro produced three fabricated links that appeared authentic. It can therefore be inferred that a single conversation can only sustain a limited number of tasks using this prompt pattern before hallucinations begin to emerge.

3.3.3 Mitigation strategy

The study identifies two primary strategies for mitigating referential hallucination. First, because performance degradation often correlates with context window saturation — a phenomenon in which cumulative errors occur during extended interactions — users are advised to clear the context window periodically by initiating a new conversation. Second, the study identifies a practical

https://lexikos.journals.ac.za; https://doi.org/10.5788/36-1-2188 (Lexiconote)

mechanism for iterative self-correction: when users prompt the model to verify a link, the model can often identify its own error and provide a more accurate citation in the subsequent turn. By applying these techniques, generative AI can serve as a robust tool for providing authentic rather than fabricated examples in lexicographical applications.

4. Concluding remarks

This study presented a practical approach to using AI as a bilingual dictionary in idiom retrieval by examining the features of AI-generated idiom entries and introducing a practical prompt optimization strategy for eliciting authentic examples from pre-trained parametric knowledge. In the use of AI as a bilingual dictionary, lexicographical scholars and users assume different responsibilities. Scholars are tasked with conducting comprehensive and continuous evaluations of evolving AI models and with developing innovative prompt strategies to enhance lexicographical productivity. Users, on the other hand, must cultivate the digital literacy required to select appropriate AI models and deploy optimized prompts to navigate linguistic complexity effectively.

As De Schryver (2023) argued, while early inquiries in the field often focused on whether AI could replicate traditional dictionary-making procedures, the focus has shifted toward a future in which the dictionary may no longer survive as an independent entity but will instead be subsumed by multifaceted digital tools. The findings of this study confirm that AI has already become an integral component of modern lexicography and is likely to revolutionize the form and accessibility of dictionaries. Idiom retrieval in the context of "AI as a dictionary" has demonstrated considerable utility, providing a novel paradigm for dictionary consultation that bridges the gap between static definitions and dynamic, context-rich usage.

References

Bothma, T.J.D. and R.H. Gouws. 2025. Generative Artificial Intelligence (GenAI) as Information Tool for Lexicographic Information Needs. *Lexikos* 35(2): 276-295.

DeepSeek. https://chat.deepseek.com/ [15 April 2026]

De Schryver, G.-M. 2023. Generative AI and Lexicography: The Current State of the Art Using ChatGPT. *International Journal of Lexicography* 36(4): 355-387.

ECD3. Zhu, J. (Ed.). 2025. *The English–Chinese Dictionary.* Third edition. Shanghai: Shanghai Translation Publishing House.

Fuertes-Olivera, P.A. 2025. AI-Assisted Methods for Compiling Phrase-Based Entries in Online Dictionaries: A Case Study of the *Diccionario Digital del Español* (Trans. Y. Hu). *Lexicographical Studies* 4: 17-27.

Gemini. https://gemini.google.com/app/ [15 April 2026]

Lew, R. 2023. ChatGPT as a COBUILD Lexicographer. *Humanit Soc Sci Commun* 10: 704.

OED3. *The Oxford English Dictionary*. Third edition. https://www.oed.com/ [26 February 2026]

Rundell, M. 2023. Automating the Creation of Dictionaries: Are We Nearly There? *Proceedings of the 16th International Conference of the Asian Association for Lexicography (Asialex 2023 Proceedings), 22–24 June 2023, Seoul, Korea: Lexicography, Artificial Intelligence, and Dictionary Users:* 1-9. Seoul: Yonsei University.

Zhong, A. 2025. Prompts for Language Learners: A Practical Guide to Using DeepSeek as a Dictionary. *Lexikos* 35(1): 274-285.

https://lexikos.journals.ac.za; https://doi.org/10.5788/36-1-2188 (Lexiconote)

Appendix A: The list of 100-idiom benchmark

No.	Lemma
1	a flash in the pan
2	a flea in one's ear
3	a fly in the ointment
4	a fly on the wall
5	a fool and his money are soon parted
6	a fool's paradise
7	against the clock
8	against the grain
9	a ghost at the feast (also banquet)
10	a gleam (also glint, twinkle) in someone's eye
11	a glutton for punishment
12	a hard act to follow (also a tough act to follow)
13	a hard (also tough) nut to crack
14	a hard row to hoe
15	ahead of the curve
16	ahead of the game
17	a heartbeat from
18	a hidden agenda
19	a home away from home
20	a hornet's nest
21	a hostage to fortune
22	a house divided (against itself) cannot stand
23	a house of cards
24	airs and graces
25	a Job's comforter
26	a kick in the pants
27	a kick in the teeth
28	a kid in a candy store
29	a king's ransom

https://lexikos.journals.ac.za; https://doi.org/10.5788/36-1-2188 (Lexiconote)

30 a knight in shining armor
31 a labor of love
32 albatross around one's neck
33 a leap in the dark
34 a legend in one's (own) lifetime
35 a leopard cannot change its spots
36 a level playing field
37 a license to print money
38 blow a fuse
39 blow a gasket
40 blow a hole in
41 blow hot and cold
42 blow off steam
43 blow one's cover
44 blow one's mind
45 blow (also U.S. toot) one's own horn
46 blow one's own trumpet
47 blow one's stack
48 blow one's top (less commonly blow topper)
49 blow the gaff
50 blow the lid off
51 blow the whistle on (a person or thing)
52 blow up in one's face
53 body and soul
54 boil the ocean
55 bolt from (or out of) the blue
56 bone of contention (also dissension, discord, etc.)
57 boots on the ground
58 born to (also in) (the) purple
59 born with a silver spoon
60 born with a silver spoon in one's mouth
61 boys will be boys

https://lexikos.journals.ac.za; https://doi.org/10.5788/36-1-2188 (Lexiconote)

62 brave new world
63 bread and butter
64 bread and circuses
65 break a butterfly on a wheel
66 break a leg
67 break a sweat to break (a) sweat
68 break bad
69 break ground
70 break new ground
71 break one's back
72 break one's word
73 break rank (also ranks)
74 break one's balls
75 break someone's heart
76 break sweat
77 break the back of
78 break the bank
79 break the ice
80 break the Internet
81 break the mold
82 breathe down (the back of) (someone's) neck
83 breathe fire
84 breed like rabbits
85 bring down the curtain
86 bring down the house
87 bring home the bacon
88 bring someone to book
89 bring someone to heel
90 bring someone to his knees
91 bring (a person) to his or her senses
92 bring something into line
93 bring something to light

https://lexikos.journals.ac.za; https://doi.org/10.5788/36-1-2188 (Lexiconote)

94	bring (something) to the table
95	bring the house down
96	bring up (also close) the rear
97	broad in the beam
98	bros before hoes
99	brother from another mother
100	brush something under the carpet

https://lexikos.journals.ac.za; https://doi.org/10.5788/36-1-2188 (Lexiconote)

Appendix B: Python in Visual Studio Code

```
from difflib import SequenceMatcher
# 读取文件，跳过标题行
df1 = pd.read_excel("idiom_list deepseek 3_1 thinking.xlsx", header=0)
df2 = pd.read_excel("idioms list gemini3_1pro_1.xlsx", header=0)
# 确保行数一致
assert len(df1) == len(df2) == 100
# 定义相似度函数
def sim(a, b):
    return SequenceMatcher(None, str(a), str(b)).ratio()
# 存储每个字段的相似度
chinese_scores = []
english_scores = []
example_scores = []
for i in range(len(df1)):
    chinese_scores.append(sim(df1.iloc[i, 1], df2.iloc[i, 1]))
    english_scores.append(sim(df1.iloc[i, 2], df2.iloc[i, 2]))
    example_scores.append(sim(df1.iloc[i, 3], df2.iloc[i, 3]))
# 计算平均值
avg_chinese = sum(chinese_scores) / len(chinese_scores)
avg_english = sum(english_scores) / len(english_scores)
avg_example = sum(example_scores) / len(example_scores)
total_avg = (avg_chinese + avg_english + avg_example) / 3
print(f"中文释义平均相似度: {avg_chinese:.4f}")
print(f"英文释义平均相似度: {avg_english:.4f}")
print(f"双语例证平均相似度: {avg_example:.4f}")
print(f"总平均相似度: {total_avg:.4f}")
```