Do LLMs Know and Understand Domain Conceptual Knowledge?

Sijia Shen[1], Feiyan Jiang[1]*, Peiyan Wang[1]†, Yubo Feng[2], Yuchen Jiang[1],Chang Liu[1]
[1]Shenyang Aerospace University, [2]Dalian University of Technology
{shensijia, jiangfeiyan, jiangyuchen2, liuchang82}@stu.sau.edu.cn,
wangpy@sau.edu.cn, argmax@126.com

Abstract

This paper focuses on the task of generating concept sememe trees to study whether Large Language Models (LLMs) can understand and generate domain conceptual knowledge. Concept sememe tree is a hierarchical structure that represents lexical meaning by combining sememes and their relationships. To this end, we introduce the Neighbor Semantic Structure (NSS) and Chain-of-Thought (CoT) prompting method to evaluate the effectiveness of various LLMs in generating accurate and comprehensive sememe trees across different domains. The NSS, guided by conceptual metaphors, identifies terms that exhibit significant external systematicity within a hierarchical relational network and incorporates them as examples in the learning process of LLMs. Meanwhile, the CoT prompting method guides LLMs through a systematic analysis of a term's intrinsic core concepts, essential attributes, and semantic relationships, enabling the generation of concept sememe trees. We conduct experiments using datasets drawn from four authoritative terminology manuals and evaluate different LLMs. The experimental results indicate that LLMs possess the capability to capture and represent the conceptual knowledge aspects of domain-specific terms. Moreover, the integration of NSS examples with a structured CoT process allows LLMs to explore domain conceptual knowledge more profoundly, leading to the generation of highly accurate concept sememe trees.

1 Introduction

Large Language Models (LLMs) are regarded as versatile tools for various tasks, such as recommendations, language learning, and writing (Moskvoretskii et al., 2024; Kasneci et al., 2023). Previous works propose that LLMs are capable of encoding a significant amount of knowledge (Petroni et al., 2019). This has led researchers to determine to explore the kinds of knowledge within LLMs. Existing probing works mainly focus on factual knowledge(Safavi and Koutra, 2021), ontology knowledge (Wu et al., 2023) , lexical semantics knowledge (Moskvoretskii et al., 2024), and terminology knowledge(Jhirad et al., 2023a). Although existing research investigates terminological knowledge through generating term definitions in specialized domains like finance, current literature systematically neglects domain conceptual knowledge.

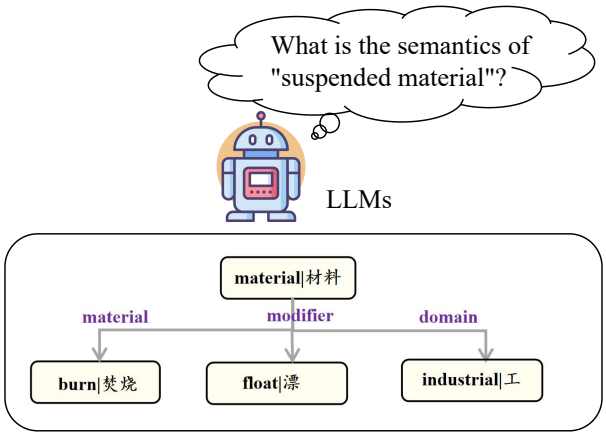


Figure 1: The conceptual knowledge of the term "suspended material".

Domain conceptual knowledge refers to a collection of core concepts, terms, and their interrelationships within a specific field, typically conveyed through terminology (Liu and Wang, 2019). Terms facilitate the transmission of domain knowledge, encompassing their definitions, relationships, and practical significance. For instance, the conceptual knowledge of the term "suspended material" (as shown in Figure 1) can be deconstructed into a three-dimensional semantic framework: its foundational definition ("burn" as material prop-

*Sijia Shen and Feiyan Jiang contributed equally to this work.
†Corresponding author.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 5967–5976
November 4-9, 2025

erty), functional characterization ("float" as modifier), and domain-specific context ("industrial" application). This tree structure demonstrates how domain concepts integrate core attributes, operational characteristics, and disciplinary embeddings. Domain conceptual knowledge plays a vital role in many natural language processing tasks such as machine translation (ElFqih and Monti, 2024; Ailem et al., 2021), text understanding (Piskorski and Stefanovitch.etc, 2023), and disease diagnosis normalization (Fan et al., 2024). Therefore, it is essential to explore whether LLMs can understand domain conceptual knowledge and possess a semantic understanding of it, rather than merely memorizing its superficial form.

In this paper, we systematically investigate whether LLMs possess domain conceptual knowledge and can understand the terminology. Specifically, we focus on generating concept sememe trees, defined as hierarchical structures that encode lexical semantics through systematic combinations of sememes (the smallest semantic units) and their relational mappings. To achieve this, we designed prompts for LLMs that include task introductions, output format requirements, examples of NSS, CoT (Zhang et al., 2022) guidance, and input terms. Based on terminological metaphor theory, we propose an NSS method to identify terms with similar sememe trees to the input terms, using these as examples. The CoT approach encourages LLMs to generate sememe trees through a structured process: generating the first sememe, producing additional sememes, establishing relationships, and forming the final output.

A comprehensive evaluation of 3 state-of-the-art LLMs, including GPT4o, LLAMA3-8B, and DeepSeek-V3, is conducted using benchmark datasets derived from four authoritative terminology manuals. The results of our experiments indicate that LLMs exhibit a measurable ability to generate semantic concepts associated with terms. By integrating NSS exemplars and guiding the models through step-by-step CoT reasoning, we developed a systematic framework. This framework enables LLMs to perform multi-perspective analyses that encompass intrinsic conceptual cores, defining attributes, and semantic relationships inherent in terminological structures. This methodological approach enabled the models to penetrate deeper into the foundational conceptual knowledge of the domain and accurately construct concept sememe trees. Our contributions are highlighted as follows.

- We leverage the task of concept sememe tree generation to evaluate the ability of LLMs to memorize domain concept knowledge.
- We demonstrate that LLMs possess a certain level of domain conceptual knowledge and are capable of understanding the meanings of terminology.
- Based on the theory of terminological metaphor, this study introduces examples of adjacent semantic structures to illustrate the hierarchical relationships between terms, thereby facilitating the model's understanding of metaphorical and semantic connections among terms and enhancing the effectiveness of sememe tree generation.

2 Method

Our method is divided into six stages, as shown in Figure 2. Firstly, prepare the data, terms and term definitions; Secondly, three datastores are constructed to obtain the key-value pairs with the highest cosine similarity. By summarizing the terms and counting their frequencies, we identify and select the top five terms with the highest occurrence times, and used these terms as NSS. Thirdly, CoT is added to provide a thinking process for LLMs. Finally, post-processing is carried out to obtain the sememe tree.

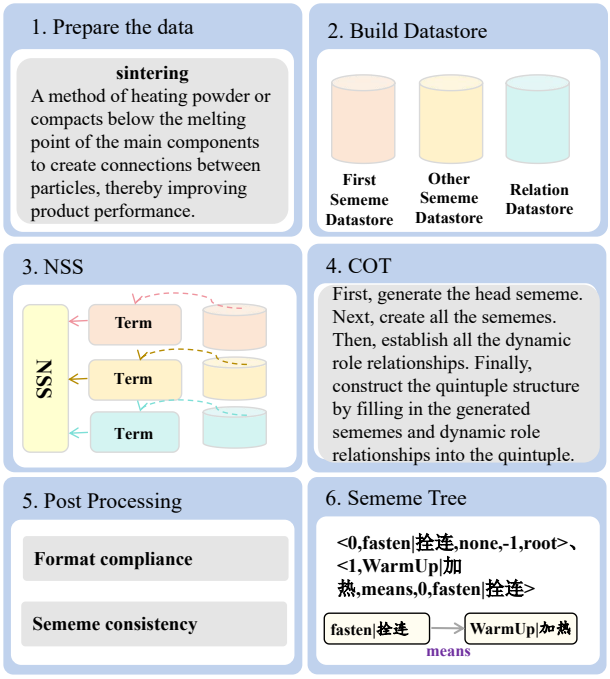
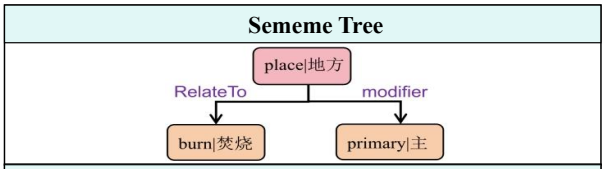


Figure 2: Our method flowchart.

2.1 Sememe Trees

To probe LLM's memorization of domain concept knowledge, we leverage the task of concept sememe tree generation. A sememe tree is a hierarchical structure that represents lexical meaning by amalgamating sememes (the smallest semantic units) with relations. This is the knowledge representation approach employed in HowNet (Dong et al., 2010). Concept sememe tree generation aim to automatically generate a sememe tree t for a given term e. For example, the sememe tree for the term "main combustion zone" is illustrated in Figure 3. "place|地方" serves as the first sememe, representing the core concept of the "main combustion zone" as a specific location, while "burn|焚烧" and "primary|主" are sememes that specify the main function and attribute of the "main combustion zone", respectively. "RelateTo" and "modifier" are the relations that reveal the correlation and modification between attributes and the core concept.



KDML

{place|地方:RelateTo={burn|焚烧},modifier={primary|主}}

List

["place|地方", ["RelateTo", "burn|焚烧"], ["modifier", "primary|主"]]

JSON

{"sememe": "place|地方", "children": [{"role": "RelateTo", "sememe": "burn|焚烧"},{"role": "modifier", "sememe": "primary|主"}]}

Quintuple

<0,place|地方,none,-1,root>、<1,burn|焚烧,RelateTo,0,place|地方>、<1,primary|主,modifier,0,place|地方>

Figure 3: The concept sememe tree of the term "main combustion zone" and the four data formats of the concept sememe tree.

2.2 Data Format of Sememe Tree

This paper employs four distinct ways of representing the sememe tree: KDML, List, JSON, and Quintuple. These four representations serve as constraints (r) on the output format of the LLM. The four data formats of the concept sememe tree for the term "main combustion zone" are illustrated in Figure 3.

- **KDML**: The Knowledge Database Markup Language is used by HowNet to represent the concept sememe tree. The format of KDML is "$\{s_f : r_1 = \{t_1\}, r_2 = \{t_2\}\}$", where s_f is the first sememe, r_1 and r_2 are the relations, t_1 and t_2 are the sub-trees which comply with KDML format.
- **List**: The concept sememe tree is represented as a nested Python list. The elements in the list are enclosed in square brackets. The first element in the list is the first sememe. For the subtrees, the first element represents the relationship, followed by the subsequent sememes.
- **JSON**: The JSON structure of the concept sememe tree represents a hierarchical object with a top-level key named "sememe" that holds the first sememe. It also includes a nested array under the key "children", where each element is an object containing two keys: "role" and "sememe". "sememe" represents the child sememe of the first sememe in the sememe tree. "role" represents the relation between the first sememe and the child sememe.
- **Quintuple**: The format of quintuple is "$< d_1, s_1, r, d_2, s_2 >$", where d_1 and d_2 are the depths of the sememes s_1 and s_2 in the concept sememe tree, respectively. And r is the relation between s_1 and s_2.

2.3 Prompt

Prompt is an intuitive method. We define the appropriate prompt p_d to leverage LLM to generate the concept sememe tree t of a given term e. This process is illustrated in Equation 1.

$$t = LLM\left(p_d\left(i, r, c, s(e), x(e)\right)\right) \quad (1)$$

where i is the task instruction that describes the task requirements. r is the rule that constrains the output format of LLMs. c refers to the CoT, which encourages LLMs to generate a sememe tree. This is accomplished through a series of steps: generating the first sememe, producing other sememes, establishing relations, and forming the final output. $s(e)$ are the few-shot examples, which include a few terms and sememe trees that may have a similar conceptual structure to the input term. $s(e)$ are

obtained through the NSS method based on terminological metaphor theory. $x(e)$ is the input term and its definition.

The prompt incorporates placeholders for various components required for selection parsing. Specifically, the symbols "[Rule]", "[CoT]", "[Example]", and "[Input]" are utilized as placeholders, representing the output format rule r, CoT c, few-shot examples $s(e)$, and the input term along with its definition $x(e)$, respectively. A prompt template example is shown in Figure 4.

2.4 Neighbor Semantic Structure

Many terms are not isolated, instead, they are interconnected within a hierarchical relational network, displaying distinct external systemic properties. During the term-mapping process, some terms retain the semantic characteristics of the source domain and evolve through hierarchical relationships based on terminological metaphors (Liu and Liu, 2024; Kasneci et al., 2019). This process transcends traditional systematicity, enabling cross-domain semantic expansion.

Based on this phenomenon, we propose that terms are structured not only through systematic hierarchical relationships but also through metaphorical associations. We introduce the concept of a "Neighbor Semantic Structure" to describe terms that share these relationships. A Neighbor Semantic Structure is a collection of terms from various domains that possess similar semantic connotations, achieved through retaining core meanings and using metaphorical expansion. For instance, the terms "wings" (biology), "airfoils" (aerospace engineering), and "cicada forewings" (entomology) originate from distinct disciplines yet share the core concept of a "wing" — an extended appendage that enables flight, gliding, or balance maintenance, as shown in Figure 5. This metaphorical mapping mechanism not only enriches the semantic dimensions of terminologies but also establishes critical cognitive pathways for cross-domain knowledge transfer.

We use NNP-TDGM (Sijia et al., 2024) to derive the NSS of a term. Given a term e, NNP-TDGM generates an initial sememe representation vector v_f, other sememe representation vectors V_s, and the relation representation vectors V_f of the sememe tree. These vectors are subsequently employed to query three distinct datastores: First Sememe Datastore, Other Sememe Datastore, and Relation Datastore. Each datastore consists of key-value pairs (k, s), where k represents the vector derived from NNP-TDGM, and s corresponds to the associated term. The (k, s) pairs with the highest cosine similarity to v_f, V_s, and V_f are obtained. Through a process of summarization and frequency counting for the term s, we identify and select the 5 terms with the most occurrences, using these terms as examples $s\,(e)$ of the NSS for the given term e.

2.5 Post Processing

Despite the constraints imposed on the output format within the prompt, the outputs from LLMs occasionally fail to fully adhere to the specified rules. To address this issue, we employ post-processing techniques to reformat the concept sememe trees generated by LLMs. This involves two key steps: format compliance processing to ensure adherence to the required structure, and sememe consistency processing to maintain the integrity and coherence of the semantic information.

- **Format compliance**. The generated concept sememe tree must adhere to the prescribed output form. Any sublists, subtrees, or quintuples that do not comply with the rules will be excluded.

- **Sememe consistency**. The sememe's Chinese and English words must correspond to the HowNet sememe set. If either the Chinese or English word is incorrect, we replace it with the corresponding correct sememe in HowNet. However, if both are incorrect and the corresponding correct sememe cannot be found, they remain unprocessed.

3 Experiment

3.1 Experimental Environment

The experiments are conducted on a Linux system running Ubuntu 16.04.7 LTS, utilizing NVIDIA A800 80G PCIe GPUs for acceleration.

3.2 Dataset

The experimental dataset is derived from four terminology manuals: Scientific Terms of Atmospheric Science, Terms in Mechanical Engineering (2nd Edition), Terms in Computer Science and Technology (Third Edition), and Terms in Electric Power (Third Edition). The dataset covers four core areas: atmospheric science, mechanical engineering, computer science and technology, and electric power engineering. Based on these, this study creates

Prompt
You are a concept semantics generator. Given a term and its definition, you generate a quintuple representation of the concept represented by the term. [Rule] ### You need to follow the steps outlined below to ultimately generate the quintuple representation of the concept represented by the term. [CoT] ### Example demonstration：[Example] ### Requirement: Given an input term and definition, please provide only the quintuple structure representing the concept of the term. [Input]
Rule
The quintuple structure is a set of "<odeep, osememe, role, fdeep, fsememe>". The structure must consist of five elements. odeep is an integer type and represents the depth of the head sememe in the conceptual semantic structure. osememe is a string type and represents the tail sememe. role denotes the dynamic role relationship. fdeep is an integer type and represents the depth of the tail sememe in the conceptual semantic structure. fsememe is a string type and represents the head sememe.
COT
First, generate the head sememe. Next, create all the sememes. Then, establish all the dynamic role relationships. Finally, construct the quintuple structure by filling in the generated sememes and dynamic role relationships into the quintuple.
Example
Term: Blade Row Definition: A series of blades arranged at equal intervals and with consistent installation angles. Thinking steps: First, generate the first sememe: ['part|部件'];Next, generate all sememes: ['root', 'part|部件', 'machine|机器', 'part|部件'];Then, generate all relationships: ['none', 'whole', 'whole'];Finally, the quintuple structure: <0, part|部件, none, -1, root>, <1, part|部件, whole, 0, part|部件>, <2, machine|机器, whole, 1, part|部件>
Input
Term: Blade Definition: A blade-shaped component that is formed by spatial stacking of the blade profile according to certain rules, or by directly modeling the aerodynamic design to create a spatial surface. It exchanges and converts energy with the airflow through this structure.
Output
<0,part|部件,none,-1,root>,<1,machine|机器,content,0,part|部件>

Figure 4: An example of prompt.

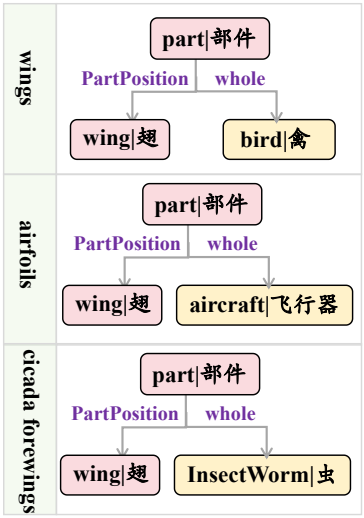


Figure 5: Sememe tree of wings, airfoils, and cicada forewings.

	Sememe	Relation	Triple
Mechanical	558	311	311
Computer	632	347	347
Electric	621	349	349
Atmospheric	537	294	294
	Term length	Definition length	Depth
Mechanical	2.60	29.46	2.34
Computer	2.99	38.14	2.75
Electric	2.92	38.10	2.75
Atmospheric	2.64	28.50	2.47

Table 1: Dataset information.

a bilingual dataset in Chinese and English to test the model's cross-language adaptability, ensuring a comprehensive and robust experimental evaluation. A total of 800 terms are evenly extracted from four datasets, with 200 terms selected from each dataset. These terms are then divided into a test set and a retrieval set in a 1:1 ratio. Table 1 presents the English dataset statistics, while its Chinese counterpart is omitted here due to complete consistency.

3.3 LLMs and Evaluation Metrics

We select three LLMs : LLAMA3-8B[1], GPT-4o(Hurst et al., 2024) and DeepSeek-V3[2]. Additionally, NNP-TDGM(Sijia et al., 2024) is selected as baseline. We use the F1 scores of triple (sememe-relation-sememe), relation, sememe, and first sememe as metrics. A higher F1 score indicates better recognition performance.

3.4 Results on Multi-Domain

Table 2, Table 3 and Table 4 show the evaluation results of multi-domain concept smeme tree gen-

[1] https://github.com/meta-llama/llama3
[2] https://github.com/deepseek-ai/DeepSeek-V3

eration on Chinese, English and Chinese-English Mixed datasets, respectively.

Dom	Model	Tri	Seme	Rela	First
A	NNP-TDGM	20.91	35.68	45.45	36.00
	LLAMA3-8B	15.10	22.96	41.99	31.30
	GPT4o	28.12	36.83	61.78	47.00
	DeepSeek-V3	30.06	35.99	58.63	52.00
M	NNP-TDGM	9.88	26.19	40.30	35.82
	LLAMA3-8B	10.82	15.20	43.57	24.22
	GPT4o	15.97	22.14	51.54	33.00
	DeepSeek-V3	21.04	25.04	54.42	44.00
C	NNP-TDGM	6.84	15.29	19.77	19.90
	LLAMA3-8B	6.78	12.99	35.59	17.94
	GPT4o	10.02	16.58	46.29	25.00
	DeepSeek-V3	12.52	18.62	43.98	34.00
E	NNP-TDGM	7.74	21.17	24.51	23.00
	LLAMA3-8B	8.76	14.34	38.25	19.72
	GPT4o	14.81	20.93	50.56	30.00
	DeepSeek-V3	14.93	23.28	48.66	29.00

Table 2: Performance evaluation of multi-domain concept sememe tree generation experiment on chinese dataset (%). The domain, Atmospheric, Mechanical, Computer, Electric, Triple F1, Sememe F1, Relation F1, and First Sememe F1 are abbreviated as Dom, A, M, C, E, Tri, Seme, Rela, and First, respectively.

The experimental results show that LLMs exhibit remarkable effectiveness in various domains. The experimental results confirm that LLMs possess domain conceptual knowledge and can understand the meanings of terms. Based on term metaphor theory, by introducing NSS and guiding the model step by step to consider the intrinsic core concepts, key attributes, and relationship types of terms, using the CoT method, the model can explore the fundamental conceptual knowledge of the domain more deeply, thus accurately and systematically describing the definitions of the terms.

In addition, under the mixed data of Chinese and English, the effect of the model in generating sememe trees is relatively good.

3.5 Results on Ablation Experiment

To assess the impact of NSS and CoT on model performance, we performed ablation experiments. The results are presented in Table 5. The symbol "−" indicates that the module is absent, while the symbol "✓" shows that the module is included. This experiment employed a quintuple prompt format and used the Qwen2.5-32B-Instruct model[3].

[3] https://github.com/QwenLM/Qwen2.5

Dom	Model	Tri	Seme	Rela	First
A	LLAMA3-8B	15.48	23.06	45.14	30.54
	GPT4o	32.84	42.53	65.26	49.00
	DeepSeek-V3	28.19	36.95	60.57	52.00
M	LLAMA3-8B	10.35	12.72	44.38	23.88
	GPT4o	19.01	21.94	58.87	31.00
	DeepSeek-V3	18.49	23.19	53.78	34.00
C	LLAMA3-8B	6.42	10.60	34.31	20.00
	GPT4o	10.47	17.33	46.57	25.00
	DeepSeek-V3	9.82	15.03	41.72	27.00
E	LLAMA3-8B	7.97	15.66	39.83	21.21
	GPT4o	13.52	19.06	51.99	28.85
	DeepSeek-V3	13.83	18.65	48.12	28.00

Table 3: Performance evaluation of multi-domain concept sememe tree generation experiment on english dataset (%). Abbreviations are consistent with table 2.

Dom	Model	Tri	Seme	Rela	First
A	LLAMA3-8B	17.63	24.36	43.59	32.73
	GPT4o	32.77	42.44	62.18	50.00
	DeepSeek-V3	31.37	42.74	61.57	56.00
M	LLAMA3-8B	12.65	16.97	47.53	26.01
	GPT4o	18.62	22.72	57.35	35.00
	DeepSeek-V3	19.67	25.00	54.00	44.00
C	LLAMA3-8B	6.51	12.71	37.52	16.35
	GPT4o	11.25	18.15	47.55	26.00
	DeepSeek-V3	11.74	18.79	43.96	31.00
E	LLAMA3-8B	10.55	16.67	41.11	18.92
	GPT4o	14.36	20.32	51.84	27.00
	DeepSeek-V3	15.97	23.32	50.48	29.00

Table 4: Performance evaluation of multi-domain concept sememe tree generation experiment on chinese-english mixed dataset (%).

Table 5 presents the experimental results showing that integrating NSS and CoT methods into the prompt significantly improves the model's understanding of terms concepts. Specifically, incorporating CoT method improved the F1 scores across multiple categories. The improvements were 4.4% for triples, 5.16% for sememes, 5.63% for relations, and 1.00% for first sememes. These results show that the CoT method guides models through a step-by-step process.This approach enables the model to explore the domain's fundamental conceptual knowledge more deeply.

NSS	CoT	Tri	Seme	Rela	First
✓	✓	28.21	37.86	53.57	52.00
✓	–	23.81	32.70	47.94	51.00
–	✓	0.00	0.00	0.00	0.00

Table 5: Ablation experiment (%).

The influence of NSS on the model is highly significant. In experiments without this structure, the four evaluation metrics dropped to zero. This phenomenon is primarily due to the lack of sufficient examples. Some large models have not been specifically trained to understand and generate definitions of specific terms or concepts. In such cases, incorporating the CoT method as a prompt makes it challenging for the model to establish a connection between the input terms and the output definitions, resulting in a sharp decline in performance. To more effectively demonstrate the advantages of the NSS, we compared NSS with fixed templates (FT). The experimental results are as shown in Table 6.

NSS	FT	Tri	Seme	Rela	First
✓	–	23.81	32.70	47.94	51.00
–	✓	2.63	7.42	30.66	11.00

Table 6: Experiment on the impact of FT and NSS (%).

Compared to the fixed template, the model's performance improved significantly. The F1 scores increased to 21.18% for triples, 25.28% for sememes, 17.28% for relations, and 40.00% for first sememes. These improvements were observed when the NSS was incorporated. The experimental results further show that providing examples enhances the model's performance. The NSS template, guided by conceptual metaphors, identifies terms with distinct external systematicity. These terms are located within a hierarchical relational network. Incorporating these terms as examples into the LLM's context learning enhances the model's ability to understand terms and generate conceptual definitions. This approach outperforms the fixed template.

3.6 Results on Different Output Format

To investigate the ability of LLMs to accept different output formats, we designed a output format study experiment.

The quintuple format achieves the highest values across all four metrics, indicating that it exhibits higher acceptance when LLMs perform tasks involving understanding and generating term concepts. This can be attributed to its straightforward, structured decomposition and the use of clear delimiters. The F1 values for the dictionary, list, and sememe tree formats are all below one percent. The reasons for this are twofold: firstly, the dictionary, list, and sememe tree formats are relatively complex, making it difficult for the model to grasp their internal structural logic. Secondly, this complexity hinders the model's ability to follow the predefined format during the generation process. In contrast, the absence of clear delimiters and structured guidance weakens the LLM's ability to adhere to the required structure during the generation process.

Format	Tri	Seme	Rela	First
JSON	0.10	2.98	5.70	4.04
List	0.03	2.62	5.88	4.02
KDML	0.00	3.51	9.90	6.00
Quintuple	2.63	7.42	30.66	11.00

Table 7: Results on different format (%).

3.7 Case Analysis

As shown in Table 8, when using the fixed template to generate the quintuple, the format "<3,motion|运动,patient,1,operate|操作>" contains a structural error. Based on the hierarchical structure of concept definitions, the subnode of the first-level node "operate|操作" cannot be directly located at the third level. If the error "<4, solid|固态, toState, 2, material|物质>" is excluded, the generated structure comprises three layers, which is inconsistent with the single-layer architecture of the correct answer.

Correct	<0,manual|非自动,none,-1,root>
FT	<0,action|动作,none,-1,root>、<1,operate|操作,content,0,action|动作>、<2,human|人类,agent,1,operate|操作>、<3,motion|运动,patient,1,operate|操作>
NS	<0,manual|非自动,none,-1,root>、<1,human|人力,agent,0,manual|非自动>
Ours	<0,manual|非自动,none,-1,root>

Table 8: Case analysis.

After incorporating the NSS, as presented in Table 9, the model successfully learns the quintuple forms of neighboring terms, such as "interconnection," "braking," "sliding," "starting," and "mo-

tion." The accurate generation of the structure"<0, manual |非自动, none, -1, root>" demonstrates the effectiveness of the NSS.

Input Terms	**Correct quintuple**
Manual	<0,manual|非自动,none,-1,root>
Neighbor term	The quintuple corresponding to the neighboring terms
Interconnection	<0,respond|回应,none,-1,root>、<1,EachOther|相互,manner,0,respond|回应>
Braking	<0,TurnOff|止动,none,-1,root>、<1,vehicle|交通工具,patient,0,TurnOff|止动>
Sliding	<0,slide|滑,none,-1,root>
Starting	<0,start|开始,none,-1,root>
Motion	<0,cease|停做,none,-1,root>

Table 9: Retrieval results of NSS.

3.8 Impact of Term Popularity

To assess the popularity of terms, we input them as search queries into the Google search engine and used the number of returned search results as a metric.

Table 10 presents the correspondence between term popularity percentiles and F1 scores of different evaluation metrics. The horizontal dimension of the table displays popularity percentile segments at 10% intervals, while the vertical dimension arranges four evaluation metrics with their F1 scores: Triple (Tri), Sememe (Seme), Relation (Rela), and First Sememe (First).

	0-10	10-20	20-30	30-40	40-50
Tri	10.62	14.87	17.22	16.24	13.66
Seme	15.93	21.10	22.01	20.42	18.51
Rela	43.36	47.48	49.44	48.26	47.62
First	27.50	32.05	35.04	32.70	29.08
	50-60	60-70	70-80	80-90	90-100
Tri	14.54	14.31	16.21	15.87	15.79
Seme	19.44	19.71	21.70	21.47	21.44
Rela	48.96	49.22	50.71	49.98	49.47
First	30.51	30.91	32.59	31.44	32.14

Table 10: Relationship between term popularity and model performance.

Table 10 reveals that terms in the high-popularity range (top) and the mid-to-low popularity range (40th−50th percentile) exhibit relatively low F1 scores. This phenomenon may be attributed to the higher ambiguity of terms in these ranges.

4 Related Work

A sememe tree represents a hierarchical structure designed to express lexical meaning through the combination of semantic primitives (the smallest semantic units) and the relationships among them. HowNet has been widely applied to various tasks, including word embeddings, word sense disambiguation, language modeling, and reverse dictionary construction. In recent years, research has increasingly focused on the automatic generation of semantic primitive trees for specific terms. (Zhang et al., 2014) developed a semantic primitive knowledge base for aviation-related terms using manual and semi-automatic methods. (Ye et al., 2022) proposed a Transformer-based model for generating sememe trees. While this method can determine whether a relationship exists between semantic primitives, it does not specify the type of relationship. (Sijia et al., 2024) introduced a term DEF generation model based on NNP-TDGM to address two key challenges: insufficient decoder training on low-frequency samples and limited encoding capacity. However, existing methods still have limitations. For instance, they rely heavily on labeled data and face challenges in leveraging large volumes of unlabeled data. Additionally, they exhibit substantial domain specificity. In contrast, large models offer improvements in these areas through large-scale pretraining and enhanced semantic understanding.

LLMs capabilities in lexical semantics and ontology construction have attracted significant scholarly interest. (Jain and Anke, 2022) proposed a zero-shot classification-based induction method to extract hypernym relations from LLMs, demonstrating that prompt-based guidance can effectively capture hierarchical relationships. This demonstrates that LLMs possess the ability to infer ontological knowledge, thereby facilitating term structuring. (Jhirad et al., 2023b) assessed the performance of LLMs in understanding financial terminology. Using definition modeling, they demonstrated the models' ability to generate precise domain-specific definitions through zero-shot and few-shot learning. (Moskvoretskii et al., 2024) analyzed the performance of LLaMA-2 and Mistral in classification learning tasks, revealing their capacity to acquire hierarchical knowledge and in-

fer conceptual relationships. Although existing research predominantly examines word semantics and ontology tasks, term definition modeling remains underexplored. This paper explores whether LLMs can comprehend terms and articulate their underlying conceptual semantics, offering valuable insights for term definition modeling.

5 Conclusion

In this study, we comprehensively investigate whether LLMs can effectively acquire domain-specific conceptual knowledge, moving beyond mere surface-level recognition to attain a deeper semantic understanding. Our experiments demonstrate that LLMs possess a measurable ability to generate semantic concepts associated with terms. Grounded in the theory of term metaphor, we introduce examples of semantically related structures and employ a chain-of-thought approach. This method guides the model to systematically analyze the core concepts, attributes, and relationships of terms in a sequential manner, thereby enhancing its ability to uncover underlying meanings.

Limitation

This approach facilitates meticulous extraction of domain-specific conceptual knowledge and methodical derivation of term definitions. However, it is important to note that both the understanding and reasoning about the semantics of term concepts by LLMs are imperfect, and the challenges they face when processing terms over the long term are evident. These observations indicate that their understanding of term concepts remains limited. Therefore, enhancing LLMs' understanding of term concept semantics represents a significant direction for future research.

Acknowledgement

This work was supported by the China National Committee for Terminology in Science and Technology (No. YB2022015) and the Applied Basic Research Program of Liaoning Province (No. 2022JH2/101300248).

References

Melissa Ailem, Jingshu Liu, and Raheel Qader. 2021. Encouraging neural machine translation to satisfy terminology constraints. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 1450–1455. Association for Computational Linguistics.

Zhendong Dong, Qiang Dong, and Changling Hao. 2010. Hownet and its computation of meaning. In *COLING 2010, 23rd International Conference on Computational Linguistics, Demonstrations Volume, 23-27 August 2010, Beijing, China*, pages 53–56. Demonstrations Volume.

Khadija ElFqih and Johanna Monti. 2024. Large language models as legal translators of arabic legislatives: Does chatgpt and gemini care for context and terminology? In *Proceedings of The Second Arabic Natural Language Processing Conference, ArabicNLP 2024, Bangkok, Thailand, August 16, 2024*, pages 111–122. Association for Computational Linguistics.

Yongqi Fan, Yansha Zhu, and Kui Xue.etc. 2024. Rrnorm: A novel framework for chinese disease diagnoses normalization via llm-driven terminology component recognition and reconstruction. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 9162–9175. Association for Computational Linguistics.

Aaron Hurst, Adam Lerer, and Adam P. Goucher. 2024. Gpt-4o system card. *CoRR*, abs/2410.21276.

Devansh Jain and Luis Espinosa Anke. 2022. Distilling hypernymy relations from language models: On the effectiveness of zero-shot taxonomy induction. In *Proceedings of the 11th Joint Conference on Lexical and Computational Semantics, *SEM@NAACL-HLT 2022, Seattle, WA, USA, July 14-15, 2022*, pages 151–156. Association for Computational Linguistics.

James Jhirad, Edison Marrese-Taylor, and Yutaka Matsuo. 2023a. 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.

James Jhirad, Edison Marrese-Taylor, and Yutaka Matsuo. 2023b. 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.

Enkelejda Kasneci, Kathrin Sessler, and Stefan K ü chemann.etc. 2019. Metaphor in terminology: Finding tools for efficient professional communicatio. *Fachsprache. Journal of Professional and Scientific Communication Special Issue 2019*, page 6586.

Enkelejda Kasneci, Kathrin Sessler, and Stefan K ü chemann.etc. 2023. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103:102274.

Chengpan Liu and Dongliang Liu. 2024. Systematicity of conceptual metaphors in aviation terminology. *Chinese Scientific and Technical Terminology*, 26(03):57–64.

Yungang Liu and Fenglong Wang. 2019. Decomposing the concept of territory in political geography. *Human Geography*, 34(01):14–19.

Viktor Moskvoretskii, Alexander Panchenko, and Irina Nikishina. 2024. Are large language models good at lexical semantics? A case of taxonomy learning. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pages 1498–1510. ELRA and ICCL.

Fabio Petroni, Tim Rocktaschel, and Sebastian Riedel.etc. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 2463–2473. Association for Computational Linguistics.

Jakub Piskorski and Nicolas Stefanovitch.etc. 2023. Multilingual multifaceted understanding of online news in terms of genre, framing, and persuasion techniques. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 3001–3022. Association for Computational Linguistics.

Tara Safavi and Danai Koutra. 2021. Relational world knowledge representation in contextual language models: A review. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 1053–1067. Association for Computational Linguistics.

Shen Sijia, Wang Peiyan, and Shengren.etc. 2024. NNP-TDGM: Nearest neighbor prompt term DEF generation model). In *Proceedings of the 23rd Chinese National Conference on Computational Linguistics (Volume 1: Main Conference)*, pages 57–70, Taiyuan, China. Chinese Information Processing Society of China.

Weiqi Wu, Chengyue Jiang, and Yong Jiang.etc. 2023. Do plms know and understand ontological knowledge? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 3080–3101. Association for Computational Linguistics.

Yining Ye, Fanchao Qi, Zhiyuan Liu, and Maosong Sun. 2022. Going "deeper": Structured sememe prediction via transformer with tree attention. In *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 128–138. Association for Computational Linguistics.

Guiping Zhang, Lina Diao, and Peiyan Wang. 2014. Construction of aviation terminology semantic knowledge base based on hownet. *Chinese Journal of Information Science*, 28(05):92–101.

Zhuosheng Zhang, Aston Zhang, and Mu Li.etc. 2022. Automatic chain of thought prompting in large language models. *CoRR*, abs/2210.03493.