# Inference Helps PLMs' Conceptual Understanding: Improving the Abstract Inference Ability with Hierarchical Conceptual Entailment Graphs

**Juncai Li**[1], **Ru Li**[1]*, **Xiaoli Li**[2,3], **Qinghua Chai**[1], **Jeff Z. Pan**[4]*,
[1]School of Computer and Information Technology, Shanxi University, China
[2]Institute for Infocomm Research, A*STAR, Singapore,
[3]A*STAR Centre for Frontier AI Research, Singapore
[4]ILCC, School of Informatics, University of Edinburgh, UK
juncaisev@163.com, {liru, charles}@sxu.edu.cn, xlli@i2r.a-star.edu.sg
http://knowledge-representation.org/j.z.pan

## Abstract

The abstract inference capability of the Language Model plays a pivotal role in boosting its generalization and reasoning prowess in Natural Language Inference (NLI). Entailment graphs are crafted precisely for this purpose, focusing on learning entailment relations among predicates. Yet, prevailing approaches overlook the *polysemy* and *hierarchical nature of concepts* during entity conceptualization. This oversight disregards how arguments might entail differently across various concept levels, thereby missing potential entailment connections. To tackle this hurdle, we introduce the *concept pyramid* and propose the HiCon-EG (Hierarchical Conceptual Entailment Graph) framework, which organizes arguments hierarchically, delving into entailment relations at diverse concept levels. By learning entailment relationships at different concept levels, the model is guided to better understand concepts so as to improve its abstract inference capabilities. Our method enhances scalability and efficiency in acquiring common-sense knowledge through leveraging statistical language distribution instead of manual labeling, Experimental results show that entailment relations derived from HiCon-EG significantly bolster abstract detection tasks. Our code is available at https://github.com/SXUCFN/HiCon-EG

## 1 Introduction

Cognitive research underscores *abstract inference ability* as the cornerstone of human cognition, empowering us to extrapolate and interpolate from past encounters, distill patterns, and adapt to novel scenarios (Saitta and Zucker, 2013). For instance, when humans comprehend "John presents his friend a book", they invariably perceive "John" and "his friend" as *Person*, "book" as an *Entity*, and abstract the event as "*PersonX present PersonY Entity*". This event can be further abstracted as "*PersonX give PersonY Entity*". In Natural Language Processing (NLP), "*PersonX present PersonY Entity*" is defined as the given premise and "*PersonX give PersonY Entity*" as the inferred hypothesis, constituting an textual entailment relationship.

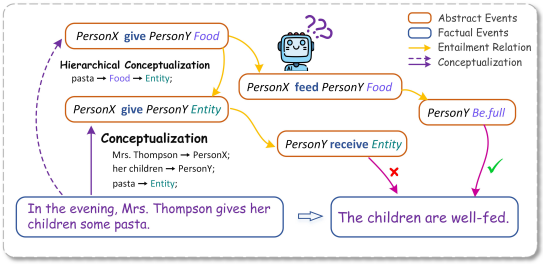



Figure 1: After conceptualizing events, models can infer more information from conceptualized events. However, different levels of conceptualization may lead to different entailment relationships. For example, when *pasta* is conceptualized as *food*, we can infer that *PersonY Be.full*.

In the evolution of Natural Language Inference (NLI), numerous studies delve into abstract challenges across various domains such as common sense reasoning (He et al., 2024, 2023; Romero et al., 2019), question answering (Zheng et al., 2024; Chen et al., 2022b), knowledge base explanation (Mellish and Pan, 2008), argumentation mining (Saadat-Yazdi et al., 2023), machine translation (Padó et al., 2009), and beyond. In this paper, our goal is to enhance the capability of conceptual knowledge for Pre-trained Language Models (PLMs) (Pan et al., 2023). Within this landscape, a pivotal effort by Wang et al. (2024c) introduced the ABSPYRAMID benchmark, aiming to comprehensively assess the abstraction prowess of PLMs through three entailment relationship types: nouns, verbs, and events. Despite advancements, evaluations reveal that even state-of-the-art PLMs struggle with abstraction, trailing behind fine-tuned smaller models. Hence, there's a pressing need for further research to better mine entailment relationships and bolster models' abstraction capabilities.

* Contact Authors



*Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22088–22104
November 12-16, 2024 

Entailment graphs, first proposed by Berant et al. (2010), are graphs with verbs as nodes and entailment relations as edges, which can be seen as subproperty relations (in natural language form) in a schema of knowledge graphs (Pan et al., 2017b,a). Entailment graphs aim to globally discover textual entailment relationships between verbs which is different from logical entailment (Pan and Horrocks, 2002; Pan et al., 2017a), and *textual entailment* has a more relaxed definition: "t entails h" ($t \vDash h$) if, typically, a human reading $t$ would infer that $h$ is most likely true (Dagan et al., 2006). Early work explored the entailment relationships between bivalent verbs based on global transitivity constraints (Hosseini et al., 2018, 2019, 2021). Subsequently, McKenna et al. (2021) extended the Distributional Inclusion Hypothesis (Dagan et al., 1994; Kartsaklis and Sadrzadeh, 2016), allowing the discovery of entailment relationships between verbs of different valences. This evolution has enabled entailment graphs to discover more diverse entailment relationships (e.g., *PersonX give PersonY Entity* $\vDash$ *PersonY receive Entity*) beyond the entailment relationships between synonyms. In these works, to disambiguate polysemous verbs, the arguments of verbs are usually typed (conceptualised) (Lewis and Steedman, 2013; Chen et al., 2022a), that is, these arguments are mapped to a *limited* finite number of basic types, such as Person, Location, Time, etc. Therefore, the nodes in the entailment graph are essentially events after abstraction, and the graph itself can be understood as a representation of abstract relations. These relationships include the abstract relations between vocabulary and their abstract concepts (*present* $\models$ *give*) , as well as the conceptualized commonsense reasoning relations between abstract events (*PersonX give PersonY Entity* $\models$ *PersonY receive Entity*).

Nevertheless, the limited argument types (Ling and Weld, 2021) used in the conceptualization of arguments often compromises the precision of events, resulting in inaccurate entailment relationships.

A single instance can be understood through a spectrum of concepts with varying levels of granularity (Minsky, 1980). For example, an apple can be seen as an object, food, fruit, etc. Different granularity levels reveal distinct entailment relations. In Figure 1, consider the sentence "Mrs. Thompson gives her children some pasta." If pasta is conceptualized as an entity, the inference is "*PersonY receives Entity*." Viewing pasta as food allows for richer inferences, such as "*PersonX feeds PersonY Entity* .

In this paper, we argue that entailment relationships at different concept levels can supplement richer verb entailment relationships and these relationships is helpful for the model to better understand the differences between noun concepts at different concept levels. Based on the entailment relations across various levels of conceptual granularity. we create a **Hi**erarchical **Con**ceptual **E**ntailment **G**raph (HiCon-EG). In particular, we introduce a *conceptual pyramid* (Minsky, 1980) for hierarchically conceptualizing arguments. This approach enables us to uncover entailment relations under various conceptual constraints.

To mitigate the sparsity issue stemming from the abundance of concepts, we propose a concept selection method grounded in entropy principles (Liu et al., 2022) to identify the most representative concepts, thereby reducing unnecessary computations.

Our contributions can be summarized as follows: 1. We propose a novel Complex-to-Simple open information extraction method based on large language models (LLMs), which facilitates the extraction of multivalent arguments from lengthy texts. To mitigate the hallucination problem associated with LLMs, we further distill stable, smaller models. This method outperforms existing approaches on specific datasets, demonstrating superior performance. 2. We introduce the "conceptual pyramid" for the hierarchical conceptualization of arguments, enabling the mining of entailment relations under diverse conceptual constraints. To reduce computational costs, we propose an entropy-based concept selection method for identifying appropriate concepts for arguments under different predicates. Experimental results demonstrate performance comparable to GPT-4, with lower error rates. 3. We evaluate the effectiveness of our method on abstraction detection and conceptualized commonsense reasoning tasks. Results indicate significant performance enhancements on the abstraction detection task, with a slight edge over the baseline on conceptualized commonsense reasoning datasets.

## 2 Related Work

**Entailment Graph.** Berant et al. (2010) introduced a graph-based framework centered on predicates, pioneering the task of constructing a verb entailment graph (Berant et al., 2011). Subsequently, several approaches grounded in global

transitivity constraints have emerged (Hosseini et al., 2018, 2019; Chen et al., 2022a). McKenna et al. (2021) extended the interpretation of DIH to support the learning of entailment relations between differently-valenced predicates, transforming the entailment graph into a tool for mining abstract reasoning relationships. McKenna et al. (2023) proposed a smoothing theory to optimize the entailment graph; Wu et al. (2023) leveraged pre-trained language model to generate scalable entailment graphs. However, the 49 basic types (Ling and Weld, 2021) used in the argument typing process lead to the loss of original semantics.

Recently, Wang et al. (2024c) introduced the ABSPYRAMID benchmark (including an abstract detection task) to evaluate models' abstraction capabilities, revealing that abstraction remains a challenge for LLMs. Wang et al. (2024b) proposed AbsInstruct, built instructions with in-depth explanations to assist LLMs in capturing the underlying rationale of abstraction. Zhou et al. (2024) introduced the product recovery benchmark, for entailment graphs in the E-commerce setting.

**Conceptualized Commonsense Reasoning.** The abstracted events exhibit certain cognitive inference relations, which can be mined to enhance the reasoning capabilities of models. In the domain of common-sense knowledge, He et al. (2024) introduced the AbstractATOMIC abstract common sense reasoning dataset based on ATOMIC (Sap et al., 2019). Subsequently, Wang et al. (2023, 2024a) proposed various frameworks based on conceptualization and instantiation to enhance the common sense reasoning capabilities of LLMs. However, constrained by ATOMIC, such work is specific to social common sense domains.

## 3 Our Proposed Approach

The construction of the hierarchical entailment graph commences with the extraction of multivalent arguments for predicates from the multi-source NewsSpike corpus (Zhang and Weld, 2013) using our proposed C2S-OIE method. Subsequently, we engage in a multilevel conceptualization of the extracted predicate arguments, selecting the most appropriate concept for each argument governed by different predicates. Finally, we compute the relevance score between pairs of predicates to construct the entailment graph (McKenna et al., 2021).

### 3.1 C2S Open Information Extraction (Step 1)

Previous research predominantly employed heuristic methods like the Combinatory Categorial Grammar semantic parser and Dependency parsers (Steedman, 2001) for open information extraction. However, these approaches struggle with Coreference Resolution (Yu et al., 2021) when faced with complex *nested* sentence structures commonly found in news corpora. For instance, in the sentence "Bob is the last student who left the laboratory", most methods incorrectly parse "who" instead of "Bob" as the subject of "left," leading to suboptimal results.

To tackle this issue, we propose a Complex-to-Simple open information extraction (C2S-OIE) method to effectively handle the challenges posed by complex nested sentences. As illustrated in Figure 2-Step1, this approach involves two key steps:

**Complex-to-Simple**: We prompt the large language model to generate simple expressions of complex sentences. Specifically, we employ LLaMa2-7B to decompose complex sentences into multiple simple sentences using in-context learning (Brown et al., 2020), ensuring that the arguments in each simple sentence are as complete as possible. The prompt we provide is as follows:

**<INSTRUCTION>**
<EX$_1$-I><EX$_1^1$-O>$\cdots$<EX$_1^k$-O>
$\cdots$
<EX$_n$-I><EX$_n^1$-O>$\cdots$<EX$_n^k$-O>
<Q-I>

Where **<INSTRUCTION>** outlines the task of sentence simplification, <EXi-I> and <EXi-O> represent the input examples of complex sentences and their corresponding simplified outputs, respectively, and <Q-I> is the input query containing the complex sentences. Detailed prompts are provided in Appendix A.1. Given the substantial data volume, utilizing LLMs would significantly increase our costs. Moreover, since the C2S task requires LLM to generate text rather than simple discrimination, it is more susceptible to hallucinations (Ji et al., 2023; Huang et al., 2023) (as shown in figure 3). So we distill the sentence simplification capability into a BERT model by selecting high-quality results. The fine-tuned BERT model achieves 95% accuracy and demonstrates greater stability than LLaMa, making it an effective and cost-efficient substitute for this task. The detailed process is documented in Appendix A.2.

**Semantic Role Labeling**: For the extracted sim-

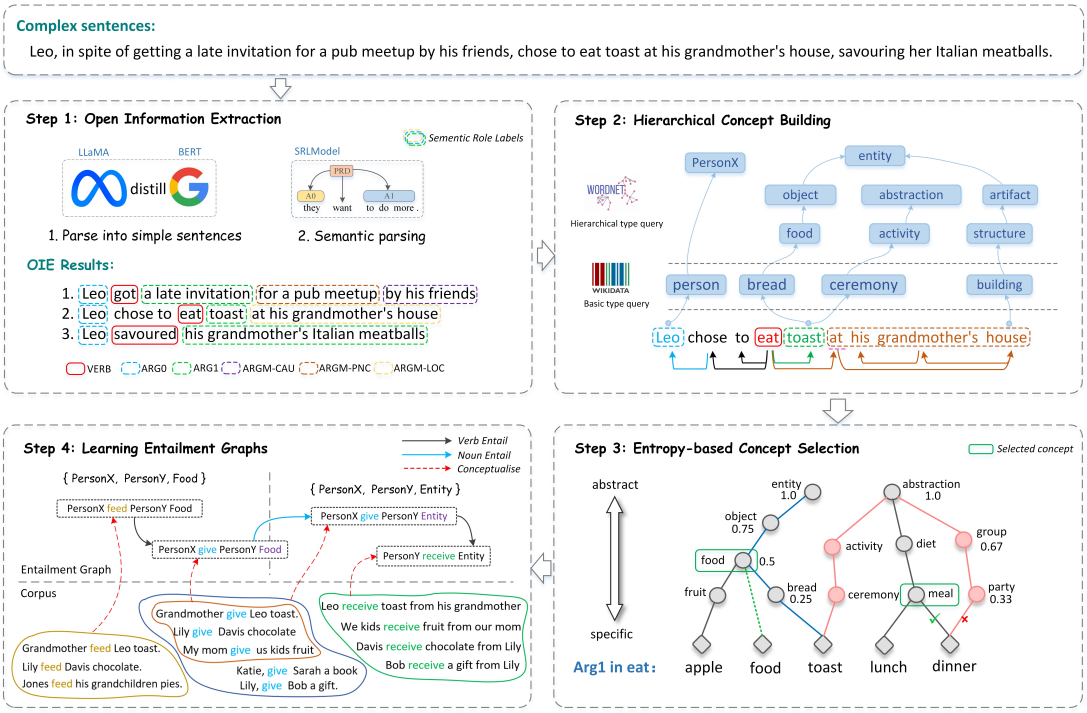



Figure 2: The summary of constructing the hierarchical concept entailment graph. The figure illustrates how a complex news corpus is processed through open information extraction to obtain arguments, conceptualised at different granularities, and finally learn entailment relations under different granularities of concepts.

plified sentences, we further analyze their semantic roles, employing a semantic role labeling model (Zhang et al., 2022b) to annotate the argument roles for each verb. The argument roles in Semantic Role Labeling are more *rich* and *detailed* (Pradhan et al., 2012), allowing us to filter out unnecessary arguments such as time and location. Due to the enhanced performance of the semantic role labeling model with simplified sentences, our proposed C2S-OIE method produces superior results compared to Open Information Extraction methods directly using long sentences (see Section 4.4).

### 3.2 Hierarchical Concept Building (Step 2)

Next, we perform hierarchical conceptualization on the verb arguments in each simplified sentence. As shown in Figure 2-Step2, 'toast' can be conceptualized into two groups like $[bread, food, entity, \ldots]$ and $[ceremony, activity, \ldots]$ from fine-grained to coarse-grained levels. This process is formalized as follows: given an argument core word $w$, hierarchical conceptualization constructs hierarchical concepts $C_i = [c_{i1}, c_{i2}, \ldots, c_{im}]$ and the set of all meanings $\rho = \{C_1, C_2, \ldots, C_n\}$ of $w$, where $c_{ij}$ represents the $j$-th level concept of meaning $C_i$. This is denoted as $\rho = HC(w)$, where $HC$ is the hierarchical conceptualization function and $\rho^{ij}$ is the

**Query Complex Sentence:** Amphibians control pests , may be a source of new medicines , and help feed many other animals .

**Output:**
✓ 1. Amphibians control pests
✓ 2. Amphibians may be a source of new medicines
✓ 3. Amphibians help feed many other animals

**Query Complex Sentence:** Hatch , raised in Fort Wayne , Indiana , moved in 2013 to his uncle 's home in California .

**Output:**
× 1. Hatch
× 2. raised in
× 3. Fort Wayne
.....

Figure 3: An example of the results of LLM simplifying complex sentences. We find hallucinations of LLM decomposing sentences into words or phrases leads to incorrect simplification results.

$j$-th level concept of $i$-th meaning of argument $w$.

**Pilot Study**: Existing knowledge bases like WordNet (Miller, 1995), Probase (Wu et al., 2012) and ConceptNet (Liu and Singh, 2004) contain extensive knowledge of lexical conceptualization, covering the meanings of general vocabulary. However, our pilot study reveals their limitations: Probase, constructed using data-driven methods, suffers from *cyclic errors*, with approximately 97% of erroneous relationships forming cycles (Liu et al., 2022) (e.g., the correct relation

isA(Paris, existing city) versus the incorrect relation isA(existing city, Paris) forming a cycle). In contrast, the hierarchical relationships in WordNet are manually constructed by language experts, ensuring higher quality. Nonetheless, WordNet's limited scale results in lower noun coverage (Wang et al., 2024c), particularly affecting: 1. Noun phrases, such as "fresh apple" and "lots of apples"; 2. Proper nouns, such as "COVID-19".

**Core-word based Conceptualization**:

Given the limited coverage of nouns in WordNet, which hinders effective querying for hierarchical conceptualization, we propose the following strategies to address this issue:

1. For certain personal pronouns (e.g., you, I, he), we assign them to the special type "Person".

2. To address the issue of phrases that cannot be directly conceptualized, as depicted in Figure 2-Step2, we retrieve the *core words* of the phrase using syntactic dependency (Manning et al., 2014) and treat them as the entire phrase (e.g. *house* from at *his grandmother's house*). Please refer to the appendix in Section A.3 for more details.

3. During conceptualization process, we consider all possible concepts for polysemous words in arguments. For instance, "toast" may represent concepts such as bread, food, or *the act of toasting*.

4. For core words not found in WordNet, we conceptualize them at the first level using Wikidata, *offering extensive noun coverage*. Subsequently, we obtain hierarchical concepts using WordNet.

### 3.3 Entropy-based Concept Selection (Step 3)

After hierarchical conceptualization, we generate many abstract concepts for each argument. Some may be incorrect, such as $(\text{eat}, \underline{\text{toast}}) \rightarrow$ activity, while others may be too broad or too specific, like $(\text{eat}, \underline{\text{toast}}) \rightarrow$ entity or $(\text{eat}, \underline{\text{toast}}) \rightarrow$ bread. This increases computational load and can lead to incorrect entailments. Thus, as shown in Figure 2-Step3, selecting a semantically accurate concept with appropriate granularity for each verb is crucial.

We define this task as follows: Given the set of all conceptualizations $\mathsf{Q} = \{\rho_1, \rho_2, \ldots, \rho_l\}$ of the core word set $W = \{w_1, w_2, \ldots, w_l\}$ for the role $r$ of the verb $v$, our objective is to select the most suitable concept for each $w$. In other words, we aim to obtain a sequence of concepts $\vartheta = \{t_1, t_2, \ldots, t_l\}$, where $t_i \in \rho_i$ represents the selected concept for the core word $w_i$. However, attaining the appropriate concept presents challenges that must be addressed:

1. How to discern the various semantic meanings of a polysemous verb based on the distinctions among argument concepts.

2. How to choose the correct semantic meaning for the arguments of a polysemous core word. For instance, "apple" can denote both a fruit and a company, but when paired with the verb "eat," consuming a company is clearly absurd.

3. How to ensure the selected concept can generalize across many core words in a series of similar instances, thereby enhancing generalization ability.

4. In practice, some argument concepts, such as "things" and "food" (as shown in 2-Step3), are already sufficiently abstract and may not require further conceptualization.

In addressing Challenge 1, Zhang et al. (2022a) suggests that, backed by extensive data, the frequency of correct and cognitively consistent $(verb, concept)$ pairs is significantly higher than that of incorrect combinations. Following this insight, when selecting concepts, we prioritize higher-frequency concepts to ensure consistency across selected concepts. Thus, when encountering pairs like $(eat, apple)$, we can confidently infer that the type of apple should be categorized as "food" rather than a "company," given that pairs like $(eat, banana)$ and $(eat, bread)$ are more prevalent in the corpus compared to $(eat, company)$.

Hence, entropy serves as a measure of uniformity for evaluating concept selection outcomes (Cover, 1999). Here, we define the objective function as:

$$\mathbb{L}(\vartheta) = H(\chi|v, W) = -\sum_{\tau_i \in \mathsf{S}} P(\tau_i) \log P(\tau_i)$$

Here, $\chi$ represents a discrete random variable conforming to the distribution of elements in the sequence $\vartheta$. $S$ denotes the value space of the sequence $\vartheta$, and $p(\tau_i) = n(\tau_i)/n$, where $\tau_i \in S$, is the probability of the type $\tau_i$ in the sequence $\vartheta$.

Meanwhile, we ensure the generalizability of the selected concept by optimizing our goal to cover as many instances as possible. However, this can result in overly abstract concepts, as extremely coarse-grained concepts like "entity" can encompass most argument words.

To address this issue, as depicted in Figure 2-Step3, we introduce a Hierarchical-Depth regularization term to constrain the model's selection. We define the hierarchical depth of $t_i$ as follows:

$$\text{ds}(t_i) = \frac{\text{idx}(t_i, C)}{\text{len}(C)}$$

**Algorithm 1** GA for ECS

**Input:** Verb $v$ , semantic types set $T_i$ for each argument words $w_i \in W$ specific to role $a$, the population size $S$, the number of iterations $G_{max}$

**Output:** The optimal type sequence $\vartheta_o$.

**while** current population size $p_{cur} < S$ **do**
  **for** each $w_i \in W$ **do**
    randomly select $t_i$ from $T_i$;
  **end for**
  set the initial solutions as $\vartheta = \{t_1, \ldots, t_n\}$
**end while**
**while** current generation $G < G_{max}$ **do**
  Calculate fitness $\mathbb{L}(\vartheta)$ of each individual;
  Retain several individual with higher fitness;
  Reproduce to generate new individuals;
  Integrate the population to $S$;
  $G = G + 1$;
**end while**
Set $\vartheta_o$ as the individual with the highest fitness level;
return $\vartheta_o$

where $\text{idx}(t_i, C)$ represents the depth of $t_i$ in its concept hierarchy $C$ (i.e., the index of $t_i$ in $C$), and $\text{len}(C)$ denotes the length of concept hierarchy $C$.

Next, the hierarchical depth score as a regularization term is integrated into the objective function:

$$\mathbb{L}(\vartheta) = H(\chi|v, W) + \|\text{ds}(t)\|_2$$

This regularization term effectively constrains the model, encouraging the selection of finer-grained concepts. Finer-grained concepts are more adept at distinguishing arguments with varied meanings, particularly when the target verb is polysemous. Moreover, these refined concepts enhance the accuracy of our search for entailment relationships. However, as the number of arguments $n$ increases, the regularization term grows exponentially (the proof process is documented in the appendix, see Section A.4), leading to an imbalance among the terms of $\mathbb{L}$. To address this, we add a coefficient to the regularization term and introduce a hyperparameter $\lambda$ between the two terms. This allows us to balance the two objectives and control the granularity of concept selection.

Finally, we employ the genetic algorithm (Algorithm 1) as a heuristic to find the optimal solution.

$$\mathbb{L}(\vartheta) = \lambda H(\chi|v, W) + (1-\lambda)\frac{1}{\sqrt{n}}\|\text{ds}(t)\|_2$$

### 3.4 Learning Entailment Graphs (Step 4)

For an event in the corpus, we denote it as $E^u = (v, A^u)$, where $v$ represents the predicate in the event, and $A^u = \{(r_1, w_1), ..., (r_n, w_n)\}$ represents all the arguments of event $E^u$, with $w_i$ being the argument word with the role $r_i$ in $E^u$. Additionally, we define: 1. $E^h = (v, A^h), A^h = \{(r_1, \rho_1), ..., (r_n, \rho_n)\}$, where $\rho_1 = HC(w_i)$ denotes the hierarchical conceptualization result of the core word $w_i$. 2. $E^c = (v, A^c), A^c = \{(r_1, t_1), ..., (r_n, t_n)\}$, where $t_i$ represents the type determined after hierarchical concept selection for the core word $w_i$. We limit each verb to a maximum of three arguments.

Given a set of conceptualized argument constraints $A^c$, we filter the event set $\mathbf{E}$ from the corpus, where $E^u \in \mathbf{E}$ must meet the following criteria:

1. The number of roles in the event $E^u$ should be less than or equal to the number of roles in the constraint $A^c$.

2. For each role $r_i$ and its argument $w_i$ of the event $E^u$, we require that $t_i \in \rho_i$, where $t_i$ represents the type of the role corresponding to the given constraint $A^c$, and $\rho_i$ denotes the hierarchical conceptualization result of the argument $w_i$.

Subsequently, we adopt (McKenna et al., 2021; Hosseini et al., 2018) to construct an entailment graph, with typed predicates $A$ as nodes and entailment relationships as edges based on the multi-valued distribution containment hypothesis. To maintain data integrity, we only mark the edges with a BInc score (Weeds and Weir, 2003) exceeding 0.9 as entailment relationships.

Moreover, according to criteria 2, due to the existence of hierarchical conceptualization, an event $E^u$ in the corpus may simultaneously satisfy the conditions of multiple argument type constraints. As shown in Figure 2-Step4, in the event *Grandmother give Leo toast*., the term *toast* has entailment relationships across different hierarchical conceptualizations. We connect these relationships to construct noun entailment connections. (Wang et al., 2024c).

## 4 Experiment

Due to lack of Entailment graph datasets pertinent to problem, we construct data that conforms to hierarchical concept entailment based on existing datasets (Section 4.1) and verify the effectiveness of our method in Section 4.3. Furthermore, to assess the efficacy of our open information extraction and hierarchical concept selection, we conduct

| Methods | Backbone | Noun | | | Verb | | | ABS-HC | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Acc | Ma-F1 | AUC | Acc | Ma-F1 | AUC | Acc | Ma-F1 | AUC |
| NLI + Zero | BART-large-mnli | 71.24 | 68.13 | 75.67 | 56.25 | 47.17 | 62.33 | 65.68 | 72.17 | 72.52 |
| | RoBERTa-large-mnli | 68.66 | 63.18 | 75.42 | 55.73 | 45.54 | 61.27 | 64.62 | 72.30 | 72.68 |
| | DeBERTa-base-mnli | 68.77 | 65.81 | 72.79 | 56.42 | 48.08 | 61.55 | 64.96 | 71.00 | 69.98 |
| | DeBERTa-large-mnli | 73.18 | 71.08 | 78.12 | 56.93 | 49.28 | 63.16 | 68.38 | 73.42 | 73.09 |
| NLI + FT | BART-large-mnli | 85.75 | 85.12 | 90.80 | 64.96 | 64.96 | 68.60 | 79.52 | 80.52 | 87.15 |
| | RoBERTa-large-mnli | 86.15 | 85.34 | 90.87 | 64.61 | 64.26 | 69.46 | 79.13 | 80.46 | 86.96 |
| | DeBERTa-base-mnli | 85.59 | 84.61 | 90.43 | 65.50 | 65.47 | 69.87 | 77.10 | 78.89 | 85.73 |
| | DeBERTa-large-mnli | 86.62 | 85.83 | 91.00 | 66.04 | 65.96 | 70.51 | 79.83 | 80.80 | 87.51 |
| PLM + FT | RoBERTa-base | 84.23 | 83.25 | 89.58 | 63.55 | 63.53 | 68.12 | 79.13 | 80.19 | 86.69 |
| | RoBERTa-large | 85.27 | 84.44 | 90.59 | 64.98 | 64.98 | 69.23 | 79.65 | 80.82 | 87.34 |
| | DeBERTa-base | 84.09 | 83.03 | 89.74 | 63.50 | 63.45 | 68.03 | 78.85 | 79.95 | 86.78 |
| | DeBERTa-large | 86.89 | 86.11 | 90.98 | 65.54 | 65.52 | 69.11 | 80.32 | 81.17 | 87.76 |
| LLM+LoRA | Falcon (7B) | 87.06 | 86.36 | 91.42 | 63.92 | 63.79 | 68.06 | 77.50 | 79.04 | 85.94 |
| | Falcon-Ins (7B) | 86.04 | 85.43 | 91.10 | 64.00 | 63.96 | 68.53 | 76.64 | 78.41 | 85.27 |
| | Llama2 (7B) | 87.56 | 86.82 | 91.52 | 65.07 | 64.79 | 69.27 | 79.20 | 80.52 | 87.28 |
| | Llama2-Chat (7B) | 86.71 | 86.17 | **91.79** | 64.96 | 64.54 | 68.95 | 79.41 | 80.78 | 87.51 |
| | Llama3-Ins (8B) | 87.34 | 89.91 | 91.47 | 64.51 | 64.61 | 69.11 | 78.23 | 79.81 | 86.82 |
| LLM API | ChatGPT | 74.00 | 72.27 | - | 56.30 | 55.71 | - | 68.13 | 68.32 | - |
| | ChatGPT (CoT) | 62.90 | 62.88 | - | 56.20 | 53.89 | - | 60.11 | 61.29 | - |
| | ChatGPT (10-shot ICL) | 76.10 | 74.60 | - | 58.60 | 58.51 | - | 70.46 | 70.39 | - |
| | GPT-4 | 80.50 | 78.70 | - | 56.30 | 53.84 | - | 65.30 | 70.21 | - |
| | GPT-4o | 78.10 | 83.32 | - | 58.00 | 66.56 | - | 66.40 | 72.94 | - |
| HiCon-EG | DeBERTa-large-mnli | 87.46 | 89.55 | 91.37 | **66.73** | 67.22 | **70.90** | 81.52 | **82.87** | 89.35 |
| | DeBERTa-base | 87.30 | 89.77 | 91.56 | 65.36 | **67.90** | 69.40 | 81.60 | 82.70 | **89.62** |
| | DeBERTa-large | **87.60** | **89.98** | 91.60 | 65.77 | 66.76 | 70.13 | **81.88** | 82.79 | 89.59 |

Table 1: Main results on ABSPYRAMID dataset. We evaluate the model performance across noun, verb, and HC datasets of ABSPYRAMID using Acc, Ma-F1, and AUC. Bold highlights the best performance, while underlining indicates the second-best.

| Type | # Total | # Train | # Valid | # Test |
|---|---|---|---|---|
| Noun | 100783 | 81,034 | 9,874 | 9,875 |
| Verb | 61542 | 49,669 | 5,939 | 5,934 |
| ABS-HC | 157948 | 94,753 | 31,584 | 31,611 |

Table 2: Some statistical results of the ABSPYRAMID, where Noun entailment and Verb entailment are consistent with the original dataset, and ABS-HC dataset is the dataset we re-divided.

verification experiments in Sections 4.4 and 4.5, respectively.

### 4.1 Dataset Construction

First, we develop a dataset to validate HiCon-EG. ABSPYRAMID (Wang et al., 2024c) consolidates a comprehensive entailment graph dataset comprising fundamental events from ASER (Zhang et al., 2022a) and abstract concepts curated with guidance from WordNet (Miller, 1995) and ChatGPT. We extract verb and noun entailment data from this dataset, filtering out entries with inconsistent entailment relationships across different conceptualization levels (Appendix B.1). Subsequently, we partition the ABSPYRAMID dataset, denoting the resulting subsets as ABSPYRAMID-HC (ABS-HC). Table 2 illustrates the statistical breakdown of this partition.

### 4.2 Baselines

We fine-tunes some models with HiCon-EG and then compare the results with the following baselines: 1.NLI model + Zero Shot, 2.NLI model + FT, 3.PLM + FT, 4.LLM + LoRA, 5.LLM API. Considering these methods are fine-tuned on the complete ABSPYRAMID dataset, we do not compare the sampling instruction-tuning method of AbsInstruct as a baseline.

Overall, we follow the experimental details in ABSPYRAMID, hyperparameters in fine-tuning and LoRA, prompts in LLMs. to ensure consistency with our Baseline.

### 4.3 Abstraction Detection task

We establish an Abstraction Detection task, where the model discerns whether an Abstraction relationship exists between given premise $A_1$ and hypothesis $A_2$. Model performance is assessed based on three evaluation metrics: accuracy, F1 score, and AUC value.

**Main Results**: We conduct experiments on the three entailment relationship datasets of ABSPYRAMID (Noun, Verb, ABS-HC), with results presented in Table 1.

We observed that HiCon-EG enhances the PLMs' overall abstraction capabilities to a certain extent. This is attributable to the following two aspects: HiCon-EG ,on one hand, can effectively mine richer verb entailment relationships with different abstract levels of noun concepts, thereby improving the model's verb abstraction capabilities; on the other hand, the rich entailment relationships between verbs can be conducive to the model fully mining hierarchical noun concepts, thus upgrading the model's noun abstraction capabilities. The mutual promotion of the two types of relationships in HiCon-EG is well illustrated by the model's notably improved performance on the ABS-HC dataset.

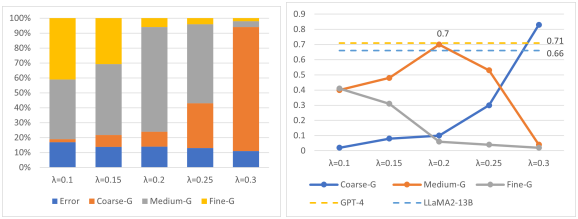
Even on the ABSPYRAMID-Noun dataset, where existing models have shown strong performance, HiCon-EG still demonstrates notable improvements, particularly in F1 scores. We attribute this enhancement to our dataset's ability to effectively address sample imbalances and aid the model in identifying incorrect entailment relationships.

**NLI models Ability**:Moreover, NLI demonstrates a certain zero-shot capability on the ABS-HC dataset, with DeBERTa-large-mnli achieving an F1 score of 73.42 (He et al., 2021). This suggests that NLI, due to its pre-training task similarity, has acquired knowledge, particularly regarding noun entailment, pertinent to our task. However, we also note that fine-tuning the NLI model on our dataset yields performance comparable to PLM+FT. This indicates the distinctiveness and necessity of our task relative to NLI.

**LLM models Ability**: We fine-tuned LLMs with LoRA (Hu et al., 2022) to assess their performance on the ABS-HC dataset, including LLaMA (Touvron et al., 2023), Falcon (Penedo et al., 2023), etc. While LLMs generally exhibit strong performance, they do not surpass the HiCon-EG method. This might stem from the fact that LLMs does not specifically learn diverse entailment relationships under different hierarchical concepts during the pre-training phase. Similarly, we supplemented the performance of ChatGPT on the ABS-HC dataset and obtained similar conclusions.

| Models | LSOIE-wiki | LSOIE-sci |
|---|---|---|
| BERT | 47.5 | 57 |
| BERT+Dep-GCN | 48.7 | 58.1 |
| SMiLe-OIE | 51.7 | 60.5 |
| Chunk-OIE | 52.8 | **61.5** |
| CRF | 52.57 | 58.49 |
| +C2S | **53.92** | 59.86 |

Table 3: The performance of our C2S method on the LSOIE-wiki and LSOIE-sci datasets. We evaluate the performance of all models using the F1 value. Our method outperforms current state-of-the-art (SOTA) models on the LSOIE-wiki dataset, particularly notable for its longer average sentence length.




(a) The proportion of different granularity concepts in the concept selection results

(b) The proportion of moderate granularity concepts in the concept selection results as the parameter $\lambda$ changes

Figure 4: the human evaluation results of hierarchical concept selection

### 4.4 Open Information Extraction (OIE)

To validate the efficacy of the OIE method proposed in this paper, we conducted experiments on the LSOIE datasets (Solawetz and Larson, 2021), with results presented in Table 3. Compared to existing methods (Dong et al., 2023, 2022), our approach yielded superior performance in the open information extraction task. Particularly on the LSOIE-wiki dataset, characterized by longer average sentence length, our method outperforms current SOTA models. Simultaneously, we performed ablation studies on the C2S process, revealing its significant contribution to the OIE task.

| | senses | depth | total |
|---|---|---|---|
| original (WordNet) | 9.14 | 5.56 | 23.4 |
| selected(HiCon-EG) | 1.72 | 1.98 | 3.56 |

Table 4: Comparison of the average values before and after concept selection. *Sense* represents the number of senses of polysemous nouns, *depth* indicates the average depth of concepts of all senses, and *total* shows the average number of concepts.

### 4.5 Hierarchical Concept Selection

Assessing the effectiveness of the Entropy-based Concept Selection method is pivotal to our research. In this section, we define the task of evaluating the semantic granularity of concepts as follows:

Annotators are tasked with assessing the correctness and semantic granularity of a conceptualization result $C$ for a given triple (Verb, argument, concept), consisting of a verb $V$, an argument $W$, and $C$. Evaluation labels encompass: correct, too abstract, too specific, and moderate.

We enlisted three master's students as annotators and randomly sampled 500 conceptualization results from our dataset. Detailed numerical information regarding the evaluation results is documented in Appendix B.2. To ensure annotation accuracy, we assessed inter-annotator consistency, yielding a Fleiss' Kappa result of 0.80.

The results depicted in Figure 4(b) demonstrate the efficacy of the parameter $\lambda$ in regulating semantic granularity. At a value of 0.2, our method achieves a moderate granularity selection rate comparable to GPT-4, while maintaining lower cost and higher efficiency compared to LLM.

In addition, we evaluated the filtering effect of the Entropy-based Concept Selection method on WordNet which has a large number of hierarchical concepts, considering that concept selection can effectively reduce the complexity of our subsequent calculations.We compared the number of synsets, the average depth of hierarchical concepts, and the average number of concepts before and after concept selection. As shown in Table 4, the Entropy-based Concept Selection method greatly reduce the number of concepts, and this enables our efficient calculations even with a large number of concepts.

### 4.6 Commonsense Reasoning

Since HiCon-EG constructs entailment relationships through Distributional Inclusion Hypothesis, it can not only discover abstract relationships between concepts at different levels, such as *PersonX present PersonY Entity* $\models$ *PersonX give PersonY Entity*, but also explore relations that are akin to commonsense knowledge, such as *PersonX give PersonY Entity* $\models$ *PersonY receive Entity*. To evaluate the impact of HiCon-EG on conceptualized commonsense reasoning tasks, we conduct experiments using the AbstractATOMIC dataset (He et al., 2024) and CAT as baseline (Wang et al., 2023) . Comparisons with SOTA models using AUC and ACC metrics show that HiCon-EG slightly outperforms existing methods, as indicated in Table 5. More details was shown in appendix B.3

| Models | Validation | | Testing | |
|---|---|---|---|---|
| | AUC | ACC | AUC | ACC |
| RoBERTa-large | 75.3 | 81.77 | 76.9 | 82.69 |
| DeBERTa-large | 76.9 | 82.18 | 78 | 82.96 |
| CAT | 78.7 | 82.88 | 80 | 83.52 |
| CANDLE | - | 83.64 | - | **84.64** |
| VERA-T5+FT | - | 80.13 | - | 81.25 |
| LLAMA2+LoRA | - | 79.89 | - | 82.15 |
| HiCon-EG | **78.95** | **83.94** | **80.15** | 84.53 |

Table 5: The performance of HiCon-EG on the AbstractATOMIC dataset: Comparative Analysis with State-of-the-Art Models. We assessed all models' performance using AUC and ACC metrics.

## 5 Conclusion

In this paper, we propose a method for constructing a Hierarchical Conceptual Entailment Graph. This approach aids the model in identifying entailment relationships across diverse hierarchical concepts, thereby enhancing the abstract reasoning capabilities of existing models. We validate the value of our method across Conceptualized Commonsense Reasoning and abstraction detection tasks, demonstrating the effectiveness of both the Complex-to-Simple Open Information Extraction (C2S OIE) method and the Entropy-based concept selection method proposed in this paper. The experimental results show that the entailment relationships under different levels of concepts in HiCon-EG can effectively help language models improve their understanding of concepts, thereby enhancing language models' performance on commonsense reasoning tasks.

## Limitations

The method of this paper is based on open information extraction of the corpus, and constructs a hierarchical concept entailment graph through hierarchical conceptualization and multi-valued distribution containment hypothesis. However, compared with knowledge bases such as ASER, the entailment graph we constructed has a single relation, and more abundant relations can be added in the future.

Our method has achieved good results in entailment reasoning and abstract commonsense reasoning. However, such data are all abstract-level datasets. In the future, we will try to use this method to verify on more instance-level datasets to examine whether abstract reasoning ability can be extended to factual reasoning tasks, or to enhance the model's abstract reasoning ability through factual reasoning.

In addition, although our method effectively improves the model's abstract reasoning ability, our method is still an unsupervised construction method based on the corpus, and the entailment relationships generated in this way cannot guarantee their accuracy. In the future, we hope to introduce more supervised information and evaluation methods to ensure the accuracy of the extracted abstract reasoning relationships.

## Acknowledgments

This work has been supported by the National Natural Science Foundation of China (No.61936012), by the Science and Technology Cooperation and Exchange Special Project of ShanXi Province (No.202204041101016), by the Chang Jiang Scholars Program (J2019032), and by the Key Research and Development Program of Shanxi Province (No.202102020101008).

## References

Jonathan Berant, Ido Dagan, and Jacob Goldberger. 2010. Global learning of focused entailment graphs. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1220–1229, Uppsala, Sweden. Association for Computational Linguistics.

Jonathan Berant, Ido Dagan, and Jacob Goldberger. 2011. Global learning of typed entailment rules. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 610–619, Portland, Oregon, USA. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Zhibin Chen, Yansong Feng, and Dongyan Zhao. 2022a. Entailment graph learning with textual entailment and soft transitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5899–5910, Dublin, Ireland. Association for Computational Linguistics.

Zhuo Chen, Yufeng Huang, Jiaoyan Chen, Yuxia Geng, Yin Fang, Jeff Z. Pan, Ningyu Zhang, and Wen Zhang. 2022b. Lako: Knowledge-driven visual question answering via late knowledge-to-text injection. In *IJCKG*, pages 20–29.

Thomas M Cover. 1999. *Elements of information theory*. John Wiley & Sons.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The pascal recognising textual entailment challenge. In *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment*, pages 177–190, Berlin, Heidelberg. Springer Berlin Heidelberg.

Ido Dagan, Fernando Pereira, and Lillian Lee. 1994. Similarity-based estimation of word cooccurrence probabilities. In *32nd Annual Meeting of the Association for Computational Linguistics*, pages 272–278, Las Cruces, New Mexico, USA. Association for Computational Linguistics.

Kuicai Dong, Aixin Sun, Jung-Jae Kim, and Xiaoli Li. 2022. Syntactic multi-view learning for open information extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4072–4083, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Kuicai Dong, Aixin Sun, Jung-jae Kim, and Xiaoli Li. 2023. Open information extraction via chunks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15390–15404, Singapore. Association for Computational Linguistics.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith.

2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Jie He, Simon Chi Lok U, Víctor Gutiérrez-Basulto, and Jeff Z. Pan. 2023. BUCA: A Binary Classification Approach to Unsupervised Commonsense Question Answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL 2023)*.

Mutian He, Tianqing Fang, Weiqi Wang, and Yangqiu Song. 2024. Acquiring and modeling abstract commonsense knowledge via conceptualization. *Artificial Intelligence*, 333:104149.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention.

Xavier Holt. 2019. Probabilistic models of relational implication.

Mohammad Javad Hosseini, Nathanael Chambers, Siva Reddy, Xavier R. Holt, Shay B. Cohen, Mark Johnson, and Mark Steedman. 2018. Learning typed entailment graphs with global soft constraints. *Transactions of the Association for Computational Linguistics*, 6:703–717.

Mohammad Javad Hosseini, Shay B. Cohen, Mark Johnson, and Mark Steedman. 2019. Duality of link prediction and entailment graph induction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4736–4746, Florence, Italy. Association for Computational Linguistics.

Mohammad Javad Hosseini, Shay B. Cohen, Mark Johnson, and Mark Steedman. 2021. Open-domain contextual link prediction and its complementarity with entailment graphs. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2790–2802, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12).

Dimitri Kartsaklis and Mehrnoosh Sadrzadeh. 2016. Distributional inclusion hypothesis for tensor-based composition. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2849–2860, Osaka, Japan. The COLING 2016 Organizing Committee.

Mike Lewis and Mark Steedman. 2013. Combined Distributional and Logical Semantics. *Transactions of the Association for Computational Linguistics*, 1:179–192.

Xiao Ling and Daniel Weld. 2021. Fine-grained entity recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 26(1):94–100.

H. Liu and P. Singh. 2004. Conceptnet — a practical commonsense reasoning tool-kit. *BT Technology Journal*, 22(4):211–226.

Jingping Liu, Tao Chen, Chao Wang, Jiaqing Liang, Lihan Chen, Yanghua Xiao, Yunwen Chen, and Ke Jin. 2022. Vocsk: Verb-oriented commonsense knowledge mining with taxonomy-guided induction. *Artificial Intelligence*, 310:103744.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

Nick McKenna, Liane Guillou, Mohammad Javad Hosseini, Sander Bijl de Vroe, Mark Johnson, and Mark Steedman. 2021. Multivalent entailment graphs for question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10758–10768, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Nick McKenna, Tianyi Li, Mark Johnson, and Mark Steedman. 2023. Smoothing entailment graphs with language models. 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 (Volume 1: Long Papers)*, pages 551–563, Nusa Dua, Bali. Association for Computational Linguistics.

Chris Mellish and Jeff Z. Pan. 2008. Natural Language Directed Inference from Ontologie". In *Artificial Intelligence Journal*.

George A. Miller. 1995. Wordnet: a lexical database for english. *Commun. ACM*, 38(11):39–41.

Marvin Minsky. 1980. K-lines: A theory of memory. *Cognitive Science*, 4(2):117–133.

Sebastian Padó, Daniel Matthew Cer, Michel Galley, Dan Jurafsky, and Christopher D. Manning. 2009. Measuring machine translation quality as semantic equivalence: A metric based on entailment features. *Machine Translation*, 23:181–193.

Jeff Z. Pan and Ian Horrocks. 2002. Reasoning in the SHOQ(Dn) Description Logic. In *Proc. of the 2002 Int. Workshop on Description Logics (DL-2002)*.

Jeff Z. Pan, Simon Razniewski, Jan-Christoph Kalo, Sneha Singhania, Jiaoyan Chen, Stefan Dietze, Hajira Jabeen, Janna Omeliyanenko, Wen Zhang, Matteo Lissandrini, ussa Biswas, Gerard de Melo, Angela Bonifati, Edlira Vakaj, Mauro Dragoni, and amien Graux. 2023. Large language models and knowledge graphs: Opportunities and challenges. *Transactions on Graph Data and Knowledge*.

J.Z. Pan, D. Calvanese, T. Eiter, I. Horrocks, M. Kifer, F. Lin, and Y. Zhao, editors. 2017a. *Reasoning Web: Logical Foundation of Knowledge Graph Construction and Querying Answering*. Springer.

J.Z. Pan, G. Vetere, J.M. Gomez-Perez, and H. Wu, editors. 2017b. *Exploiting Linked Data and Knowledge Graphs for Large Organisations*. Springer.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The refinedweb dataset for falcon LLM: outperforming curated corpora with web data, and web data only. *CoRR*, abs/2306.01116.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. In *Joint Conference on EMNLP and CoNLL - Shared Task*, pages 1–40, Jeju Island, Korea. Association for Computational Linguistics.

Julien Romero, Simon Razniewski, Koninika Pal, Jeff Z. Pan, Archit Sakhadeo, and Gerhard Weikum. 2019. Commonsense Properties from Query Logs and Question Answering Forums. In *Proc. of 28th ACM International Conference on Information and Knowledge Management (CIKM 2019)*, pages 1411–1420.

Ameer Saadat-Yazdi, Jeff Z. Pan, and Nadin Kökciyan. 2023. Uncovering Implicit Inferences for Improved Relational Argument Mining. In *EACL*, pages 2476–2487.

Lorenza Saitta and Jean-Daniel Zucker. 2013. *Abstraction in Different Disciplines*, pages 11–47. Springer New York, New York, NY.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. 2019. Atomic: an atlas of machine commonsense for if-then reasoning. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19. AAAI Press.

Jacob Solawetz and Stefan Larson. 2021. LSOIE: A large-scale dataset for supervised open information extraction. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2595–2600, Online. Association for Computational Linguistics.

Mark Steedman. 2001. *The Syntactic Process*. The MIT Press.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, and Amjad Almahairi. 2023. Llama 2: Open foundation and fine-tuned chat models.

Weiqi Wang, Tianqing Fang, Chunyang Li, Haochen Shi, Wenxuan Ding, Baixuan Xu, Zhaowei Wang, Jiaxin Bai, Xin Liu, Jiayang Cheng, Chunkit Chan, and Yangqiu Song. 2024a. Candle: Iterative conceptualization and instantiation distillation from large language models for commonsense reasoning.

Weiqi Wang, Tianqing Fang, Baixuan Xu, Chun Yi Louis Bo, Yangqiu Song, and Lei Chen. 2023. Cat: A contextualized conceptualization and instantiation framework for commonsense reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, ACL 2023*.

Zhaowei Wang, Wei Fan, Qing Zong, Hongming Zhang, Sehyun Choi, Tianqing Fang, Xin Liu, Yangqiu Song, Ginny Wong, and Simon See. 2024b. AbsInstruct: Eliciting abstraction ability from LLMs through explanation tuning with plausibility estimation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 973–994, Bangkok, Thailand. Association for Computational Linguistics.

Zhaowei Wang, Haochen Shi, Weiqi Wang, Tianqing Fang, Hongming Zhang, Sehyun Choi, Xin Liu, and Yangqiu Song. 2024c. AbsPyramid: Benchmarking the abstraction ability of language models with a unified entailment graph. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3991–4010, Mexico City, Mexico. Association for Computational Linguistics.

Julie Weeds and David Weir. 2003. A general framework for distributional similarity. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, pages 81–88.

Wentao Wu, Hongsong Li, Haixun Wang, and Kenny Q. Zhu. 2012. Probase: a probabilistic taxonomy for text understanding. *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*.

Yuting Wu, Yutong Hu, Yansong Feng, Tianyi Li, Mark Steedman, and Dongyan Zhao. 2023. Align-then-enhance: Multilingual entailment graph enhancement

with soft predicate alignment. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 881–894, Toronto, Canada. Association for Computational Linguistics.

Xintong Yu, Hongming Zhang, Yangqiu Song, Changshui Zhang, Kun Xu, and Dong Yu. 2021. Exophoric pronoun resolution in dialogues with topic regularization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3832–3845, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Congle Zhang and Daniel S. Weld. 2013. Harvesting parallel news streams to generate paraphrases of event relations. In *Conference on Empirical Methods in Natural Language Processing*.

Hongming Zhang, Xin Liu, Haojie Pan, Haowen Ke, Jiefu Ou, Tianqing Fang, and Yangqiu Song. 2022a. ASER: towards large-scale commonsense knowledge acquisition via higher-order selectional preference over eventualities. *Artificial Intelligence*, 309:103740.

Yu Zhang, Qingrong Xia, Shilin Zhou, Yong Jiang, Guohong Fu, and Min Zhang. 2022b. Semantic role labeling as dependency parsing: Exploring latent tree structures inside arguments. In *Proceedings of COLING*, pages 4212–4227, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Huaixiu Steven Zheng, Swaroop Mishra, Xinyun Chen, Heng-Tze Cheng, Ed H. Chi, Quoc V Le, and Denny Zhou. 2024. Take a step back: Evoking reasoning via abstraction in large language models.

Wendi Zhou, Tianyi Li, Pavlos Vougiouklis, Mark Steedman, and Jeff Z Pan. 2024. A Usage-centric Take on Intent Understanding in E-Commerce. In *Proc. of Empirical Methods in Natural Language Processing (EMNLP 2024)*.

# A Hierarchical Concept Entailment Graph Details

## A.1 Prompts for C2S

The input of our C2S process is a complex sentence, and we require the model to decompose it into multiple simple sentences. In the prompts, the prompt we give is shown in Table 6.

| |
|---|
| **Task Instruction**: Given a long sentence, parse all events in it and generate corresponding simple sentences. Here are some examples |
| **Example Input**: With time winding down , Avs defenseman Greg Zanon tried to clear the puck from behind his net , but it hit a referee 's stake in the corner and bounced to Kyle Chipchura . |
| **Example Output**: 1. Time was winding down. 2. Greg Zanon tried to clear the puck from behind his net 3. the puck hit a referee's stake in the corner. 4. the puck bounced to Kyle Chipchura |
| **More Examples**: ... |
| **Query Input**: Now, extract the events in the following sentences according to the format of the above example: **[Sentence]** |

Table 6: The C2S process prompt. The placeholder **[Sentence]** will be replaced with real sentence.

## A.2 Distillation process

In order to reduce the computational cost and obtain more stable results, we distilled the ability of generate simple sentences from Llama2 into Bert. First, we filtered the data generated by it according to the following strategy:

1. The length of the generated clause must be greater than 5, so that short phrases generated by large models can be effectively filtered out.

2. The generated sentence must contain a verb.

3. Each word in the generated sentence must appear in the original sentence.

4. In the generated sentence, the verb must have at least one argument.

After filtering out higher-quality clauses, we denote the clause as $S_c$ and the original sentence as $S_o$, and we construct the dataset according to the following steps:

1. We follow the method of A.3 to retrieve the core verb in the clause, denoted as $v$.

2. For each word in the clause, we retrieve its position in the original sentence and mark it as 1. If a word appears multiple times, we choose the one closest to the verb $v$ in the original sentence.

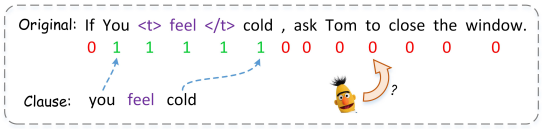



Figure 5: An example of the sentence simplification dataset we constructed, where the model is required to mark the arguments related to the verb in the sentence as 1 and other words as 0.

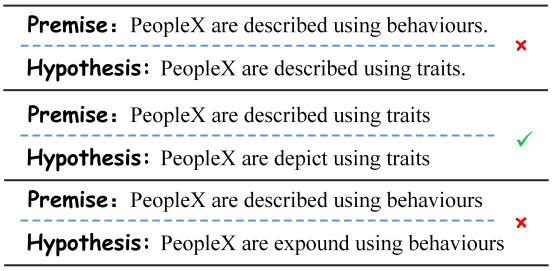



Figure 6: Some examples of the ABSPYRAMID dataset, where the model is required to determine whether the given Premise can infer the Hypothesis.

3. For other words in the original sentence, we mark them as 0.

We then designed a Sequence Labeling task (Figure 5). For a given sentence $S_o$ and the verb $v$ in the sentence, the model needs to mark each argument related to the verb as 1 and other words as 0. Specifically, we used the BERT model to complete this task. In the final application, we first obtained the verb in the original sentence $S_o$ through part-of-speech tagging, and then simplified the sentence through the trained model.

## A.3 Core words retrieved process

In this section, we introduce how to obtain linguistic fragments in the sentence, as shown in Step 2 of Figure 2. We obtain the dependency relationship in the argument through syntactic dependency analysis. Then, for each word $w_i$ in an argument $a = \{w_1, \cdots, w_n\}$, we query its parent node $w_p$ and find $w_t$ that satisfies $w_p \notin a$. At this time, if $w_t$ is not a preposition, we denote $w_t$ as the core word. Otherwise, we query the child node $w_c$ of $w_t$ and mark it as the core word.

## A.4 The proof process of the exponential growth of the regularization term

In this section, we prove the exponential growth property of the regularization term in our hierarchical concept selection.

**Remark 1.** *The regularization term* $\|\vec{a}\|_2 = \mathrm{O}(\sqrt{n})$*, where n is the number of concepts.* $\vec{a}$ *is*

| Model Type | PLM/Method | Validation | | Testing | |
|---|---|---|---|---|---|
| | | AUC | ACC | AUC | ACC |
| Pre-trained Langauge Models | RoBERTa-large 340M | 75.30 | 81.77 | 76.90 | 82.69 |
| | DeBERTa-v3-large 435M | 76.90 | 82.18 | 78.00 | 82.96 |
| | GPT2-XL 1.5B | 62.20 | 47.65 | 61.50 | 47.21 |
| | PseudoReasoner (BERT-base) | 73.00 | 79.69 | 74.10 | 80.27 |
| | PseudoReasoner(RoBERTa-large) | 76.30 | 79.89 | 77.20 | 80.07 |
| | CAT (RoBERTa-large)340M | 78.20 | 82.27 | 79.40 | 83.02 |
| | CAT (DeBERTa-v3-large) 435M | <u>78.70</u> | 82.88 | <u>80.00</u> | 83.52 |
| | CANDLE Distilled (RoBERTa-large) | - | 83.11 | - | 84.50 |
| | CANDLE Distilled (DeBERTa-v3-large) | - | <u>83.64</u> | - | **84.64** |
| Large Langauge Models | ChatGPT (openai/gpt-3.5-turbo) | - | 70.27 | - | 72.08 |
| | LLAMA2 7B | - | 74.67 | - | 76.80 |
| | LLAMA2 13B | - | 80.67 | - | 82.08 |
| | Mistral-v0.1 7B | - | 65.09 | - | 69.80 |
| | LLAMA2 (LoRA Fine-tuned) 7B | - | 79.89 | - | 82.15 |
| | Mistral-v0.1 (LoRA Fine-tuned) 7B | - | 79.59 | - | 80.35 |
| | VERA-T5 5B | - | 72.60 | - | 76.85 |
| | VERA-T5 (Fine-tuned) 5B | - | 80.13 | - | 81.25 |
| Our HiCon-EG | RoBERTa-large 340M | 78.32 | 82.96 | 79.11 | 83.79 |
| | DeBERTa-v3-large 435M | **78.95** | **83.94** | **80.15** | <u>84.53</u> |

Table 7: The performance of our HiCon-EG on the AbstractATOMIC dataset. We compared it with existing methods and mainstream LLMS. We evaluated the performance of all models using AUC and ACC. Our method achieved the best results on most indicators.

*the vector of the concept hierarchical depth score vector and* $a_i \in (0,1]$.

*Proof.* Set $\varepsilon \in (0,1]$ as the minimum value of $a_i$, then we have:

$$\|\vec{a}\|_2 = \sqrt{\sum_{i=1}^{n} a_i^2} \geqslant \sqrt{\sum_{i=1}^{n} \varepsilon^2} = \sqrt{n\varepsilon^2} = \varepsilon\sqrt{n} = \mathrm{O}(\sqrt{n}) \tag{1}$$

similarly, we have:

$$\|\vec{a}\|_2 = \sqrt{\sum_{i=1}^{n} a_i^2} \leqslant \sqrt{\sum_{i=1}^{n} 1^2} = \sqrt{n} = \mathrm{O}(\sqrt{n}) \tag{2}$$

Therefore, by the squeeze theorem, the regularization term $\|\vec{a}\|_2 = \mathrm{O}(\sqrt{n})$. □

| | Coarse-G | Medium-G | Fine-G | Error |
|---|---|---|---|---|
| GPT-4 | 0.03 | **0.71** | 0.10 | 0.15 |
| LLaMA | 0.08 | 0.66 | 0.07 | **0.19** |
| $\lambda$=0.3 | **0.83** | 0.04 | 0.02 | 0.11 |
| $\lambda$=0.25 | 0.30 | 0.53 | 0.04 | 0.13 |
| $\lambda$=0.2 | 0.10 | <u>0.70</u> | 0.06 | 0.14 |
| $\lambda$=0.15 | 0.08 | 0.48 | 0.31 | 0.14 |
| $\lambda$=0.1 | 0.02 | 0.40 | **0.41** | 0.17 |

Table 8: The proportion of concepts of different granularity in the model annotation results under different models/parameters, where Coarse-G represents coarse-grained concepts, Medium-G represents medium-grained concepts, Fine-G represents fine-grained concepts, and Error represents the proportion of incorrect annotations.

## B Experiment Details

### B.1 Dataset Construction

We selected data with different entailment relationships under different hierarchical concepts. The specific screening rules are as follows:

We first selected event pairs $(e_1, e_2)$ with different hierarchical concepts from the noun entailment dataset. Then, we queried the verb entailment relationship sets $E_1$ and $E_2$ of $e_1$ and $e_2$ respectively. We selected the difference set of the two sets $R = (A \setminus B) \cap (B \setminus A) = \{x \mid x \in A \text{ and } x \notin B \vee x \in B \text{ and } x \notin A\}$. Finally, we divided the selected data into the ABSPYRAMID-HC test set. The examples of the data are shown in Figure 6.

### B.2 Hierarchical Concept Selection

To verify the effectiveness of our hierarchical concept selection method, we hired three master's students as annotators. We asked them to evaluate the correctness and semantic granularity of each conceptualized result. Specifically, the annotators need to determine whether each conceptualized result is (Coarse-grained, Medium-grained, Fine-grained, error). We calculated the proportion of each label, and the results are shown in Table 8.

Through the results, we observed that as the parameter $\lambda$ increases, the proportion of fine-grained concepts gradually decreases, and the proportion of coarse-grained concepts gradually increases. When $\lambda = 0.8$, the proportion of moderate granularity concepts is the largest, which indicates that our method is effective in controlling the semantic granularity.

We also tested the effect of LLMs on the concept selection task. Specifically, we selected LLAMA2 7B and GPT-4 for comparison. The results show that GPT-4 achieved better results in selecting moderate granularity concepts, but the error rate of LLMs is relatively high.

### B.3 Commonsense Reasoning

In this task, we use the AbstractATOMIC (He et al., 2024) dataset which is a conceptualized commonsense reasoning dataset built on ATOMIC. We selected the conceptualized data of abstract knowledge triplets in the dataset (as shown in Table 10). In this data, each head event **[Head]** is obtained through instance recognition and conceptualization of the original event **[Sent]** in ATOMIC, and the manual annotation process ensures the reliability of the data.

We conducted experiments on the AbstractATOMIC dataset and compared it with existing work. Since HiCon-EG is a graph of reasoning relationships between events, we only conducted experiments on the "Triple Conceptualization" part of the AbstractATOMIC dataset. The results are shown in Table 5. HiCon-EG achieved better results on all indicators, slightly surpassing the existing methods overall.

We believe that this reflects that HiCon-EG also contains information about commonsense reasoning in the construction process, rather than simply the relationship between the granularity of synonyms.

| | Accuracy | AUC | Macro F1 |
|---|---|---|---|
| bert-base-cased | 85.68 | 87.91 | 66.15 |
| bert-large-cased | 86.68 | 88.92 | 70.11 |
| roberta-base | 84.09 | 87.08 | 60.63 |
| roberta-large | 87.43 | 89.94 | 71.05 |
| deberta-v3-base | 85.72 | 89.95 | 67.43 |
| deberta-v3-large | **89.30** | **93.14** | **75.03** |

Table 9: The results of HiCon-EG on the Levy/Holt dataset. We compared different pre-trained models. We evaluated the performance of all models using Accuracy, AUC, and Macro F1.

| **Sent** | PersonX wins [the costume contest] | |
|---|---|---|
| **Head** | PersonX wins [event] | |
| **relation** | **tail** | **Label** |
| oReact | upset | 1 |
| oWant | congratulate them | 0 |
| xEffect | personx takes home the prize | 1 |
| xIntent | to impress others | 1 |

Table 10: An example in the AbstractATOMIC dataset, where we show the original sentence **[Sent]** in ATOMIC, its conceptualization result as the head node **[Head]**, the relationship **[Relation]**, the tail node **[Tail]**, and the label **[Label]**.

### B.4 Entailment discrimination task

To verify the effectiveness of our method in the traditional entailment graph construction task, we followed the method of Wang et al. (2024c) and fine-tuned the model using the enhanced data of HiCon-EG. We conducted experiments on the Levy/Holt dataset (Gururangan et al., 2018; Holt, 2019) to verify the results. The results are shown in Table 9. Our method achieved good results on the Levy/Holt dataset.

| Methods | Backbone | Noun | | | Verb | | | Merged DataSet | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Acc | Ma-F1 | AUC | Acc | Ma-F1 | AUC | Acc | Ma-F1 | AUC |
| NLI + Zero | BART-large-mnli | 71.24 | 68.13 | 75.67 | 56.25 | 47.17 | 62.33 | 65.68 | 72.17 | 72.52 |
| | RoBERTa-large-mnli | 68.66 | 63.18 | 75.42 | 55.73 | 45.54 | 61.27 | 64.62 | 72.30 | 72.68 |
| | DeBERTa-base-mnli | 68.77 | 65.81 | 72.79 | 56.42 | 48.08 | 61.55 | 64.96 | 71.00 | 69.98 |
| | DeBERTa-large-mnli | 73.18 | 71.08 | 78.12 | 56.93 | 49.28 | 63.16 | 68.38 | 73.42 | 73.09 |
| NLI + FT | BART-large-mnli | 85.75 | 85.12 | 90.80 | 64.96 | 64.96 | 68.60 | 79.52 | 80.52 | 87.15 |
| | +HiCon-EG | 87.04 | 89.47 | 91.27 | 65.99 | 68.33 | 69.75 | 80.43 | 80.91 | 88.76 |
| | RoBERTa-large-mnli | 86.15 | 85.34 | 90.87 | 64.61 | 64.26 | 69.46 | 79.13 | 80.46 | 86.96 |
| | +HiCon-EG | 87.14 | 89.66 | 91.14 | 65.52 | 67.13 | 69.80 | 80.81 | 81.67 | 88.83 |
| | DeBERTa-base-mnli | 85.59 | 84.61 | 90.43 | 65.50 | 65.47 | 69.87 | 77.10 | 78.89 | 85.73 |
| | +HiCon-EG | 85.45 | 88.32 | 90.41 | 66.15 | 67.07 | 70.06 | 80.61 | 81.39 | 88.56 |
| | DeBERTa-large-mnli | 86.62 | 85.83 | 91.00 | 66.04 | 65.96 | 70.51 | 79.83 | 80.8 | 87.51 |
| | +HiCon-EG | 87.46 | 89.55 | 91.37 | **66.73** | 67.22 | **70.90** | 81.52 | **82.87** | 89.35 |
| PLM + FT | BERT-base | 85.09 | 84.14 | 89.94 | 64.26 | 64.20 | 68.06 | 76.73 | 78.58 | 85.39 |
| | +HiCon-EG | 85.78 | 87.72 | 90.02 | 64.13 | 62.83 | 68.53 | 79.52 | 80.66 | 87.81 |
| | BERT-large | 85.94 | 85.12 | 90.37 | 63.58 | 63.58 | 68.03 | 77.28 | 79.29 | 86.06 |
| | +HiCon-EG | 86.77 | 88.42 | 90.73 | 64.89 | 66.54 | 69.73 | 80.67 | 81.56 | 88.67 |
| | RoBERTa-base | 84.23 | 83.25 | 89.58 | 63.55 | 63.53 | 68.12 | 79.13 | 80.19 | 86.69 |
| | +HiCon-EG | 84.07 | 86.97 | 89.54 | 64.09 | 65.83 | 68.34 | 80.96 | 81.71 | 88.90 |
| | RoBERTa-large | 85.27 | 84.44 | 90.59 | 64.98 | 64.98 | 69.23 | 79.65 | 80.82 | 87.34 |
| | +HiCon-EG | 86.52 | 89.06 | 90.82 | 65.17 | 65.62 | 69.53 | 81.36 | 82.46 | 89.19 |
| | DeBERTa-base | 84.09 | 83.03 | 89.74 | 63.50 | 63.45 | 68.03 | 78.85 | 79.95 | 86.78 |
| | +HiCon-EG | 87.30 | 89.77 | 91.56 | 65.36 | **67.90** | 69.40 | 81.60 | 82.70 | **89.62** |
| | DeBERTa-large | 86.89 | 86.11 | 90.98 | 65.54 | 65.52 | 69.11 | 80.32 | 81.17 | 87.76 |
| | +HiCon-EG | **87.60** | **89.98** | **91.60** | 65.77 | 66.76 | 70.13 | **81.88** | 82.79 | 89.59 |

Table 11: The complete experimental results of HiCon-EG on the ABSPYRAMID dataset.