On the Role of Entity and Event Level Conceptualization in Generalizable Reasoning: A Survey of Tasks, Methods, Applications, and Future Directions

Anonymous ACL submission

Abstract

Entity- and event-level conceptualization, as fundamental elements of human cognition, plays a pivotal role in generalizable reasoning. This process involves abstracting specific instances into higher-level concepts and forming abstract knowledge that can be applied in unfamiliar or novel situations, which can enhance models' inferential capabilities and support the effective transfer of knowledge across various domains. Despite its significance, there is currently a lack of a systematic overview that comprehensively examines existing works in the definition, execution, and application of conceptualization to enhance reasoning tasks. In this paper, we address this gap by presenting the first comprehensive survey of 150+ papers, categorizing various definitions, resources, methods, and downstream applications related to conceptualization into a unified taxonomy, with a focus on the entity and event levels. Furthermore, we shed light on potential future directions in this field and hope to garner more attention from the community.

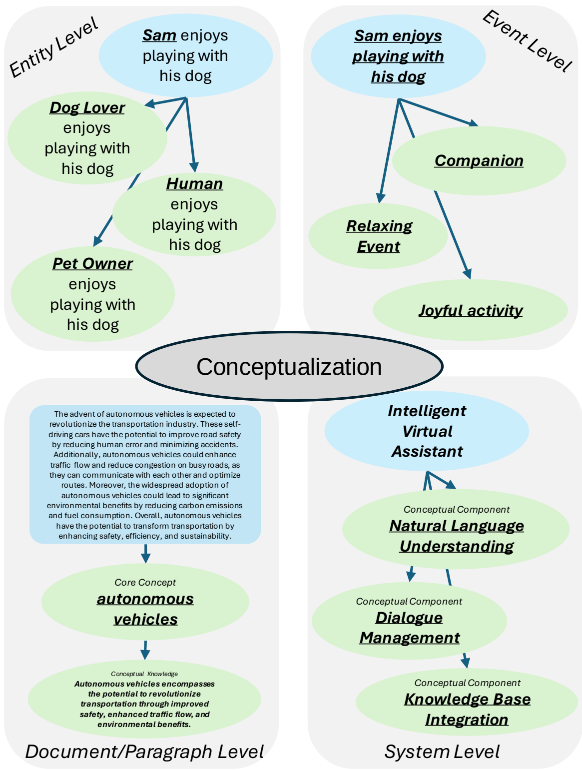

Figure 1: Examples of performing conceptualization at different semantic levels.

1 Introduction

> "*Concepts are the glue that holds our mental world together.*"– Murphy (2004)

Conceptualization has been widely recognized as a fundamental component of human intelligence, spanning fields from psychology (Kahneman, 2011; Evans, 2003; Bransford and Franks, 1971) to computational linguistics (Bengio et al., 2021; Tenenbaum et al., 2011; Lachmy et al., 2022). In the era of deep learning, numerous studies have emerged focusing on conceptualization as a means to achieve generalizable reasoning with (Large) Language Models (LLMs; OpenAI, 2022, 2023; Touvron et al., 2023a,b; Mesnard et al., 2024; Reid et al., 2024) in areas such as commonsense reasoning (Wang et al., 2023b,a, 2024a), causal reasoning (Feder et al., 2021; Kunda et al., 1990), physical reasoning (Bisk et al., 2020; Wang et al., 2023c; Hong et al., 2021), and more.

In general terms, conceptualization refers to the process of consolidating specific instances with shared properties or characteristics into a cohesive concept that represents a vast collection of instances. It is a sub-type of abstraction (Giunchiglia and Walsh, 1992), but specifically requires the presence of a concept as the base for such abstraction. With proper conceptualization, abstract knowledge can be subsequently derived by associating original knowledge at the instance level with that concept. When encountering unfamiliar or novel scenarios, concepts in abstract knowledge can be instantiated to new instances to support downstream reasoning (Tenenbaum et al., 2011). This process can occur at various levels, including entity (Wu et al., 2012; Liang et al., 2017; Alukaev et al., 2023; Liu et al., 2023c), event (He et al., 2024; Wang et al., 2024a,c), paragraph/document (Falke and Gurevych, 2019; Falke et al., 2017), and system levels (Subramonian et al., 2023; Kadioglu and Kleynhans, 2024), ultimately forming a hierarchy that contribute to a comprehensive understanding and representation of knowledge.

Despite its significance, the field lacks a comprehensive and unified taxonomy to categorize existing research on conceptualization. This has led to several drawbacks, such as the inconsistent use of the term "conceptualization" across different studies, resulting in varying definitions despite a common underlying meaning. Additionally, the methods for conceptualizing different types of instances in a scalable and accurate manner remain unclear. Finally, it is essential to summarize the benefits that conceptualization can bring to downstream tasks to gather insights for future applications and new research directions.

To address these issues, we present the first-ever survey that systematically taxonomizes conceptualization. Firstly, in Section 2, we present a hierarchical definition of conceptualization based on different semantic levels of instances being conceptualized, namely: entity, event, paragraph/document, and system. In later sections, we focus on two main types of conceptualization based on the entity and event levels, as they are most prevalent in existing literature and play a key role in human reasoning. We review more than 150 papers and organize them into four main categories, as shown in Figure 2. We summarize the main representative tasks and datasets available for these types of conceptualization in Section 3. Subsequently, in Section 4, we categorize conceptualization acquisition methods into extraction, retrieval, and generative-based methods. The downstream benefits of conceptualization are discussed in Section 5, with a specific focus on several reasoning tasks. Finally, in Section 6, we propose two future directions that can be benefited from conceptualization. We hope our work can serve as a practical handbook for researchers and pave the way for further advancements in conceptualization.

2 The Hierarchy of Conceptualization

We first define the hierarchy of conceptualization according to the type of instances being conceptualized. They are categorized into four levels: entity level, event level, document level, and system level. Running examples are shown in Figure 1.

Entity Level: Entity-level conceptualization involves grouping multiple entities under a shared concept (Yang et al., 2021; Peng et al., 2022). It is the most common form of conceptualization in human cognition and is frequently applied for knowledge acquisition (Carey, 1991; Murphy, 2004). For instance, entities like "apple," "pear," and "grape," can be categorized together under the broader concept of "fruit." By doing so, abstract knowledge can be derived by reintegrating the concept into the context of specific instances, such as the assertion "fruit is delicious," with "apple is delicious" serving as the specific source. When someone encounters an unknown fruit, they can quickly understand its properties by associating it with the abstract knowledge of fruit, such as its possible taste or nutrition.

Event Level: While a concept can capture the semantic meaning of a group of entities, it can also represent events at a higher level of conceptualization. Event-level conceptualization aims to broaden the scope from entities to include events as well (He et al., 2024; Wang et al., 2024c). It seeks to associate different events under a shared concept that preserves the original semantic meaning to the maximum extent possible. For instance, activities like "Sam playing with his dog," "Alex dancing in the club," and "Bob doing yoga" can all be conceptualized as "relaxing events." Abstract knowledge can then follow, stating that "If someone engages in relaxing events, they feel happy and relaxed." When someone encounters an unknown or unfamiliar event, such as "Charlie likes painting the sunset," they can infer that painting the sunset is a relaxing event and that Charlie feels happy and relaxed when doing so.

Document Level: Document-level conceptualization further extends the scope of the instance from entities and events to paragraphs or entire documents. One representative task in this category is abstractive summarization (Ladhak et al., 2022; Wang et al., 2019; Lin and Ng, 2019), where the objective is to generate a summary that captures the main ideas and essential information while maintaining the overall meaning and context of the original text. For example, documents on topics like "A Study on Climate Change," "The Impact of Global Warming," and "The Future of Renewable Energy" can all be abstracted under the concept of

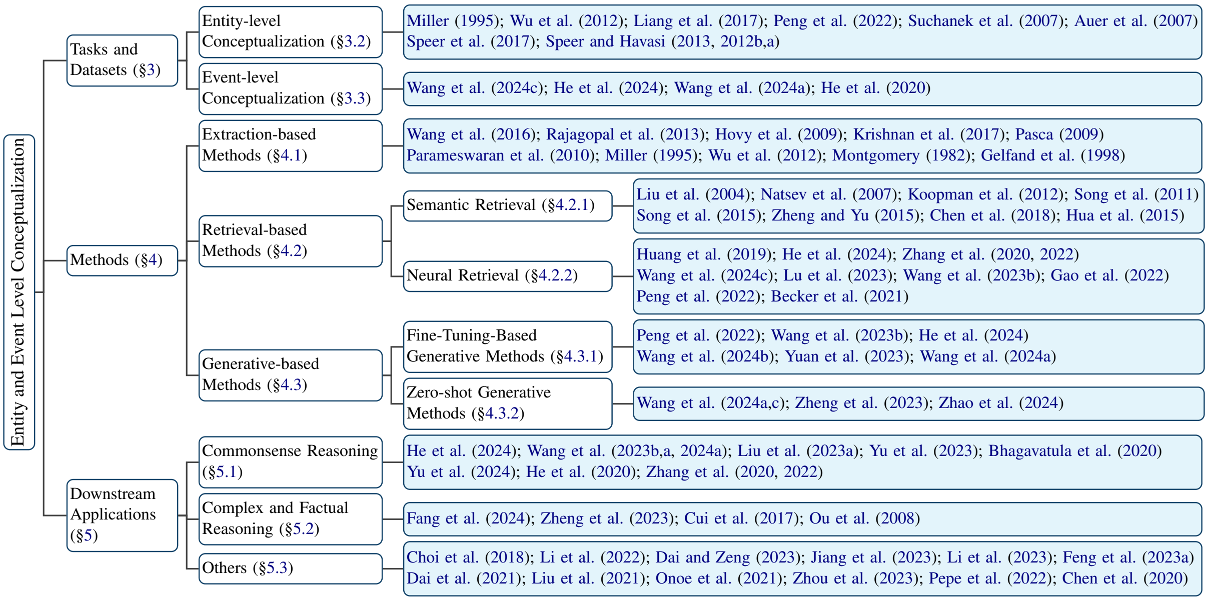

Figure 2: Taxonomy of representative works in entity and event level **conceptualization** categorized by tasks and datasets (§3), methods in performing conceptualization (§4), and downstream applications (§5).

"Environmental Studies." Abstract knowledge can then be derived, such as "Environmental studies discuss the impact of human activities on the environment and potential solutions." This process is challenging because it requires a delicate balance between preserving essential details related to the concept and reducing the length of the text.

System Level: Finally, system-level conceptualization aims to simplify the understanding of a complex system by abstracting its behavior and functionality into a higher-level representation. There is no fixed definition of system-level conceptualization, as it can vary depending on the context and the system being considered. A representative example in NLP is recent work by Subramonian et al. (2023), where the authors provide a systematic categorization of NLP tasks based on their objectives and characteristics while neglecting the detailed format of input/output and the dataset the tasks are evaluated on. Abstract knowledge typically comes in the form of knowledge or facts associated with the derived conceptualization, which may vary depending on the context.

3 Tasks and Datasets

While all levels of conceptualization play a pivotal role in knowledge representation, those at the entity and event levels are the most fundamental due to their unique importance in human cognition and generalizable reasoning. Therefore, in later sections, we specifically focus on entity and event level conceptualizations. We first discuss existing literature on the concept linking task and examine currently available resources for these two types of conceptualizations. Statistical comparisons between different resources are shown in Table 1. For datasets that also serve as evaluation benchmarks, we mark their associated tasks with classification task (`CLS`) and generation task (`GEN`).

3.1 Concept Linking Tasks

The main task of conceptualization can be formulated as a concept linking task, where the goal is to link an instance i to a concept c such that i can be semantically represented by c. It is challenging due to the infinite number of possible instance-concept pairs. Previous approaches, such as those byBrauer et al. (2010); Yates et al. (2015), have attempted to further restrict the task to linking instances to a limited set of strict ontologies using heuristic or statistical methods. The task can also be formulated with a generative objective, which requires a model to generate c directly given i as input.

3.2 Entity-level Conceptualization

To conceptualize different entities into concepts, multiple large-scale concept taxnomies have been constructed as resources for this type of conceptualization. WordNet (Miller, 1995) is the first and most well-known concept taxonomy, which is a large lexical database of English. It is a network of concepts, where each concept is a set of synonyms. Probase (Wu et al., 2012; Liang et al., 2017) is a later built concept taxonomy, which is a large-scale probabilistic taxonomy of concepts.

Type	Dataset	#Instance	#Concept.	Tasks
Entity	WordNet	82,115	84,428	N/A
	Probase	10,378,743	16,285,393	N/A
	Probase+	10,378,743	21,332,357	N/A
	YAGO	143,210	352,297	N/A
	DBPedia	1,000,000	1,000,000	N/A
	ConceptNet	21,000,000	8,000,000	N/A
	COPEN	24,000	393	`CLS`
Event	Abs.ATM.	21,493	503,588	`CLS,GEN`
	Abs.Pyr.	17,000	220,797	`CLS,GEN`
	CANDLE	21,442	6,181,391	N/A

Table 1: Statistical comparisons between different datasets with entity and event level conceptualizations.

It is constructed by analyzing a large amount of web pages and search logs. YAGO (Suchanek et al., 2007) is a semantic knowledge base, which is a large-scale concept taxonomy of entities and events. It is constructed by extracting information from Wikipedia (Merity et al., 2017) and WordNet. DBPedia (Auer et al., 2007) is a large-scale knowledge base which is built by extracting structured information from Wikipedia. It also contains structured conceptual knowledge about entities and events. ConceptNet (Speer et al., 2017) is the most recent concept taxonomy, featuring a large-scale semantic network of concepts. It is constructed by extracting structured information from various sources, including Wikipedia, WordNet, and Open Mind Common Sense (Singh et al., 2002). Recently, Peng et al. (2022) introduced COPEN, a entity level conceptualization benchmark that is constructed by probing language models to retrieve concepts of an entity from a pre-defined set of concepts. All of them are important knowledge bases that are rich in entity conceptualizations.

3.3 Event-level Conceptualization

Compared to abstracting entities, there are fewer resources available for event-level conceptualizations. The most notable is the AbstractATOMIC dataset (He et al., 2024), which was constructed by filtering head events from the ATOMIC dataset and identifying instance candidates within each event using syntactic parsing and human-defined rules. These instances are matched against Probase and WordNet to acquire candidate concepts using GlossBERT (Huang et al., 2019), which are then verified by a supervised model and human annotations. AbsPyramid (Wang et al., 2024c) extends the AbstractATOMIC pipeline to ASER (Zhang et al., 2020, 2022), a large-scale eventuality knowledge graph, by incorporating candidate concepts generated by ChatGPT to complement Probase and WordNet. It also extends coverage to verbs in addition to nouns and events, and broadens the domain of events from social aspects to all aspects. Both datasets provide rich event conceptualizations sourced from diverse origins.

4 Conceptualization Acquisition Methods

In this section, we discuss methods for performing or collecting entity and event-level conceptualizations. We categorize them into three types: extraction, retrieval, and generative-based methods, which are briefly demonstrated in Figure 3. We provide more discussions in Appendix A.

4.1 Extraction-Based Methods

Extracting concepts from text is the earliest paradigm for systematically collecting conceptualizations (Montgomery, 1982; Gelfand et al., 1998). It typically involves first extracting all possible concepts from the text, followed with identifying the relationships between these concepts. In this process, concepts are recognized either by looking for the most frequent words or by matching against a predefined list of patterns, such as "is a,", "is a type of", etc. Instances are then matched by looking for the subject of these patterns in the text, which forms instance-conceptualization pairs. The main advantages of extraction-based methods (Wang et al., 2016; Parameswaran et al., 2010; Rajagopal et al., 2013; Hovy et al., 2009; Krishnan et al., 2017; Pasca, 2009) are easy implementation, high processing speed, and free of training data. This has facilitated the development of many large-scale concept taxonomies and knowledge bases, such as WordNet (Miller, 1995), ConceptNet (Speer et al., 2017; Speer and Havasi, 2013, 2012b,a), Probase (Wu et al., 2012; Liang et al., 2017), and DBpedia (Auer et al., 2007; Bizer et al., 2009). However, these methods, while successful in extracting conceptual relationships from text, are limited by text quality, reliance on predefined concepts, lack of semantic understanding, difficulty handling ambiguous words, and poor generalization to new domains or unseen concepts.

4.2 Retrieval-Based Methods

4.2.1 Semantic-Based Retrieval

Semantic-based retrieval methods aim to obtain conceptualizations by looking at the semantic similarity between the input instance and the concepts in a pre-defined concept taxonomy. It typically in-

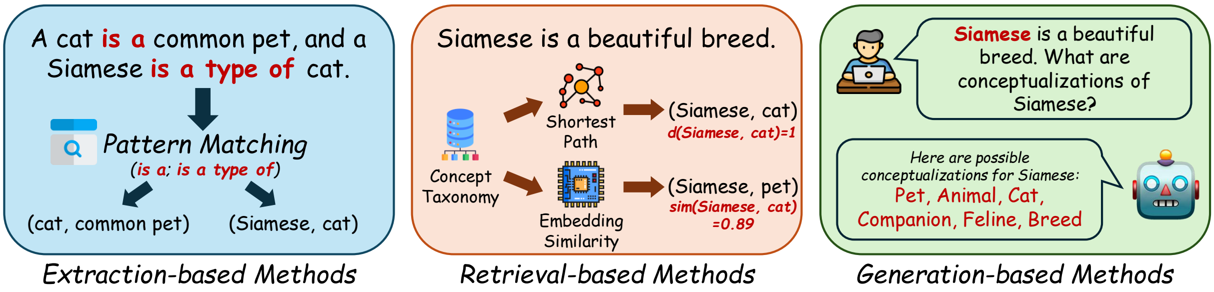

Figure 3: Conceptual demonstration of different types of methods in performing or collecting entity and event level conceptualizations. Instance and conceptualization pairs can be obtained at the end of each type of method.

volves representing both the instance and a set of concepts into a shared semantic space and calculating the similarity between them. One representative approach is to use WordNet (Miller, 1995), a large lexical database of English words, to calculate semantic similarity between two words as their shortest path in the WordNet hierarchy (Liu et al., 2004). Other methods (Natsev et al., 2007; Song et al., 2011, 2015; Koopman et al., 2012; Zheng and Yu, 2015; Chen et al., 2018; Hua et al., 2015) also share similar aspirations and define their own way of calculating such similarities. However, these methods are usually limited by the need for comprehensive and accurate knowledge bases, high computational costs, the inability to handle unseen concepts, and the loss of important semantic context, prompting the development of neural-based retrieval methods.

4.2.2 Neural-Based Retrieval

Neural-based retrieval methods overcome previous limitations by leveraging neural networks (or language models) to learn the semantic representations of the input instance and the concepts in the knowledge base or concept taxonomy. Then, the similarity between the input instance and the concepts can be calculated based on the learned representation embeddings. This approach can be benefitted by the advancement in language modeling, such as BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and DeBERTa (He et al., 2021, 2023). The most representative work in neural-based concept retrieval is AbstractATOMIC (He et al., 2024). It uses GlossBERT (Huang et al., 2019) to encode concepts (from WordNet and Probase) and instances (extracted from events in ATOMIC (Sap et al., 2019)) into embeddings and leverage cosine similarity and human annotations to collect conceptualizations in a large scale manner. Other methods (Wang et al., 2024c; Zhang et al., 2020, 2022; Lu et al., 2023; Wang et al., 2023b; Gao et al., 2022; Becker et al., 2021) similarly adopt different strategies in leveraging LMs as encoders, expanding the coverage of instances,training retrieval models. Despite their promising results, these methods are limited by their need for extensive labeled data, reliance on the completeness and accuracy of the knowledge base, and inability to retrieval new concepts that are out of training data.

4.3 Generative-Based Methods

4.3.1 Fine-Tuning-Based Generative Methods

Fine-tuning-based generative methods aim to take an entity or event as input and generate the concept directly via a fine-tuned generative language model. This approach allows the model to generate conceptualizations for new instances and offers maximum flexibility of the input. Several methods (Peng et al., 2022; Yuan et al., 2023; He et al., 2024; Wang et al., 2024c,b, 2023b) have adopted this paradigm in training generative conceptualizers, based on models such as GPT2 (Radford et al., 2019), BART (Lewis et al., 2020), and T5 (Raffel et al., 2020), for automated conceptualization acquisition. These methods typically train LMs on human-annotated or pre-existing conceptualization resources and yield outstanding results. However, fine-tuning-based generative methods are limited by their high computational cost, time-consuming and resource-intensive data collection, uncertain performance across diverse domains, and relatively low quality of novel concepts compared to human annotations. While these are common limitations associated with fine-tuned generative models, zero-shot generative methods using powerful LLMs and advanced prompting techniques potentially address these issues.

4.3.2 Zero-Shot Generative Methods

Finally, zero-shot generative-based methods leverage powerful LLMs (Brown et al., 2020; OpenAI, 2022, 2023; Reid et al., 2024; Touvron et al., 2023a,b) to generate the concept directly from an input instance. They rely on the vast amount of internal knowledge within the model and human-crafted prompts to efficiently distill conceptualizations and abstract knowledge from the models. This is particularly useful when training data is scarce or when the domain is new and there are no existing training data available. Existing methods (Wang et al., 2024a,c; Zheng et al., 2023; Zhao et al., 2024) all share similar aspirations in collecting conceptualizations. The benefits are significant, as these methods can collect conceptualizations efficiently and at low cost without specific fine-tuning. The resulting conceptualization knowledge base are thus scalable and downstream models trained on them typically have improved generalization ability to new instances and domains. However, to ensure high-quality generated conceptualizations, it is recommended to implement quality control mechanisms such as human evaluation or discriminators as post-filters. Recent studies (Wang et al., 2024a; Fang et al., 2024) have shown that commonsense plausibility estimators (Liu et al., 2023b) are effective for such quality control.

5 Downstream Applications

In this section, we discuss downstream tasks that can be benefited from applying conceptualizations. An overview of performances by different methods that leverage conceptualization, evaluated on various benchmarks, are shown in Figure 4.

5.1 Commonsense Reasoning

Commonsense reasoning is the ability to make inferences about the world based on common knowledge, which involves reasoning about everyday events and situations (Davis, 1990; Davis and Marcus, 2015). In this section, we discuss how conceptualizations benefit models in performing commonsense reasoning tasks.

Generative Commonsense Inference Modeling: The task of generative commonsense inference modeling (COMET; (Bosselut et al., 2019; Hwang et al., 2021)) aims to complete an inferential commonsense knowledge given a head event and a commonsense relation. State-of-the-art methods for COMET mainly fine-tune language models on large-scale commonsense knowledge bases, which suffer from data sparsity and lack of diversity in commonsense knowledge. Although transfer from LLMs helps (West et al., 2022, 2023), distilled knowledge tends to be too easy for models to learn and converge to trivial inferences. To address these issues, Wang et al. (2023b) proposed to leverage conceptualization as knowledge augmentation tools to improve COMET. Conceptualizations are first derived from head events to obtain abstracted events. Then, the tail of the original commonsense knowledge is placed back to the abstracted event to form abstracted commonsense knowledge. These derived abstract knowledge are then integrated with the original knowledge in commonsense knowledge bases to enrich the diversity of commonsense knowledge. Experiments results show consistent improvement in models' performances. Wang et al. (2024a) further show that, by instantiating conceptualizations in abstract knowledge back to other novel instances, models can be further improved by training with newly instantiated knowledge. Liu et al. (2023a) also proposed a task that aims to generate diverse sentences describing concept relationships in various everyday scenarios. Conceptualizations and associated abstract knowledge can further boost models' performances on this task.

Commonsense Question Answering: The task of commonsense question answering aims to answer questions that require commonsense knowledge. Various benchmarks and datasets have been proposed to evaluate LMs' performances, such as Abductive NLI (aNLI; (Bhagavatula et al., 2020)), CommonsenseQA (CSQA; (Talmor et al., 2019)), PhysicalIQA (PIQA; (Bisk et al., 2020)), SocialIQA (SIQA; (Sap et al., 2019)), and WinoGrande (WG; (Sakaguchi et al., 2021)). To obtain a generalizable model for commonsense question answering, the most effective pipeline fine-tunes language models on QA pairs synthesized from knowledge in commonsense knowledge bases (Ma et al., 2021; Shi et al., 2023; Wang et al., 2023a). The head h_o and relation r of a (h_o, r, t) triple are transformed into a question using natural language prompts, with the tail t serving as the correct answer option. Distractors or negative examples are generated by randomly sampling tails from triples that do not share common keywords with the head. To leverage conceptualization into the QA synthesis process, Wang et al. (2023a); Fang et al. (2024) have proposed two strategies: On the one hand, they improve distractor sampling by incorporating conceptualizations of head events into common words of the question, thereby enabling selection of more relevant distractors that improve the model's ability to discern correct answers from dis-

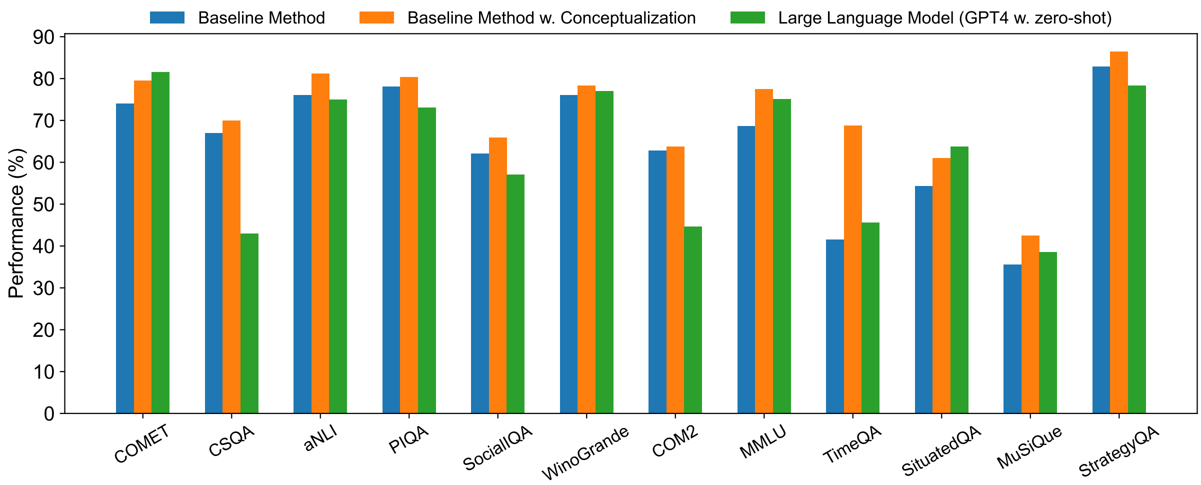

Figure 4: Empirical benefits of conceptualization in methods across various benchmarks. All reported results are sourced from respective original papers.

tractors. On the other hand, abstract knowledge derived from head events are integrated into original synthesized QA pairs, akin to COMET, to enrich the training data with diverse information, thereby enhancing the model's generalization capability in commonsense question answering tasks. Experimental results show that the proposed strategies significantly improve the performance of commonsense question answering with conceptualization.

5.2 Complex and Factual Reasoning

Complex reasoning refers to the ability to solve intricate problems that necessitate multiple steps of reasoning, which involves reasoning upon intricate scenarios, which may encompass multiple entities, events, and relations. Fang et al. (2024) proposed to synthesize complex queries based on commonsense knowledge triples from ATOMIC. Both human-defined rules and tails generated by large language models are utilized to generate these complex queries. The model is subsequently trained on these complex queries to enhance its capability to solve complex reasoning problems. In this context, conceptualizations of head events can be used as augmentations to generate more diverse and complex queries (Cui et al., 2017). This can assist the model in learning to solve more intricate problems. Simultaneously, conceptualizations of head events can also be used to generate more informative distractors. This can aid the model in learning to distinguish more effectively between correct answers and distractors.

Zheng et al. (2023) also developed a prompting method to improve the performance of LLMs on general and factual QA tasks. It involves instructing the model with a simple zero-shot prompt to consider each question abstractly by generating and probing relevant concepts, then using this knowledge in the prompt to generate the answer. This simple prompting method has been shown to significantly improve the performance of large language models on general QA tasks, including MMLU (Physics and Chemistry) (Hendrycks et al., 2021), TimeQA (Chen et al., 2021), StrategyQA (Geva et al., 2021), and MuSiQue (Trivedi et al., 2022). This work is interesting as it demonstrates that a simple prompting method can significantly enhance the performance of LLMs on general QA tasks.

5.3 Others

Aside from those two types of tasks, the line of works focusing on ultra-fine entity (Choi et al., 2018; Li et al., 2022; Dai and Zeng, 2023; Jiang et al., 2023; Li et al., 2023; Feng et al., 2023a; Dai et al., 2021; Liu et al., 2021; Onoe et al., 2021) and event typing (Zhou et al., 2023; Pepe et al., 2022; Chen et al., 2020) can also be benefited by conceptualization. These tasks aim to type named entities, nominal nouns, and pronouns into a set of free-form phrases. Conceptualizations can serve as a bridge between the surface form and the target type, which is crucial for these tasks.

6 Future Directions and Conclusions

Finally, we conclude our work by discussing two interesting future directions.

6.1 Controllable Generation and Hallucination Reduction

Firstly, we envision that conceptualization can assist controllable text generation (Feng et al., 2023b; Huang et al., 2023; Zhang et al., 2024). In some formulations, the task requires the model to generate a brief piece of text that remains consistent within a specific context or scope (Meng et al., 2022). Conceptualizations can be applied as additional supervision signals or constraints that guide the model to generate text whose conceptualizations align with those in the input theme, thereby enhancing the controllability of the generated text. This could be achieved by training a pair of conceptualization generator and discriminator, which could be used to generate the conceptualizations and evaluate their consistency between input and output text. Conceptualization can also serve as data augmentation tools to provide more training data, preferably guided with human annotation or large language models as loose teachers, for training more robust text generators that better align with the controllable targeting data.

Similarly, it may also benefit hallucination reduction (Choubey et al., 2023; Dale et al., 2023; Ji et al., 2023b; Sun et al., 2023). Hallucination (Ji et al., 2023a) refers to generating text that is unsupported by the input context, such as introducing information that is not present in the context or even contradicts it. In many reasoning scenarios, hallucination can be detrimental to the model's performance, and neutralizing it is crucial for ensuring the reliability of the generated text. Towards this objective, conceptualization can be similarly applied as external signals to verify the generated text and ensure its accuracy. By measuring the semantic distance of conceptualizations between the given input and generated contents, hallucinations can possibly be detected by finding clearly unrelated concepts appearing at both ends. Empirical metrics to measure such distance can be the shortest path length of concepts in taxonomies such as WordNet (Miller, 1995) and Probase (Wu et al., 2012), or even embedding similarity between different concepts. However, it's important to build a comprehensive set of conceptualizations of a given text to support such a verification process, as incomplete conceptualizations may cause erroneously detected hallucinations due to human-caused errors. We leave detailed implementations to future work.

6.2 Modeling Changes in Distribution

Conceptualization also plays a pivotal role in building reasoning systems that can capture situational changes in distribution to achieve System II reasoning (Sloman, 1996; Kahneman, 2011). Among the several components that make up System II reasoning, a key element is the ability to reason with situational changes in distribution (Bengio et al., 2021, 2019). These changes are triggered by environmental factors and actions by the agents themselves or others, especially when dealing with nonstationarities (Bengio, 2017). This ability can be achieved by dynamically recombining existing concepts in the given environment or action and learning from the resultant situational changes (Lake and Baroni, 2018; Bahdanau et al., 2019; de Vries et al., 2019). For instance, consider the event "PersonX is driving a car on a sunny day." A change in the weather from sunny to rainy could cause a different outcome, such as "PersonX becomes more cautious and drives slower." This illustrates that a change in weather conditions can lead to a change in the driver's behavior, representing an environmental change that triggers situational changes within the distribution of different weather conditions. In this process, the model is required to infer different changes that can possibly occur within a single event as the context, and reason about the potential outcome of each change. To model the distribution of different changes within an event, conceptualization can be used to represent the different states of the environment or action (Wang and Song, 2024). The model can then reason about the changes in distribution by manipulating the granularity of conceptualized changes. This type of distributional conceptualization not only provides an ontology for modeling the distribution of different changes within an event, but also assists the model in reasoning about the potential outcomes with appropriate abstract knowledge. Future works can leverage LLMs to curate benchmark datasets via sequential conceptualization generation and develop advanced systems for System II reasoning.

6.3 Conclusions

In conclusion, this work surveys conceptualizations by proposing a four-level hierarchical definition and reviewing representative works in acquiring, leveraging, and applying entity and event-level conceptualization to downstream reasoning tasks. We also propose several intriguing ideas related to conceptualizations that may inspire further research. We hope our work paves the way for more research works toward generalizable machine intelligence through conceptualization and fosters the development of more advanced systems that can capture, organize, and learn world knowledge through connection between concepts, much like humans do.

Limitations

The main limitations of our survey are two-fold. First, due to the vast amount of literature on conceptualization and conceptual knowledge across various datasets, we only cover the most representative works that stand out for their exceptional value and uniqueness in our taxonomy. Most of the papers are sourced from ACL Anthology[1], ACM Digital Library[2], and proceedings of leading artificial intelligence and machine learning conferences. Consequently, it is possible that some other related works are not included, but we aim to cover them in future versions. Second, our survey specifically focuses on entity and event level conceptualization, leaving document/paragraph level and system level conceptualization unaddressed. However, it is impossible to survey everything within one single submission. Future research can expand the scope of our survey to include more types of conceptualizations and modalities, such as categorization in the vision modality (Chen and Wang, 2004).

Ethics Statement

Our paper presents a comprehensive survey of conceptualization, with a specific focus on entity and event levels. All datasets and models reviewed in this survey are properly cited and are available under free-access licenses for research purposes. We did not conduct additional dataset curation or human annotation work. Therefore, to the best of our knowledge, this paper does not yield any ethical concerns.

References

Danis Alukaev, Semen Kiselev, Ilya Pershin, Bulat Ibragimov, Vladimir Ivanov, Alexey Kornaev, and Ivan Titov. 2023. Cross-modal conceptualization in bottleneck models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 5241–5253. Association for Computational Linguistics.

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary G. Ives. 2007. Dbpedia: A nucleus for a web of open data. In *The Semantic Web, 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007 + ASWC 2007, Busan, Korea, November 11-15, 2007*, volume 4825 of *Lecture Notes in Computer Science*, pages 722–735. Springer.

[1]https://aclanthology.org/
[2]https://dl.acm.org/

Dzmitry Bahdanau, Shikhar Murty, Michael Noukhovitch, Thien Huu Nguyen, Harm de Vries, and Aaron C. Courville. 2019. Systematic generalization: What is required and can it be learned? In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Jiaxin Bai, Xin Liu, Weiqi Wang, Chen Luo, and Yangqiu Song. 2023. Complex query answering on eventuality knowledge graph with implicit logical constraints. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Maria Becker, Katharina Korfhage, and Anette Frank. 2021. COCO-EX: A tool for linking concepts from texts to conceptnet. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations, EACL 2021, Online, April 19-23, 2021*, pages 119–126. Association for Computational Linguistics.

Yoshua Bengio. 2017. The consciousness prior. *CoRR*, abs/1709.08568.

Yoshua Bengio, Yann LeCun, and Geoffrey E. Hinton. 2021. Deep learning for AI. *Commun. ACM*, 64(7):58–65.

Yoshua Bengio et al. 2019. From system 1 deep learning to system 2 deep learning. In *Neural Information Processing Systems*.

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Wen-tau Yih, and Yejin Choi. 2020. Abductive commonsense reasoning. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Ning Bian, Xianpei Han, Le Sun, Hongyu Lin, Yaojie Lu, and Ben He. 2023. Chatgpt is a knowledgeable but inexperienced solver: An investigation of commonsense problem in large language models. *CoRR*, abs/2303.16421.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2020. PIQA: reasoning about physical commonsense in natural language. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 7432–7439. AAAI Press.

Christian Bizer, Jens Lehmann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann. 2009. Dbpedia - A crystallization point for the web of data. *J. Web Semant.*, 7(3):154–165.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. COMET: commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 4762–4779. Association for Computational Linguistics.

John D Bransford and Jeffery J Franks. 1971. The abstraction of linguistic ideas. *Cognitive psychology*, 2(4):331–350.

Falk Brauer, Michael Huber, Gregor Hackenbroich, Ulf Leser, Felix Naumann, and Wojciech M. Barczynski. 2010. Graph-based concept identification and disambiguation for enterprise search. In *Proceedings of the 19th International Conference on World Wide Web, WWW 2010, Raleigh, North Carolina, USA, April 26-30, 2010*, pages 171–180. ACM.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared 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 M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher 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 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Susan Carey. 1991. Knowledge acquisition: Enrichment or conceptual change. *The epigenesis of mind: Essays on biology and cognition*, pages 257–291.

Chunkit Chan, Jiayang Cheng, Weiqi Wang, Yuxin Jiang, Tianqing Fang, Xin Liu, and Yangqiu Song. 2024a. Exploring the potential of chatgpt on sentence level relations: A focus on temporal, causal, and discourse relations. In *Findings of the Association for Computational Linguistics: EACL 2024, St. Julian's, Malta, March 17-22, 2024*, pages 684–721. Association for Computational Linguistics.

Chunkit Chan, Cheng Jiayang, Yauwai Yim, Zheye Deng, Wei Fan, Haoran Li, Xin Liu, Hongming Zhang, Weiqi Wang, and Yangqiu Song. 2024b. Negotiationtom: A benchmark for stress-testing machine theory of mind on negotiation surrounding. *CoRR*, abs/2404.13627.

Lihan Chen, Jiaqing Liang, Chenhao Xie, and Yanghua Xiao. 2018. Short text entity linking with fine-grained topics. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM 2018, Torino, Italy, October 22-26, 2018*, pages 457–466. ACM.

Muhao Chen, Hongming Zhang, Haoyu Wang, and Dan Roth. 2020. What are you trying to do? semantic typing of event processes. In *Proceedings of the 24th Conference on Computational Natural Language Learning, CoNLL 2020, Online, November 19-20, 2020*, pages 531–542. Association for Computational Linguistics.

Wenhu Chen, Xinyi Wang, and William Yang Wang. 2021. A dataset for answering time-sensitive questions. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*.

Yihan Chen, Benfeng Xu, Quan Wang, Yi Liu, and Zhendong Mao. 2024. Benchmarking large language models on controllable generation under diversified instructions. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 17808–17816. AAAI Press.

Yixin Chen and James Ze Wang. 2004. Image categorization by learning and reasoning with regions. *J. Mach. Learn. Res.*, 5:913–939.

Jiayang Cheng, Lin Qiu, Tsz Ho Chan, Tianqing Fang, Weiqi Wang, Chunkit Chan, Dongyu Ru, Qipeng Guo, Hongming Zhang, Yangqiu Song, Yue Zhang, and Zheng Zhang. 2023. Storyanalogy: Deriving story-level analogies from large language models to unlock analogical understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 11518–11537. Association for Computational Linguistics.

Eunsol Choi, Omer Levy, Yejin Choi, and Luke Zettlemoyer. 2018. Ultra-fine entity typing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 87–96. Association for Computational Linguistics.

Prafulla Kumar Choubey, Alexander R. Fabbri, Jesse Vig, Chien-Sheng Wu, Wenhao Liu, and Nazneen Rajani. 2023. Cape: Contrastive parameter ensembling for reducing hallucination in abstractive summarization. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 10755–10773. Association for Computational Linguistics.

Wanyun Cui, Yanghua Xiao, Haixun Wang, Yangqiu Song, Seung-won Hwang, and Wei Wang. 2017. KBQA: learning question answering over QA corpora and knowledge bases. *Proc. VLDB Endow.*, 10(5):565–576.

Hongliang Dai, Yangqiu Song, and Haixun Wang. 2021. Ultra-fine entity typing with weak supervision from a masked language model. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 1790–1799. Association for Computational Linguistics.

Hongliang Dai and Ziqian Zeng. 2023. From ultra-fine to fine: Fine-tuning ultra-fine entity typing models to fine-grained. 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 2259–2270. Association for Computational Linguistics.

Dhairya Dalal, Paul Buitelaar, and Mihael Arcan. 2023. Calm-bench: A multi-task benchmark for evaluating causality-aware language models. In *Findings of the Association for Computational Linguistics: EACL 2023, Dubrovnik, Croatia, May 2-6, 2023*, pages 296–311. Association for Computational Linguistics.

David Dale, Elena Voita, Loïc Barrault, and Marta R. Costa-jussà. 2023. Detecting and mitigating hallucinations in machine translation: Model internal workings alone do well, sentence similarity even better. 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 36–50. Association for Computational Linguistics.

Ernest Davis. 1990. *Representations of commonsense knowledge*. notThenot Morgan Kaufmann series in representation and reasoning. Morgan Kaufmann.

Ernest Davis and Gary Marcus. 2015. Commonsense reasoning and commonsense knowledge in artificial intelligence. *Commun. ACM*, 58(9):92–103.

Harm de Vries, Dzmitry Bahdanau, Shikhar Murty, Aaron C. Courville, and Philippe Beaudoin. 2019. CLOSURE: assessing systematic generalization of CLEVR models. In *Visually Grounded Interaction and Language (ViGIL), NeurIPS 2019 Workshop, Vancouver, Canada, December 13, 2019*.

Zheye Deng, Weiqi Wang, Zhaowei Wang, Xin Liu, and Yangqiu Song. 2023. Gold: A global and local-aware denoising framework for commonsense knowledge graph noise detection. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 3591–3608. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Jonathan St BT Evans. 2003. In two minds: dual-process accounts of reasoning. *Trends in cognitive sciences*, 7(10):454–459.

Tobias Falke and Iryna Gurevych. 2019. Fast concept mention grouping for concept map-based multi-document summarization. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 695–700. Association for Computational Linguistics.

Tobias Falke, Christian M. Meyer, and Iryna Gurevych. 2017. Concept-map-based multi-document summarization using concept coreference resolution and global importance optimization. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing, IJCNLP 2017, Taipei, Taiwan, November 27 - December 1, 2017 - Volume 1: Long Papers*, pages 801–811. Asian Federation of Natural Language Processing.

Tianqing Fang, Zeming Chen, Yangqiu Song, and Antoine Bosselut. 2024. Complex reasoning over logical queries on commonsense knowledge graphs. *CoRR*, abs/2403.07398.

Tianqing Fang, Weiqi Wang, Sehyun Choi, Shibo Hao, Hongming Zhang, Yangqiu Song, and Bin He. 2021a. Benchmarking commonsense knowledge base population with an effective evaluation dataset. 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 8949–8964. Association for Computational Linguistics.

Tianqing Fang, Hongming Zhang, Weiqi Wang, Yangqiu Song, and Bin He. 2021b. DISCOS: bridging the gap between discourse knowledge and commonsense knowledge. In *WWW '21: The Web Conference 2021, Virtual Event / Ljubljana, Slovenia, April 19-23, 2021*, pages 2648–2659. ACM / IW3C2.

Amir Feder, Nadav Oved, Uri Shalit, and Roi Reichart. 2021. Causalm: Causal model explanation through counterfactual language models. *Comput. Linguistics*, 47(2):333–386.

Yanlin Feng, Adithya Pratapa, and David R. Mortensen. 2023a. Calibrated seq2seq models for efficient and generalizable ultra-fine entity typing. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 15550–15560. Association for Computational Linguistics.

Yuxi Feng, Xiaoyuan Yi, Xiting Wang, Laks V. S. Lakshmanan, and Xing Xie. 2023b. Dunst: Dual noisy self training for semi-supervised controllable text generation. 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 8760–8785. Association for Computational Linguistics.

Silin Gao, Jena D. Hwang, Saya Kanno, Hiromi Wakaki, Yuki Mitsufuji, and Antoine Bosselut. 2022. Comfact: A benchmark for linking contextual commonsense knowledge. In *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 1656–1675. Association for Computational Linguistics.

Boris Gelfand, Marilyn Wulfekuler, and WF Punch. 1998. Automated concept extraction from plain text. In *AAAI 1998 Workshop on Text Categorization*, pages 13–17.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did aristotle use a laptop? A question answering benchmark with implicit reasoning strategies. *Trans. Assoc. Comput. Linguistics*, 9:346–361.

Fausto Giunchiglia and Toby Walsh. 1992. A theory of abstraction. *Artif. Intell.*, 57(2-3):323–389.

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

Mutian He, Yangqiu Song, Kun Xu, and Dong Yu. 2020. On the role of conceptualization in commonsense knowledge graph construction. *CoRR*, abs/2003.03239.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: decoding-enhanced bert with disentangled attention. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Yining Hong, Li Yi, Josh Tenenbaum, Antonio Torralba, and Chuang Gan. 2021. PTR: A benchmark for part-based conceptual, relational, and physical reasoning. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 17427–17440.

Eduard H. Hovy, Zornitsa Kozareva, and Ellen Riloff. 2009. Toward completeness in concept extraction and classification. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, EMNLP 2009, 6-7 August 2009, Singapore, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 948–957. ACL.

Wen Hua, Zhongyuan Wang, Haixun Wang, Kai Zheng, and Xiaofang Zhou. 2015. Short text understanding through lexical-semantic analysis. In *31st IEEE International Conference on Data Engineering, ICDE 2015, Seoul, South Korea, April 13-17, 2015*, pages 495–506. IEEE Computer Society.

Luyao Huang, Chi Sun, Xipeng Qiu, and Xuanjing Huang. 2019. Glossbert: BERT for word sense disambiguation with gloss knowledge. 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 3507–3512. Association for Computational Linguistics.

Xuancheng Huang, Zijun Liu, Peng Li, Tao Li, Maosong Sun, and Yang Liu. 2023. An extensible plug-and-play method for multi-aspect controllable text generation. 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 15233–15256. Association for Computational Linguistics.

Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. 2021. (comet-) atomic 2020: On symbolic and neural commonsense knowledge graphs. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 6384–6392. AAAI Press.

Raghav Jain, Daivik Sojitra, Arkadeep Acharya, Sriparna Saha, Adam Jatowt, and Sandipan Dandapat. 2023. Do language models have a common sense regarding time? revisiting temporal commonsense reasoning in the era of large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 6750–6774. Association for Computational Linguistics.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2023a. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12):248:1–248:38.

Ziwei Ji, Zihan Liu, Nayeon Lee, Tiezheng Yu, Bryan Wilie, Min Zeng, and Pascale Fung. 2023b. RHO: reducing hallucination in open-domain dialogues with knowledge grounding. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 4504–4522. Association for Computational Linguistics.

Chengyue Jiang, Wenyang Hui, Yong Jiang, Xiaobin Wang, Pengjun Xie, and Kewei Tu. 2023. Recall, expand, and multi-candidate cross-encode: Fast and accurate ultra-fine entity typing. 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 11597–11609. Association for Computational Linguistics.

Zhijing Jin, Yuen Chen, Felix Leeb, Luigi Gresele, Ojasv Kamal, Zhiheng Lyu, Kevin Blin, Fernando Gonzalez Adauto, Max Kleiman-Weiner, Mrinmaya Sachan, and Bernhard Schölkopf. 2023. Cladder: A benchmark to assess causal reasoning capabilities of language models. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Serdar Kadioglu and Bernard Kleynhans. 2024. Building higher-order abstractions from the components of recommender systems. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 22998–23004. AAAI Press.

Daniel Kahneman. 2011. *Thinking, fast and slow*. macmillan.

Bevan Koopman, Guido Zuccon, Peter Bruza, Laurianne Sitbon, and Michael Lawley. 2012. An evaluation of corpus-driven measures of medical concept similarity for information retrieval. In *21st ACM International Conference on Information and Knowledge Management, CIKM'12, Maui, HI, USA, October 29 - November 02, 2012*, pages 2439–2442. ACM.

Adit Krishnan, Aravind Sankar, Shi Zhi, and Jiawei Han. 2017. Unsupervised concept categorization and extraction from scientific document titles. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, CIKM 2017, Singapore, November 06 - 10, 2017*, pages 1339–1348. ACM.

Ziva Kunda, Dale T. Miller, and Theresa Claire. 1990. Combining social concepts: The role of causal reasoning. *Cogn. Sci.*, 14(4):551–577.

Royi Lachmy, Valentina Pyatkin, Avshalom Manevich, and Reut Tsarfaty. 2022. Draw me a flower: Processing and grounding abstraction in natural language. *Trans. Assoc. Comput. Linguistics*, 10:1341–1356.

Faisal Ladhak, Esin Durmus, He He, Claire Cardie, and Kathleen R. McKeown. 2022. Faithful or extractive? on mitigating the faithfulness-abstractiveness tradeoff in abstractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 1410–1421. Association for Computational Linguistics.

Brenden M. Lake and Marco Baroni. 2018. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 2879–2888. PMLR.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7871–7880. Association for Computational Linguistics.

Bangzheng Li, Wenpeng Yin, and Muhao Chen. 2022. Ultra-fine entity typing with indirect supervision from natural language inference. *Trans. Assoc. Comput. Linguistics*, 10:607–622.

Na Li, Zied Bouraoui, and Steven Schockaert. 2023. Ultra-fine entity typing with prior knowledge about labels: A simple clustering based strategy. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 11744–11756. Association for Computational Linguistics.

Jiaqing Liang, Yanghua Xiao, Haixun Wang, Yi Zhang, and Wei Wang. 2017. Probase+: Inferring missing links in conceptual taxonomies. *IEEE Trans. Knowl. Data Eng.*, 29(6):1281–1295.

Hui Lin and Vincent Ng. 2019. Abstractive summarization: A survey of the state of the art. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 9815–9822. AAAI Press.

Chenzhengyi Liu, Jie Huang, Kerui Zhu, and Kevin Chen-Chuan Chang. 2023a. Dimongen: Diversified generative commonsense reasoning for explaining concept relationships. 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 4719–4731. Association for Computational Linguistics.

Jiacheng Liu, Wenya Wang, Dianzhuo Wang, Noah A. Smith, Yejin Choi, and Hannaneh Hajishirzi. 2023b. Vera: A general-purpose plausibility estimation model for commonsense statements. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 1264–1287. Association for Computational Linguistics.

Qing Liu, Hongyu Lin, Xinyan Xiao, Xianpei Han, Le Sun, and Hua Wu. 2021. Fine-grained entity typing via label reasoning. 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 4611–4622. Association for Computational Linguistics.

Shuang Liu, Fang Liu, Clement T. Yu, and Weiyi Meng. 2004. An effective approach to document retrieval via utilizing wordnet and recognizing phrases. In *SIGIR 2004: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Sheffield, UK, July 25-29, 2004*, pages 266–272. ACM.

Yihong Liu, Haotian Ye, Leonie Weissweiler, Philipp Wicke, Renhao Pei, Robert Zangenfeind, and Hinrich Schütze. 2023c. A crosslingual investigation of conceptualization in 1335 languages. 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 12969–13000. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Mengying Lu, Yuquan Wang, Jifan Yu, Yexing Du, Lei Hou, and Juanzi Li. 2023. Distantly supervised course concept extraction in moocs with academic discipline. 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 13044–13059. Association for Computational Linguistics.

Kaixin Ma, Filip Ilievski, Jonathan Francis, Yonatan Bisk, Eric Nyberg, and Alessandro Oltramari. 2021. Knowledge-driven data construction for zero-shot evaluation in commonsense question answering. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 13507–13515. AAAI Press.

Joshua Maynez, Priyanka Agrawal, and Sebastian Gehrmann. 2023. Benchmarking large language model capabilities for conditional generation. 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 9194–9213. Association for Computational Linguistics.

Tao Meng, Sidi Lu, Nanyun Peng, and Kai-Wei Chang. 2022. Controllable text generation with neurally-decomposed oracle. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2017. Pointer sentinel mixture models. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, and et al. 2024. Gemma: Open models based on gemini research and technology. *CoRR*, abs/2403.08295.

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

Christine A Montgomery. 1982. Concept extraction. *American journal of computational linguistics*, 8(2):70–73.

Gregory Murphy. 2004. *The big book of concepts*. MIT press.

Apostol Natsev, Alexander Haubold, Jelena Tesic, Lexing Xie, and Rong Yan. 2007. Semantic concept-based query expansion and re-ranking for multimedia retrieval. In *Proceedings of the 15th International Conference on Multimedia 2007, Augsburg, Germany, September 24-29, 2007*, pages 991–1000. ACM.

Yasumasa Onoe, Michael Boratko, Andrew McCallum, and Greg Durrett. 2021. Modeling fine-grained entity types with box embeddings. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual*

Event, August 1-6, 2021, pages 2051–2064. Association for Computational Linguistics.

OpenAI. 2022. Chatgpt: Optimizing language models for dialogue. *OpenAI*.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Shiyan Ou, Viktor Pekar, Constantin Orasan, Christian Spurk, and Matteo Negri. 2008. Development and alignment of a domain-specific ontology for question answering. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2008, 26 May - 1 June 2008, Marrakech, Morocco*. European Language Resources Association.

Aditya G. Parameswaran, Hector Garcia-Molina, and Anand Rajaraman. 2010. Towards the web of concepts: Extracting concepts from large datasets. *Proc. VLDB Endow.*, 3(1):566–577.

Marius Pasca. 2009. Outclassing wikipedia in open-domain information extraction: Weakly-supervised acquisition of attributes over conceptual hierarchies. In *EACL 2009, 12th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference, Athens, Greece, March 30 - April 3, 2009*, pages 639–647. The Association for Computer Linguistics.

Hao Peng, Xiaozhi Wang, Shengding Hu, Hailong Jin, Lei Hou, Juanzi Li, Zhiyuan Liu, and Qun Liu. 2022. COPEN: probing conceptual knowledge in pre-trained language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 5015–5035. Association for Computational Linguistics.

Sveva Pepe, Edoardo Barba, Rexhina Blloshmi, and Roberto Navigli. 2022. STEPS: semantic typing of event processes with a sequence-to-sequence approach. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 11156–11164. AAAI Press.

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is chatgpt a general-purpose natural language processing task solver? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 1339–1384. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Dheeraj Rajagopal, Erik Cambria, Daniel Olsher, and Kenneth Kwok. 2013. A graph-based approach to commonsense concept extraction and semantic similarity detection. In *22nd International World Wide Web Conference, WWW '13, Rio de Janeiro, Brazil, May 13-17, 2013, Companion Volume*, pages 565–570. International World Wide Web Conferences Steering Committee / ACM.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy P. Lillicrap, Jean-Baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, Ioannis Antonoglou, Rohan Anil, Sebastian Borgeaud, Andrew M. Dai, Katie Millican, Ethan Dyer, Mia Glaese, Thibault Sottiaux, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, James Molloy, Jilin Chen, Michael Isard, Paul Barham, Tom Hennigan, Ross McIlroy, Melvin Johnson, Johan Schalkwyk, Eli Collins, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, Clemens Meyer, Gregory Thornton, Zhen Yang, Henryk Michalewski, Zaheer Abbas, Nathan Schucher, Ankesh Anand, Richard Ives, James Keeling, Karel Lenc, Salem Haykal, Siamak Shakeri, Pranav Shyam, Aakanksha Chowdhery, Roman Ring, Stephen Spencer, Eren Sezener, and et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *CoRR*, abs/2403.05530.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Winogrande: an adversarial winograd schema challenge at scale. *Commun. ACM*, 64(9):99–106.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. Social iqa: Commonsense reasoning about social interactions. 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 4462–4472. Association for Computational Linguistics.

Haochen Shi, Weiqi Wang, Tianqing Fang, Baixuan Xu, Wenxuan Ding, Xin Liu, and Yangqiu Song. 2023. QADYNAMICS: training dynamics-driven synthetic QA diagnostic for zero-shot commonsense question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 15329–15341. Association for Computational Linguistics.

Push Singh, Thomas Lin, Erik T. Mueller, Grace Lim, Travell Perkins, and Wan Li Zhu. 2002. Open mind common sense: Knowledge acquisition from the general public. In *On the Move to Meaningful Internet*

Systems, 2002 - DOA/CoopIS/ODBASE 2002 Confederated International Conferences DOA, CoopIS and ODBASE 2002 Irvine, California, USA, October 30 - November 1, 2002, Proceedings, volume 2519 of *Lecture Notes in Computer Science*, pages 1223–1237. Springer.

Steven A Sloman. 1996. The empirical case for two systems of reasoning. *Psychological bulletin*, 119(1):3.

Yangqiu Song, Haixun Wang, Zhongyuan Wang, Hongsong Li, and Weizhu Chen. 2011. Short text conceptualization using a probabilistic knowledgebase. In *IJCAI 2011, Proceedings of the 22nd International Joint Conference on Artificial Intelligence, Barcelona, Catalonia, Spain, July 16-22, 2011*, pages 2330–2336. IJCAI/AAAI.

Yangqiu Song, Shusen Wang, and Haixun Wang. 2015. Open domain short text conceptualization: A generative + descriptive modeling approach. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina, July 25-31, 2015*, pages 3820–3826. AAAI Press.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 4444–4451. AAAI Press.

Robyn Speer and Catherine Havasi. 2012a. Conceptnet 5. *Tiny Trans. Comput. Sci.*, 1.

Robyn Speer and Catherine Havasi. 2012b. Representing general relational knowledge in conceptnet 5. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012, Istanbul, Turkey, May 23-25, 2012*, pages 3679–3686. European Language Resources Association (ELRA).

Robyn Speer and Catherine Havasi. 2013. Conceptnet 5: A large semantic network for relational knowledge. In Iryna Gurevych and Jungi Kim, editors, *The People's Web Meets NLP, Collaboratively Constructed Language Resources*, Theory and Applications of Natural Language Processing, pages 161–176. Springer.

Arjun Subramonian, Xingdi Yuan, Hal Daumé III, and Su Lin Blodgett. 2023. It takes two to tango: Navigating conceptualizations of NLP tasks and measurements of performance. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 3234–3279. Association for Computational Linguistics.

Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. Yago: a core of semantic knowledge. In *Proceedings of the 16th International Conference on World Wide Web, WWW 2007, Banff, Alberta, Canada, May 8-12, 2007*, pages 697–706. ACM.

Bin Sun, Yitong Li, Fei Mi, Fanhu Bie, Yiwei Li, and Kan Li. 2023. Towards fewer hallucinations in knowledge-grounded dialogue generation via augmentative and contrastive knowledge-dialogue. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 1741–1750. Association for Computational Linguistics.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4149–4158. Association for Computational Linguistics.

Qingyu Tan, Hwee Tou Ng, and Lidong Bing. 2023. Towards benchmarking and improving the temporal reasoning capability of large language models. 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 14820–14835. Association for Computational Linguistics.

Joshua B Tenenbaum, Charles Kemp, Thomas L Griffiths, and Noah D Goodman. 2011. How to grow a mind: Statistics, structure, and abstraction. *science*, 331(6022):1279–1285.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas

Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. Musique: Multi-hop questions via single-hop question composition. *Trans. Assoc. Comput. Linguistics*, 10:539–554.

Shuting Wang, Alexander Ororbia, Zhaohui Wu, Kyle Williams, Chen Liang, Bart Pursel, and C. Lee Giles. 2016. Using prerequisites to extract concept maps fromtextbooks. In *Proceedings of the 25th ACM International Conference on Information and Knowledge Management, CIKM 2016, Indianapolis, IN, USA, October 24-28, 2016*, pages 317–326. ACM.

Weiqi Wang, Tianqing Fang, Wenxuan Ding, Baixuan Xu, Xin Liu, Yangqiu Song, and Antoine Bosselut. 2023a. CAR: conceptualization-augmented reasoner for zero-shot commonsense question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 13520–13545. Association for Computational Linguistics.

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. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*. Association for Computational Linguistics.

Weiqi Wang, Tianqing Fang, Baixuan Xu, Chun Yi Louis Bo, Yangqiu Song, and Lei Chen. 2023b. CAT: A contextualized conceptualization and instantiation framework for commonsense reasoning. 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 13111–13140. Association for Computational Linguistics.

Weiqi Wang and Yangqiu Song. 2024. MARS: Benchmarking the metaphysical reasoning abilities of language models with a multi-task evaluation dataset. *CoRR*, abs/2406.02106.

Wenbo Wang, Yang Gao, Heyan Huang, and Yuxiang Zhou. 2019. Concept pointer network for abstractive summarization. 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 3074–3083. Association for Computational Linguistics.

Yi Ru Wang, Jiafei Duan, Dieter Fox, and Siddhartha S. Srinivasa. 2023c. NEWTON: are large language models capable of physical reasoning? In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 9743–9758. Association for Computational Linguistics.

Zhaowei Wang, Wei Fan, Qing Zong, Hongming Zhang, Sehyun Choi, Tianqing Fang, Xin Liu, Yangqiu Song, Ginny Y. Wong, and Simon See. 2024b. Absinstruct: Eliciting abstraction ability from llms through explanation tuning with plausibility estimation. *CoRR*, abs/2402.10646.

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.

Zihao Wang, Weizhi Fei, Hang Yin, Yangqiu Song, Ginny Y. Wong, and Simon See. 2023d. Wasserstein-fisher-rao embedding: Logical query embeddings with local comparison and global transport. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 13679–13696. Association for Computational Linguistics.

Zihao Wang, Yangqiu Song, Ginny Y. Wong, and Simon See. 2023e. Logical message passing networks with one-hop inference on atomic formulas. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Zihao Wang, Hang Yin, and Yangqiu Song. 2021. Benchmarking the combinatorial generalizability of complex query answering on knowledge graphs. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*.

Peter West, Chandra Bhagavatula, Jack Hessel, Jena D. Hwang, Liwei Jiang, Ronan Le Bras, Ximing Lu, Sean Welleck, and Yejin Choi. 2022. Symbolic knowledge distillation: from general language models to commonsense models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 4602–4625. Association for Computational Linguistics.

Peter West, Ronan Le Bras, Taylor Sorensen, Bill Yuchen Lin, Liwei Jiang, Ximing Lu, Khyathi Chandu, Jack Hessel, Ashutosh Baheti, Chandra Bhagavatula, and Yejin Choi. 2023. Novacomet: Open commonsense foundation models with symbolic knowledge distillation. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 1127–1149. Association for Computational Linguistics.

Wentao Wu, Hongsong Li, Haixun Wang, and Kenny Qili Zhu. 2012. Probase: a probabilistic taxonomy for text understanding. In *Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2012, Scottsdale, AZ, USA, May 20-24, 2012*, pages 481–492. ACM.

Shan Yang, Yongfei Zhang, Guanglin Niu, Qinghua Zhao, and Shiliang Pu. 2021. Entity concept-enhanced few-shot relation extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 2: Short Papers), Virtual Event, August 1-6, 2021*, pages 987–991. Association for Computational Linguistics.

Andrew Yates, Nazli Goharian, and Ophir Frieder. 2015. Extracting adverse drug reactions from social media. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA*, pages 2460–2467. AAAI Press.

Changlong Yu, Xin Liu, Jefferson Maia, Yang Li, Tianyu Cao, Yifan Gao, Yangqiu Song, Rahul Goutam, Haiyang Zhang, Bing Yin, and Zheng Li. 2024. Cosmo: A large-scale e-commerce common sense knowledge generation and serving system at amazon. In *Companion of the 2024 International Conference on Management of Data*, SIGMOD/PODS '24, page 148–160, New York, NY, USA. Association for Computing Machinery.

Changlong Yu, Weiqi Wang, Xin Liu, Jiaxin Bai, Yangqiu Song, Zheng Li, Yifan Gao, Tianyu Cao, and Bing Yin. 2023. Folkscope: Intention knowledge graph construction for e-commerce commonsense discovery. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 1173–1191. Association for Computational Linguistics.

Chenhan Yuan, Qianqian Xie, Jimin Huang, and Sophia Ananiadou. 2024. Back to the future: Towards explainable temporal reasoning with large language models. In *Proceedings of the ACM on Web Conference 2024, WWW 2024, Singapore, May 13-17, 2024*, pages 1963–1974. ACM.

Siyu Yuan, Deqing Yang, Jinxi Liu, Shuyu Tian, Jiaqing Liang, Yanghua Xiao, and Rui Xie. 2023. Causality-aware concept extraction based on knowledge-guided prompting. 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 9255–9272. Association for Computational Linguistics.

Hanqing Zhang, Haolin Song, Shaoyu Li, Ming Zhou, and Dawei Song. 2024. A survey of controllable text generation using transformer-based pre-trained language models. *ACM Comput. Surv.*, 56(3):64:1–64:37.

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

Hongming Zhang, Xin Liu, Haojie Pan, Yangqiu Song, and Cane Wing-Ki Leung. 2020. ASER: A large-scale eventuality knowledge graph. In *WWW '20: The Web Conference 2020, Taipei, Taiwan, April 20-24, 2020*, pages 201–211. ACM / IW3C2.

Ruochen Zhao, Tan Wang, Yongjie Wang, and Shafiq Joty. 2024. Explaining language model predictions with high-impact concepts. In *Findings of the Association for Computational Linguistics: EACL 2024, St. Julian's, Malta, March 17-22, 2024*, pages 995–1012. Association for Computational Linguistics.

Huaixiu Steven Zheng, Swaroop Mishra, Xinyun Chen, Heng-Tze Cheng, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2023. Take a step back: Evoking reasoning via abstraction in large language models. *CoRR*, abs/2310.06117.

Jiaping Zheng and Hong Yu. 2015. Key concept identification for medical information retrieval. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 579–584. The Association for Computational Linguistics.

Bo Zhou, Yubo Chen, Kang Liu, and Jun Zhao. 2023. Event process typing via hierarchical optimal transport. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pages 14038–14046. AAAI Press.

Appendices

A Conceptualization Acquisition Methods

In this appendix, we elaborate further on different methods of acquiring conceptualization and provide detailed explanations of their weaknesses.

A.1 Extraction Based Methods

For methods that follow the concept extraction paradigm, Wang et al. (2016) proposed a framework to optimize both tasks simultaneously, leading to stronger performances even compared to supervised concept extraction methods. Parameswaran et al. (2010) also proposed a market-basket-based solution, which adapts statistical measures of support and confidence to design a concept extraction algorithm that achieved high precision in concept extraction. Rajagopal et al. (2013) proposed a solution to extract concepts from commonsense text, which uncovers many novel pieces of knowledge that cannot be found in the original corpora. Hovy et al. (2009); Krishnan et al. (2017); Pasca (2009) similarly proposed their solutions for large-scale concept extraction for more efficient data mining.

While these methods have been successful in extracting concepts and relationships from text, they have several limitations. First, they are heavily dependent on the quality of the text and the predefined list of concepts. If the text is noisy or contains many irrelevant words, the performance of these methods can degrade significantly, and the resulting extracted concepts may also tend to be noisy. Second, it's important to note that these methods primarily rely on parsing or pattern matching techniques on text and do not capture semantic information from the text. This potentially makes extracted concepts represented as isolated entities without any context or relationships and could result in misextraction of concepts or relationships, especially when the text contains ambiguous or polysemous words. For example, the word "bank" can refer to a financial institution, a river bank, or a memory bank, and without proper context, it's difficult to determine the correct meaning of it, thus leading to incorrect concept extraction. A low-performance parser, if wrongly parsing these words, may also lead to noisy results. Lastly, these methods are not able to generalize well to unseen concepts or text patterns that are not present in the predefined list of concepts. This limits their applicability to new domains or tasks that require the extraction of novel concepts or relationships. For example, to extract concepts from medical or legal domain text, specific patterns or extraction rules need to be designed, which may not be present when extracting normal conversational text.

A.2 Retrieval Based Methods

A.2.1 Semantic-Based Retrieval

To perform semantic-based retrieval, (Natsev et al., 2007) proposed several approaches for semantic concept-based query expansion and re-ranking in multimedia retrieval, achieving consistent performance improvement compared to text retrieval and multimodal retrieval baseline. (Song et al., 2011, 2015) improved text understanding by using a probabilistic knowledge base based on concepts and developed a Bayesian inference mechanism to conceptualize words and short text. Experimental results show significant improvements on text clustering compared to purely statistical methods and methods that use existing knowledge bases. (Koopman et al., 2012) proposed a corpus-driven approach, adapted from LSA, to retrieve medical concepts with semantic similarity measures. (Zheng and Yu, 2015) similarly used topic modeling and key concept retrieval methods to construct queries from electronic health records, which significantly improves the retrieval of tailored online consumer-oriented health education materials.

Although these methods have shown promising results in various domains, they have several limitations. First, the performance of semantic-based retrieval heavily relies on the quality of the knowledge base or concept taxonomy. In other words, it requires the knowledge base to be comprehensive, accurate, hierarchical, and up-to-date. There are very few knowledge bases that meet all these requirements, and constructing such a knowledge base is a non-trivial task. With incomplete knowledge bases, which are common in practice, the performance of semantic-based retrieval methods can be significantly degraded. Second, semantic-based retrieval methods are usually computationally expensive, as they require calculating the similarity between the input instance and all concepts in the knowledge base. This can induce exponentially increasing computational cost as the size of the knowledge base grows. When dealing with large-scale applications, this even becomes infeasible. Though caching and indexing techniques can be used to speed up the retrieval process, they are not always effective and cannot generalize well when unseen concepts or instances are encountered. Third, semantic-based retrieval methods still do not consider the semantic context of the input instance. A straightforward formulation is that the model treats the input instance as a bag of words and ignores the word order and syntactic structure. This can lead to a loss of important semantic information, especially when the input instance is long and complex. In this case, the semantic similarity between the input instance and the concepts in the knowledge base may not reflect the true semantic relevance.

A.2.2 Neural-Based Retrieval

For neural-based retrieval, aside from He et al. (2024), (Lu et al., 2023) similarly proposes a novel three-stage framework, which leverages the power of pre-trained language models explicitly and implicitly and employs discipline-embedding models with a self-train strategy based on label generation refinement across different domains.

To deal with the large amount of unlabeled data after human annotation, (Wang et al., 2023b) further proposed a semi-supervised method to unlabel the data with a supervised trained conceptualization discriminator. The discriminator is trained to rate the plausibility of unlabeled conceptualization and the model will be further refined by training on a concatenation of labeled and unlabeled data. This results in a significant improvement in the performance of the conceptualization discriminator, thus enhancing the quality of the retrieved concepts.

Despite these promising results in concept retrieval, neural-based retrieval methods have several limitations. First, these methods are usually data-hungry and require a large amount of labeled data for training. This can be a bottleneck in practice, as labeling data is often expensive and time-consuming. Human annotations are usually required to collect such data, and for models to be generalizable across different domains, the labeled data should be diverse and representative. This is even more costly and challenging to obtain. Second, neural-based retrieval methods still rely on the coverage and quality of the knowledge base or concept taxonomy. If the knowledge base is incomplete or inaccurate, the performance of neural-based retrieval methods can be significantly affected. Moreover, they cannot generate new concepts or instances that are not in the knowledge base, which limits their generalization ability.

A.3 Generative-Based Methods

A.3.1 Fine-Tuning-Based Generative Methods

While most fine-tuning based methods are explicitly discussed in the main body, we explain their limitations here. First, these methods are usually computationally expensive, as they require fine-tuning a large pre-trained language model on a specific dataset. Both the fine-tuning and the training data collection process can be time-consuming and resource-intensive. Extensive crowd-sourcing or human annotations are usually required to collect high-quality training data, which can be costly and challenging to obtain when the domain coverage scales up. Second, the feasibility of fine-tuning-based generative methods on other domains, such as medical or legal text, is still an open question. The performance of these methods heavily relies on the quality and diversity of the training data, and it's not clear how well they can generalize to new domains or tasks as text understanding abilities vary across different domains. For social commonsense, pre-trained language models have shown strong performance possibly due to a large overlap in the training data distribution, but for other domains, the performance is still unclear. Lastly, although existing studies have shown that fine-tuning based generators can deliver novel concepts that are not in the training data, such a ratio is relatively low and the quality of the generated concepts is still not as good as human annotated ones. This is expected as the models are fitted into the distribution of the training data, and it's hard for them to generate concepts that are out of the distribution.

A.3.2 Zero-Shot Generative Methods

Zero-shot generative methods aim to generate the desired output for any task's input without any task-specific fine-tuning. A very representative example of such generative models is the recently popularized LLMs (OpenAI, 2022, 2023; Touvron et al., 2023a,b; Mesnard et al., 2024; Reid et al., 2024). These models have been pre-trained on very large corpora, including those from the web, Wikipedia, books, and more, and have shown strong performance in various natural language processing tasks, including text generation (Maynez et al., 2023; Chen et al., 2024), temporal reasoning (Tan et al., 2023; Yuan et al., 2024), causal reasoning (Chan et al., 2024a; Dalal et al., 2023; Jin et al., 2023), commonsense reasoning (Jain et al., 2023; Bian et al., 2023; Fang et al., 2021b,a; Deng et al., 2023), logical reasoning (Wang et al., 2023d,e, 2021; Bai et al., 2023), and more (Qin et al., 2023; Cheng et al., 2023; Chan et al., 2024b).

In the context of conceptualization acquisition, zero-shot generative methods aim to generate conceptualizations for instances without any instance-conceptualization pairs in the training data. Wang et al. (2024a) proposed a few-shot knowledge distillation method to distill conceptualizations and associated abstract inferential knowledge from a large language model to a large-scale knowledge base. Wang et al. (2024c) also proposed acquiring conceptualizations for entities and events in ASER by instructing ChatGPT with a few-shot prompt. They further designed an instruction-tuning based method to evoke more conceptualizations from large language models by fine-tuning them with explanations on how the conceptualization is derived from the instance and their plausible reasoning chains (Wang et al., 2024b). Zheng et al. (2023) proposed a simple prompting technique, inspired by chain-of-thought reasoning, that enables LLMs to do conceptualizations to derive high-level concepts and first principles from instances containing specific details. Zhao et al. (2024) advanced this idea by proposing to extract predictive high-level features (concepts) from a large language model's hidden layer activations.

The benefits of these methods are twofold. First, such generation can introduce conceptualizations at a very low cost, as the models are pre-trained and do not require any task-specific fine-tuning. The only burden seems to be deployment and inference cost, which require a large amount of computational resources and time for large-scale generation. However, compared to all previous fine-tuning-based methods, zero-shot generative methods are much more efficient and scalable, as they do not require any training data or fine-tuning process. Second, zero-shot generative methods have shown strong generalization capabilities to new instances and domains. They can generate conceptualizations for instances that are not in the training data and have shown strong performance in various conceptualization acquisition tasks. This is particularly useful when the training data is scarce or when the domain is new, and there are no existing training data available. Since these large language models are pre-injected with vast amounts of knowledge, this makes generalization possible.