Large Language Models as a Tool for Mining Object Knowledge

Hannah Y. An*, Lenhart K. Schubert

University of Rochester, Rochester, New York, 14620, USA

Abstract

While Large language models (LLMs) are apt to confabulate facts when questioned about obscure named entities or technical domains, we hypothesize that their *general* knowledge about objects in the everyday world is largely sound. Based on that hypothesis, this paper investigates LLMs' ability to formulate explicit knowledge about common physical artifacts, focusing on their parts and materials. Our work distinguishes between the substances that comprise an entire object and those that constitute its parts—a previously underexplored distinction in knowledge base construction. Using few-shot and zero-shot multi-step prompting, we produce a repository of data on the parts and materials of about 2,300 objects and their subtypes (averaging 4–5 subtypes per object, depending on prompting strategy). The breadth of coverage and level of detail in this repository exceeds most people's knowledge about artifacts, or what can be found in other structured sources, as shown by our evaluations. Besides demonstrating the scope and reliability of LLM-derived general knowledge about artifacts and their parts and composition, our repository should prove useful in AI reasoning systems where artifacts play an important role or in multimodal learning, and serve as an explicit knowledge source (analogous to knowledge graphs) for LLMs performing multi-hop question answering.

Keywords

commonsense knowledge acquisition, lexical semantics, ontology information extraction, knowledge base construction, dataset evaluation

1. Introduction

Large language models (LLMs) have demonstrated remarkable capabilities in question answering, dialogue, summarization, and even complex reasoning, such as theory-of-mind inference and scientific explanation by analogy [1, 2, 3]. However, their tendency to confabulate facts in trying to satisfy user demands raises concerns about the reliability of their implicit knowledge. We hypothesize that LLMs' *general* commonsense knowledge is trustworthy, in contrast with their unreliability concerning little-known named entities or highly abstruse topics. To test this hypothesis, we focus on artifacts, their parts, and materials, as these concepts are deeply embedded in human cognition from an early age and play a central role in day-to-day life. Even young children learn to identify the parts and materials of everyday objects, distinguishing between glass, metal, wood, and fabric. They recognize that glass is transparent while metal is not, that metal pots withstand heat while paper plates do not, and that a tablecloth folds while a wooden tray remains rigid. They also understand functional affordances, such as using a broom's brush for sweeping and preferring sofa cushions over wooden seats for comfort.

Our work systematically extracts and evaluates LLM-derived knowledge about the part structure and material composition of common artifacts, creating a structured and interpretable resource. We compare this extracted knowledge to human-annotated data and existing linguistic resources, assessing coverage, specificity, and fidelity to human understanding. While prior work has explored mining or crowdsourcing semantic features of human conceptual knowledge, no large-scale resource explicitly distinguishes between the materials composing an entire object versus those of its individual parts.

Proceedings of the Joint Ontology Workshops (JOWO) – Episode XI: The Sicilian Summer under the Etna, co-located with the 15th International Conference on Formal Ontology in Information Systems (FOIS 2025), September 8–9, 2025, Catania, Italy

*Corresponding author.

yan2@cs.rochester.edu (H. Y. An); schubert@cs.rochester.edu (L. K. Schubert)

https://anyhannah.github.io/ (H. Y. An); https://www.cs.rochester.edu/~schubert/ (L. K. Schubert)

0009-0009-4551-7891 (H. Y. An); 0000-0001-8398-9923 (L. K. Schubert)

For instance, while a violin is broadly described as wooden, its strings are made from non-wooden materials such as catgut, nylon, or steel, a distinction often missing in prior knowledge bases.

Our approach mines part-material knowledge from LLMs and represents it in an explicit, structured format. This resource can supplement knowledge graphs and support learning and explanation, particularly in AI applications that do not rely on full-scale LLMs, such as tutorial systems that use knowledge maps to question the learner such as [4], language-enabled robots manipulating household utensils [5], and knowledge-assisted visual interpretation and question-answering systems [6, 7, 8, 9, 10].

By analyzing a subset of our resource through human judgments, which encompasses the data of 2,313 physical objects, we find that LLMs can extract structured artifact knowledge that largely aligns with human understanding, though the specificity and depth depend on prompting strategies. In some cases, LLM-generated knowledge remains superficial; in others, it surpasses typical layman conceptions. This structured dataset not only offers insight into how LLMs retrieve commonsense artifact knowledge but also serves as a valuable NLP resource for scientific research.[1]

2. Related Work

2.1. Pattern-based extraction

The pattern-based approach in knowledge acquisition identifies relationships in text by matching predefined lexical patterns. While fairly productive, it struggles with complex data patterns and implicit knowledge not reflected in surface patterns. Early efforts, like Berland and Charniak [11] and Poesio et al. [12], relied on hand-built patterns to detect part–whole relationships. Later, search by Girju et al. [13] and van Hage et al. [14] automated the process using web data, refining it further with classifiers to determine meronymy relationships. Girju et al. used a two-way classifier, while Poesio and Almuhareb [15] employed a multi-way classifier to categorize detected attributes. Later advances, such as that of Tesfaye and Zock [16], used vector similarity to cluster co-occurring nouns, enhancing both accuracy and coverage. While these methods achieved moderate success at that time, their reliance on surface patterns or co-occurrence statistics ultimately constrained their ability to capture the more implicit and syntactically diverse expressions in natural language.

2.2. Human Annotation

Human annotation often involves significant expenses and time due to the need for detailed and accurate annotations. When data is collected through crowdsourcing, ensuring quality control becomes challenging, especially when complex semantic relationships need to be precisely captured.

Several knowledge resources integrate crowd-sourced and expert annotations to address these challenges. ConceptNet [17] incorporates both methods to build a commonsense knowledge graph, linking concepts with relations like `MadeOf` and `PartOf`. However, the annotation schema can lead to ambiguity; `MadeOf` describes an object as a whole, while `PartOf` identifies components without specifying their materials, even when the parts are made of different materials from other parts. For example, the tuples (bicycle, `MadeOf`, metal) and (bicycle seat, `PartOf`, bicycle) does not specify that a bicycle saddle, which is a part of a bicycle seat, can be made of leather. WordNet [18] provides an expert-curated lexical database, where human annotators organize words into semantic hierarchies and mark part–whole relations (meronymy). Its concise definitions sometimes encode part and material information, as in the example of 'felt-tip pen,' whose definition states the writing tip is made of felt. Similarly, the ParRoT dataset, developed by Gu et al. [19], offers a fine-grained, human-annotated collection of part lists for 100 everyday objects. This dataset goes beyond major components, detailing sub-parts (e.g., the reflective glass and spring in a flashlight) to provide a comprehensive model of objects, particularly in terms of their structural and functional relationships.

Semantic feature datasets, such as McRae norms [20] and CSLB concept property norms [21], rely on human expertise to generate reliable annotations. These datasets include features like part–whole

[1]All data, code, and supplementary materials are available at `https://github.com/anyhannah/composition-miner`

relations and material composition such as `has_a` and `made_of`. The McRae norms distinguish between `made_of`, used for substances, and `made_from`, used for origins (e.g., prune made from plums). Additionally, they differentiate between essential components (e.g., an engine or a door) and non-essential parts (e.g., a bed's comforter) or functional aspects (e.g., an elevator's capacity), assigning them distinct labels. While existing human-annotated resources either prioritize broad coverage at the expense of consistency and clarity (e.g., ConceptNet) or offer detailed but narrowly scoped annotations (e.g., ParRoT, WordNet), the approach introduced here combines both scale and specificity. Later sections show that this leads to improved generalization, as reflected in external recall metrics (see Section 4.3).

2.3. Use of language models

LLMs excel at capturing complex patterns and implicit world knowledge from pre-training, enabling them to generate and retrieve detailed information with high fluency. While LLMs seem to reliably encode general knowledge about everyday entities (our hypothesis here), their accuracy declines when generating specific factoids, often leading to inconsistencies or misclassifications in certain contexts.

TransOMCS [22] constructs a large commonsense graph by extracting information from text data. However, it sometimes misinterprets relationships, leading to errors like (cement, `MadeOf`, consist) and (dog, `PartOf`, walker). These examples illustrate both the scalability of the text-based approach and the risk of incorrect data. ATOMIC$^{20}_{20}$ [23] combines crowdsourcing with LLMs, creating 1.33 million knowledge tuples. However, it introduces inaccuracies, such as merging part and material information under `MadeUpOf` and misclassifying part-related concepts under `HasProperty`, e.g., (bicycle, `HasProperty`, two wheels). ASCENT++ [24] refines web-based open knowledge extraction through clustering techniques and the use of pre-trained language models, which help group semantically similar assertions and convert free-form predicates into a structured format. Yet, it can produce bizarre errors such as (crane, `MadeOf`, many different colors) and (desk, `HasA`, multilingual staff), reflecting persistent issues of noisy outputs in text mining. Meanwhile, Hansen and Hebart [25] employ GPT-3 to generate semantic features for 1,854 objects, producing detailed outputs (e.g., 'scissors have two blades, handles, and a pivot point'). However, material descriptions about parts remain imperfect (e.g., 'scissors are made of metal', failing to mention possible plastic handles), and in some cases, the model generated incoherent outputs, abruptly shifting topics mid-sentence. These examples demonstrate LLMs' potential for structuring knowledge while highlighting issues with precision and noisy outputs. In contrast, our work focuses specifically on extracting and structuring subtype and part–material relations, providing a resource that represents composition information more explicitly than prior feature-based approaches.

3. Method

In this section, we describe the methodologies used to identify subtypes, parts, and materials of various entities. Our focus is on artifacts, as their part structure and materials are typically more heterogeneous and distinct than in natural objects. To begin, we construct a list of common physical objects. Starting with Wikidata [26] entities classified as artificial physical objects or structures, we apply keyword-based filters (e.g., excluding 'law', 'protein') and subclass-based filters (e.g., excluding 'concept', 'organism') to remove abstract entities. We then retain only entries that exist in WordNet and have links to Wikipedia. Finally, a four-stage prompting process using GPT-4 [27] is employed to select common, physical, standalone objects that are countable, resulting in a list of 2,313 unique entities. We then apply few-shot in-context learning and multi-step zero-shot learning, both leveraging GPT-4 Turbo (`gpt-4-1106-preview`) [27], to this entity list. These techniques are used independently, to generate separate datasets for outcome comparison and wide coverage in object classification. Full filtering details and prompts, along with examples of our results, are provided in Supplementary Materials A–C.

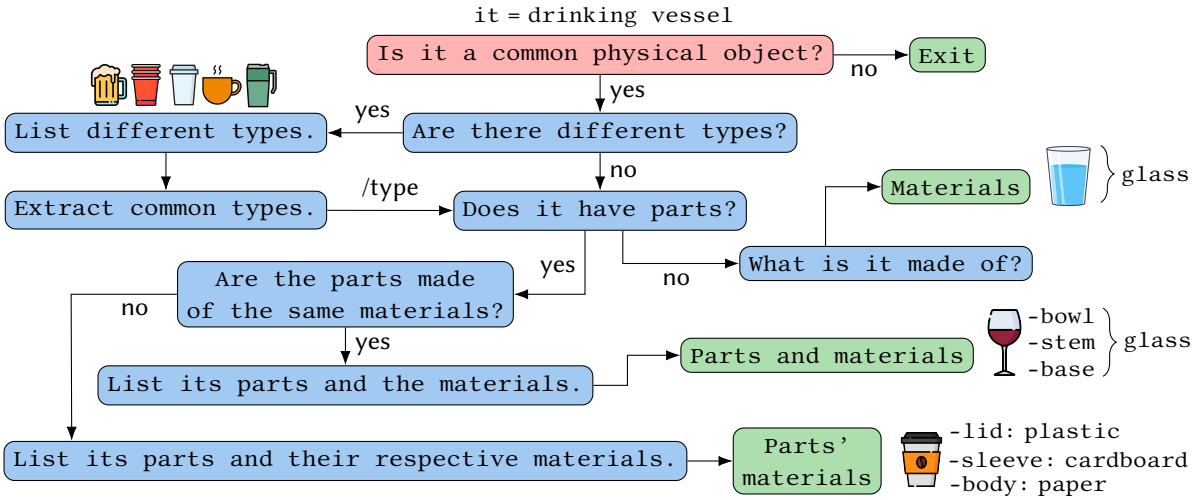

Figure 1: Illustration of our zero-shot prompting using a multi-step classification algorithm to acquire subtype, part, and material information from GPT-4 Turbo.

3.1. Multi-step zero-shot prompting

Our zero-shot prompting method employs a sequence of sub-questions to identify and categorize subtypes, parts, and materials. Although each sub-question is posed in zero-shot form, the multi-step sequence is designed to guide the LLM through complex classification tasks in a structured and incremental manner, particularly when handling multifaceted object classifications step-by-step. Figure 1 provides an overview of the multi-step zero-shot classification algorithm, illustrating how each step builds on the previous one to refine the categorization process. The process begins with the list of common physical objects identified in Section 3, followed by steps to classify their subtypes and constituent elements: (1) Determine if the entity has meaningful subtypes. (2.1) If subtypes exist, list and prioritize physically distinct variations while filtering out superficial design differences. (2.2) Identify widely recognized subtypes. (3) Assess whether the object consists of distinguishable parts. (4) If the entity has parts, check if the parts share the same material composition. (5.a) If no parts exist, list the typical materials it is made from. (5.b) If all parts share the same material composition, provide a list of the object's parts and their material composition at an overall level. (5.c) Otherwise, specify materials used for each part in detail.

3.2. Two-stage few-shot learning

We also employ a two-stage few-shot learning method to classify subtypes of entities based on their essential parts and identify the materials commonly associated with them. Each stage utilizes five in-context examples (handpicked from the human-annotated data from Section 4.2). An example of the prompts with one in-context demonstration and its corresponding output is shown in Figure 2.

In the first stage, the model is prompted to identify common subtypes of a given entity and to enumerate their constituent parts for each subtype. Subtype classification is strictly based on the presence or absence of unique, essential parts, while variations in size, shape, material, or function are explicitly excluded. If an entity has no subtypes, the model is instructed to state this explicitly and list only the parts of the base entity. The examples used in this stage were selected to represent diverse structural patterns, including cases with subsubtypes, entities lacking distinct parts or subtypes, cases involving optional components, and those comprising only essential parts. The second stage focuses on material extraction. Here, the model is prompted to identify typical materials used for each part identified in the first stage, while excluding any that are primarily used for joining, stitching, or finishing. The model is guided to use specific conjunctions (e.g., 'and,' 'or,' 'and/or') to reflect co-occurrence, mutual exclusivity, or interchangeability of materials. The in-context examples were chosen to reflect

(a) Prompt for subtypes and parts

Entity5: Frying pan
1. Stovetop frying pan: body, handle
2. Electric frying pan: body, handle, legs, lid (optional), lid knob (optional), thermostat

Entity6: Scraper (kitchen)

---

1. Flat scraper: handle, flat edge
*2. **Bowl scraper**: handle, curved edge*
3. Grill scraper: handle, sturdy flat edge, notches (optional)

(b) Prompt for materials

Entity 5: Facial tissue
Parts: none
Materials: absorbent paper

Entity 6: **Bowl scraper**
Parts: handle, curved edge
Materials:

---

1. handle : wood or plastic
2. curved edge : plastic or silicone

Figure 2: A simplified example of our two-stage few-shot prompts, with the input to the LLM above the line and the output below. (a) illustrates subtypes and parts identification, listing a bowl scraper as a subtype of a scraper. (b) focuses on material identification, listing the materials used in bowl scraper's parts extracted from (a).

variation in conjunction usage and to illustrate differences between unique and repeating materials.

4. Results & Analysis

The acquired datasets are based on 2,313 Wikipedia entries from Section 3, which serves as the foundation for the few-shot and zero-shot settings. Each entity becomes the root of a hierarchy of subtypes and subsubtypes, where applicable. The most specific level in the hierarchy—a subsubtype (if available), a subtype (if no subsubtype exists), or the entry itself (if neither applies)—serves as the primary reference for part and material information. Each classification at this level includes material details, and if the classification has distinct parts, those are also documented. In the zero-shot setting, entities that include subtypes exhibit an average of approximately 5.5 subtypes and 13.75 subsubtypes per entity. In the few-shot setting, among entities that include subtypes, there is an average of 4.41 subtypes and per entity, but only 84 distinct subsubtypes are observed in total. The few-shot data include over 6,000 instances of these most specific classifications, while the zero-shot data cover nearly 27,300 instances. Both datasets feature thousands of subtypes, though the zero-shot dataset is more extensive. Most classifications have multiple associated parts and materials, with those in the zero-shot dataset averaging around eight parts and two materials, while those in the few-shot dataset tend to have slightly fewer.

4.1. Intrinsic evaluation

We evaluate the quality of our dataset with the help of Amazon Mechanical Turk (MTurk), using two metrics: precision and recall. The evaluations focus on three categories: subtype, part, and material. We sample 120 entities from the list of common physical objects obtained in Section 3 and apply a filtering process to limit evaluated items (i.e., evaluation instances) to 1,000 in most categories. For each multiple-choice question, we collected responses from exactly three distinct MTurk workers.

4.1.1. Precision

For precision, we evaluate whether the predicted relationship (subtype, part, or material) for a given object is likely to be correct. Workers are asked to evaluate each item by selecting from multiple-choice options, including *Likely*, *Unlikely*, *Unable to answer*, and *Uncertain*. A response is marked *Likely* when the predicted relationship is considered plausible. *Unlikely* responses reflect different criteria depending on the relationship type: for subtype predictions, the predicted item may instead represent a part of the target object or an unrelated concept; for parts, the item may not be a typical or essential component; for materials, the predicted term may rarely be used in making the object or be rather considered a part than a material. *Unable to answer* applies when terms are unfamiliar or unclear, and *Uncertain* is used when workers are unsure.

Table 1
Precision for the subtypes, parts, and materials in the few-shot and zero-shot data. The table shows the percentage of responses marked as *Likely*, *Unlikely*, *Unable to answer*, and *Uncertain* for each category.

Response category	Subtype		Part		Material	
	Few	Zero	Few	Zero	Few	Zero
Likely	84.96	**91.63**	**84.79**	79.90	88.27	**88.77**
Unlikely	2.09	3.32	6.20	9.37	3.42	3.53
Unable to answer	12.02	4.64	8.90	10.47	8.30	7.67
Uncertain	0.93	0.41	0.11	0.27	0.00	0.03

Table 2
Distribution of reasons for *Unlikely* answers in precision evaluation of 'part' items. The table presents the proportion of responses for each reason across the few-shot and zero-shot data.

***Unlikely* answer detail**	**Few**	**Zero**
[Object] does not have any parts	0.00	1.07
[Part] is rather a feature than a part	47.67	44.13
[Part] is rather a material used in [object]	13.95	6.41
[Part] is not essential or not included with/attached to [object]	30.81	32.03
[Part] is irrelevant to [object]	4.65	11.03
Other reasons	2.91	5.34

The precision results in Table 1 show that most items were marked as *Likely* across all three categories. In particular, the zero-shot data received a higher portion of *Likely* responses than the few-shot data, especially for subtypes. The few-shot data resulted in 12.02% of subtype responses being marked as *Unable to answer*, all due to workers' unfamiliarity with specific items (e.g., 'fascinator' as a subtype of headgear). While the Part category shows a slightly higher rate of *Unlikely* and *Unable to answer* responses compared to other categories, these elevated rates remain relatively low overall. They appear to stem from specific challenges involved in identifying what constitutes a "part" as opposed to a material, feature, or accessory. This pattern suggests that identifying essential parts may be more challenging for both LLMs and human evaluators, possibly due to greater ambiguity and the increased complexity involved in reasoning about parts. Interestingly, all categories had a very low percentage of responses marked as *Uncertain* (less than 1%), for both few-shot and zero-shot data. This suggests that when workers feel they are familiar with a type of object, they also feel confident about their conceptual understanding of it.

To better understand the reasoning behind the significant portion of *Unlikely* responses for the part items, we further analyzed these responses. Table 2 provides a detailed breakdown of the specific reasons why certain parts were deemed unlikely, along with the respective proportions for each reason. The most dominant reason (accounting for 47.67% and 44.13% of *Unlikely* responses in the few-shot and zero-shot data, respectively) was that the workers identified the listed parts as a feature rather than a discrete part. For example, the predicted part 'neck opening' on a wetsuit, which is an absence of material forming an access point, and 'bank name' on a traveler's cheque, which is an informational label rather than a tangible component, were categorized in this way. Similarly, 'cloth' in a cloth coffee filter was judged unlikely because it refers to the material the object is made of, rather than a separate component. Annotators also noted that some items that are frequently associated with an object, such as the 'toolboxes' on a flatbed tow truck or the 'tripod mount' on a mirrorless digital camera, are not essential and thus not considered definitional parts. In other cases, predicted parts were deemed irrelevant to the object, such as 'headband' for a tennis net or 'engine room' for a ship blueprint. Taken together, these patterns indicate that worker disagreement with LLM-derived parts is often rooted in inherent fuzziness of what constitutes a part rather than outright misconceptions.

Table 3
Recall for the lists of subtypes, parts, and materials in the few-shot and zero-shot data.

Response category	Subtype		Part		Material	
	Few	Zero	Few	Zero	Few	Zero
Likely with no issues	**49.54**	44.86[2]	58.17	**74.63**	71.43	**73.67**
Likely, even though some items overlap	5.56	13.58	0.89	2.47	0.62	1.13
Unlikely with one major item missing	19.44	20.99	21.12	8.17	15.65	12.90
Unlikely with two major items missing	5.56	9.47	5.02	2.10	3.19	2.90
Unlikely with three or more major items missing	8.80	6.17	2.67	1.57	1.23	0.70
Unlikely for other reasons	0.93	0.41	0.08	0.40	0.38	0.13
Unable to answer b/c unfamiliarity or poor naming	10.19	4.53	11.97	10.67	7.42	8.50
Uncertain	0.00	0.00	0.08	0.00	0.08	0.07
Likely with no issues / some items overlap	55.10	**58.44**	59.06	**77.10**	72.05	**74.80**
Likely / Unlikely with one major item missing	74.54	**79.43**	80.18	**85.27**	**87.70**	**87.70**

4.1.2. Recall

In a manner similar to the evaluation of precision, we assess whether the provided lists for subtypes, parts, or materials of an object are complete. The summarized results of the recall evaluation are presented in Table 3. Overall, the majority of responses fall into the *Likely with no issues* category, especially for part and material identification in the zero-shot setting (74.63% and 73.67%, respectively), reflecting strong performance even without direct supervision.

To contextualize these results, we briefly illustrate what constitutes a complete versus incomplete list across the three list types. For subtypes, a complete list for the entity 'diving suit' in the few-shot setting includes *wetsuit, dry suit, dive skin, semi-dry suit, and hot water diving suit*. In contrast, the zero-shot output for 'electrical device' lists 96 subsubtypes (e.g., *power transformer, window air conditioner*) under 19 subtypes (e.g., *refrigerator, solar panel*), yet it is still judged to be missing at least three major subtypes. Upon review, we found that defining what qualifies as an 'electrical device' is inherently ambiguous. For example, are digital bathroom scales or battery-augmented bicycles electrical devices? The line between an object that *uses* electricity and one that *is* an electrical device depends heavily on context and interpretation, and it is unclear whether consistent criteria were applied by the workers.

For parts, a comprehensive list for 'ring-bound notebook' includes *cover, rings, paper, ring mechanism, spine, divider tabs* and *pocket folder*. While part lists evaluated as missing three or more major components by all three annotators are rare, one example is 'rescue squad' (a specialized truck that can be a converted fire engine) for which only *chassis, engine, compartments*, and *specialized rescue tools* were identified. This judgment raises the question of whether 'rescue squad' fire engines are a subtype with a consistent component structure. Although related terms like 'squad truck' appear in Wikipedia, their configurations vary widely; some include firefighting equipment like hoses, while others do not, depending on departmental roles. This variability suggests that 'rescue squad' may be too ambiguous or ill-defined to support firm expectations, making the omission of certain parts less clearly erroneous.

For materials, a complete list for 'handle' of an insect net includes *wood* or *plastic*; another complete list for 'sweatband' of a bucket hat comprises *cotton, polyester, elastic*, and/or *nylon*. No material lists were judged by all three annotators as missing three or more major items; the only such case flagged by two annotators was the list *steel or titanium* for the blade or head of a melee weapon.

Combining the categories *Likely* and *Likely with overlap* leads to a substantial increase in overall recall across all list types. For instance, in the few-shot setting, this combined measure reaches 55.10% for subtypes, 59.06% for parts, and 72.05% for materials. Examples of responses categorized as *Likely,*

[2]It is important to keep in mind that recall can vary considerably depending on how the task is framed, who is doing the evaluation, and which objects are being considered. As we show in Section 4.2.2 below, an alternative evaluation setup yields much higher values. This highlights that seemingly low recall in one context may not necessarily indicate poor performance, but rather reflect broader variability in human and model interpretations of completeness.

Table 4
Inter-rater reliability (Gwet's AC1) for intrinsic evaluation tasks under grouped and ungrouped responses.

Response granularity	Precision			Recall		
	Subtype	Part	Material	Subtype	Part	Material
Grouped	0.87	0.77	0.83	0.35	0.55	0.53
Ungrouped	0.87	0.77	0.83	0.23	0.53	0.54

even though some items overlap include cases where 'director's chair style camping chair' under 'portable folding camping chair' overlaps with 'folding director's chair with side table' listed under 'portable director's style camping chair.' Another such example for subtypes is the classification of 'baseball cap' under both 'hat' and 'cap' (as subtypes of the entity 'headgear'). For materials, an overlap occurs when a set such as *steel, aluminum, stainless steel* or *plastic* is listed for the side shelves of a natural gas grill.

Expanding the scope to include *Unlikely with one major item missing* further boosts recall significantly, with few-shot totals reaching 74.54% (subtype), 80.18% (part), and 87.70% (material). In most cases, LLMs retrieve the majority of critical elements with high reliability, even if one key item is omitted. For example, the few-shot subtype list for 'tong' (*kitchen, ice, forge, salad, sugar, crucible tongs*) was judged by two of three annotators to still lack one subtype. Similarly, the zero-shot part list for 'soft serve ice cream vending machine' (*freezer, dispensing head, mixing hopper, control panel, internal mechanism, cooling system, drip tray, nozzle*) was judged by all annotators to miss one major part. For materials, the few-shot output for the 'wheels' of a refrigerated trailer (*rubber, metal*) was flagged by two annotators as missing a third key material. These near-complete lists, though not perfect, indicate that the model typically identifies most salient items, despite some omissions. Note that 'one major item missing' never refers to a list of only one item judged to lack another; in dataset construction, subtypes and parts required at least two instances to be included (unless absent altogether). Cases with exactly two present and a third absent were rare (5.2% of subtype items and 2.5% parts), and most omissions occurred in longer lists. Materials allowed singletons, and about 6% had only one material judged as missing another. Additionally, 10–12% of the evaluated data were marked as *Unable to answer*, due to unclear item names or insufficient annotator knowledge. When restricted to only the subset of evaluable/answerable cases, the proportion of outputs falling into the top three categories becomes even more pronounced. For example, for the zero-shot part lists, this proportion rises from 85.27% to 95.46%.

Interestingly, performance patterns varied by item category. For part lists, the few-shot method exhibited more frequent omissions of one, two, or even three or more major items, compared to zero-shot. In contrast, subtype lists showed the opposite trend, with the few-shot condition yielding more complete lists than zero-shot. This asymmetry suggests that the prompting strategy's effectiveness can depend strongly on the nature of the target information. Material list generation consistently showed the highest overall recall values across both conditions, with fewer issues related to completeness or overlap. This may be due to the typically small and constrained set of materials associated with objects or parts, which simplifies the generation task.

While the recall values may appear modest (ranging from 44.86% to 74.63% for the *Likely with no issues* response), they should not be interpreted as outright failures, as this is a byproduct of the indeterminacy of object concepts and hierarchies and their expression in language. Seemingly missing subtypes, parts, or materials may not reflect factual deficits, but instead reflect differences in the interpretations of object-denoting phrases and the conceptions of the type and parts hierarchies associated with them.

4.1.3. Inter-rater reliability

To measure consistency among MTurk raters, we used Gwet's AC1 inter-rater reliability coefficients [28] at two response granularities: grouped and ungrouped. The grouped approach categorizes similar responses together, while the ungrouped condition treats each option independently. For example, in the grouped condition for recall, responses *Likely* and *Likely with overlap* are classified as 'Likely,' while all variants of *Unlikely* responses are grouped together. Table 4 summarizes agreement scores across

intrinsic evaluation tasks and categories. Precision agreement scores are generally higher than recall; while MTurk workers tend to agree on correct predictions, they face greater difficulty in determining whether subtype, part, or material lists are complete. This difficulty likely reflects, once again, the inherent indeterminacy of object concepts and their hierarchies, as well as the variability in their linguistic expressions. The greatest discrepancy between grouped and ungrouped responses appears in subtype recall (0.35 vs. 0.23), suggesting frequent disagreement about specific reasons behind judgments.

4.2. Dataset-Comparison evaluation

4.2.1. Human-annotated data creation

We also evaluate the quality of our dataset using comparisons with human annotations. For this, we created a human-annotated dataset capturing part structures and material compositions of 120 entities, randomly sampled from the object list provided in Section 3 without overlapping those used for intrinsic evaluation. Annotations were conducted by in-house lab members and authors following predefined criteria. Annotators primarily referenced Wikipedia but could also consult external sources, including online searches, when needed. Naming conventions for subtypes strictly followed Wikipedia entries when available. Annotators captured detailed part structures, including optional components, and used conjunctions like *and/or* to accurately reflect material combinations. When parts were composed of multiple subcomponents, the label *entity* could be used in place of a specific material. Subtypes were annotated up to two levels (subtypes and subsubtypes) when the distinctions were meaningful in terms of part structure or material composition. This depth was chosen to align with the level of granularity used in the few-shot and zero-shot methods. Additionally, uniform materials spanning multiple parts were annotated as such when applicable. Material information was categorized as explicit, indirect, or inferred, and part annotations not documented in Wikipedia were flagged with an asterisk. The annotation guidelines are provided in the supplementary materials.

4.2.2. Comparison with human-annotated data

To assess the quality and the effectiveness of the data generated by the few-shot and zero-shot extraction methods, we designed an evaluation process that compares their results with the above human-annotated data. The evaluation was also conducted on MTurk, with three different workers providing responses for each question, and a total of five workers participating overall. Due to budget constraints, we evaluated a subset of 60 entities out of 120 human-annotated entities. Each of these 60 entities appeared across three datasets, resulting in a total of 180 classifications. This subset was sufficient to provide reliable quality estimates for our purposes. To avoid bias, none of these entities was included in the few-shot prompts. For consistency, the same worker evaluated all questions related to a specific entity across the three datasets, with the order of entities randomized to prevent bias from consecutive judgments.

The evaluation questions focused on key aspects such as familiarity with subtypes, coverage comprehensiveness, level of detail, clarity and distinction, consistency in style, and focus on essential parts. Workers selected from multiple-choice options that best described each evaluation criterion, such as whether a result was comprehensive or missing major subtypes. These selections were then mapped to numerical scores ranging from -1 to $+1$, where higher values indicate better quality and lower values reflect deficiencies. The evaluation questionnaire is included in the supplementary materials.

Table 5 provides insights into the strengths and limitations of the three datasets, across various criteria. Few-shot data demonstrates a higher mean score (88.39) compared human-annotated data (85.08), exhibiting a balanced level of detail and offering clearer, more distinct classifications. Unlike zero-shot outputs, which often include overly specific or redundant details, the few-shot method avoids overgeneralization and excessive granularity. It also captures subtle elements that human annotators might overlook. Specifically, human-annotations struggled with clarity and distinction, such as omitting essential components (e.g., a lid or locking mechanism on a ballot box). These inconsistencies point to the subjective interpretation and variability inherent in manual annotation, whereas the few-shot method proves more thorough and consistent.

Table 5
Comparison of few-shot, zero-shot, and human-annotated datasets across multiple evaluation criteria. The 'Total' column represents the number of rated responses for each evaluation criterion, while the 'Mean' column reflects the average score, calculated as the ratio of the aggregate score to the total number of responses. The scores are derived based on predefined rating criteria.

Category	Evaluation criteria	Few-shot		Zero-shot		Human-ann.	
		Total	Mean	Total	Mean	Total	Mean
Subtype	Familiarity	84	62.50	117	60.68	46	**67.39**
	Coverage	91	51.10	113	**82.74**	57	21.05
	Level of detail	81	**92.59**	116	70.69	45	86.67
	Clarity & distinction	81	95.06	116	75.86	45	**100.0**
	Consistency in detail or style	82	95.12	116	94.83	45	**100.0**
Part	Focus on essential components	172	94.19	175	86.29	141	**97.16**
	Level of detail	172	91.86	175	**95.43**	143	83.22
	Clarity & distinction	179	**94.41**	178	92.13	178	75.28
	Consistency in level of detail	172	**100.0**	175	94.29	144	95.83
Material	Clarity & distinction	178	85.39	177	74.01	175	**95.43**
Total		1,292	**88.39**	1,458	83.85	1,019	85.08

In contrast, the zero-shot data achieve dramatically higher subtype coverage (82.74) than human-annotated data (21.05) and few-shot data (51.10).[3] While human annotators relied on external resources to identify answers as comprehensively as possible, the few-shot and zero-shot setups simply involved prompting the model to list as many relevant items as possible, without such aids. Despite these constraints, both LLM-based approaches, especially zero-shot, demonstrated superior coverage, highlighting the expansive capabilities of LLMs even under limited-context conditions. However, this strength in coverage comes with trade-offs. The zero-shot approach often suffers from decreased clarity and focus on essential components, producing classifications that are overly detailed or redundant. In contrast, few-shot learning strikes a practical balance between precision and completeness, bridging gaps left by human annotations and zero-shot learning.

4.3. External recall on part/material

To evaluate our dataset's coverage relative to existing datasets (external recall), we compare parts and materials information against five reference datasets: ParRoT, CSLB, McRae, WordNet, and ConceptNet. Instead of crowdsourcing, one of the authors conducted this evaluation.[4] We randomly sample 20 objects common to our dataset and each reference set. Objects missing both part and material information in either dataset are replaced (e.g., 1–5 objects in McRae and WordNet, 48 in ConceptNet). Duplicate items (e.g., *power supply cord* and *power cord* of a hair dryer), design or functional features (e.g., *armholes* in a jacket, *grip* of boots), and unattached objects (e.g., *pilot* in an airplane, *bucket* used with a mop) are removed to avoid miscounts. For ConceptNet, we exclude WordNet-imported entries and low-quality data (e.g., *accent on second syllable* as a part of an umbrella). In WordNet, parts are identified via meronym relations and glosses.

For scoring, we assign full (1 point), half (0.5 points), and no credit (0 points). **Full credit** applies to i) synonymous or contextually equivalent matches, such as *lens of spectacles* and *lenses of glasses*; ii)

[3]Note that the coverage metric used here is conceptually similar to recall discussed in Section 4.1.2, but their values differ considerably. While the coverage metric is based on mapped numerical scores and thus differs methodologically from the recall scores, the comparison remains informative. Recall ranged from 49.54% to 55.10% for few-shot settings and from 44.86% to 58.44% for zero-shot. The discrepancy between these lower recall values and the mean coverage (51.10 for few-shot and 82.74 for zero-shot) suggest that completeness assessments are more sensitive to variations in human judgment and in the data being assessed. This shows the complexity of evaluating recall or coverage in subjective tasks.

[4]The judgments of approximate synonymy, subtype relations and relative granularity needed for these comparisons, as laid out below, would have been very difficult for crowd workers.

Table 6
External recall values for our datasets, calculated based on part and material availability in external datasets. A dash ('-') indicates that the ParRoT dataset did not contain material information.

External data	Part			Material		
	Total items	Few-shot	Zero-shot	Total items	Few-shot	Zero-shot
ParRoT	131–171	56.14–57.54	87.72–89.03	-	-	-
CSLB	94–106	65.57–65.96	**93.79–94.34**	70	86.43	97.14
McRae	24–25	66.67–68.00	93.75–94.00	30	95.00	96.67
WordNet	51–63	57.19–58.73	85.46–85.71	14	82.14	**100.0**
ConceptNet	36	67.47–70.83	93.28–94.44	14	**100.0**	**100.0**

Table 7
Inverse recall values for external datasets, calculated based on parts available in our datasets. A dash ('-') indicates the dataset contained either part or material information for the initially selected 20 objects.

External data	Few-shot			Zero-shot		
	Total items	Recall	Extended	Total items	Recall	Extended
ParRoT	91-96	78.02–78.13	-	138–162	67.28–69.44	-
CSLB	84–92	64.13–67.86	-	116–142	59.15–60.10	-
McRae	65–68	25.00–26.15	23.61	83–102	**22.06–22.44**	20.83
WordNet	76–81	41.36–43.42	32.84	94–117	40.60–43.09	31.88
ConceptNet	76–82	29.28–29.88	8.00	94–121	25.44–26.03	**7.79**

cases where a target item is a specific instance of a reference item, such as *top* and *dome top* of a bird cage; iii) container–substance relations, such as *propellant of a fire extinguisher* and *pressure cartridge*; and iv) cases where a reference material, which is a supertype of another reference item, is satisfied by a target material, e.g., when *metal, aluminium, and light metal* are listed, we interpret *light metal* as "aluminium and some other metals", allowing a match with *titanium*. **Half credit** applies when the reference item is a specific type of the target item (e.g., *side carry handle* and *handle* of a suitcase; *straw* used in a hut matching *thatch*), or when one dataset uses broader categories than a specific item found in the other dataset, with the specific item serving as a defined component or a focused role (e.g., *seats* of an airplane matches *cabin* of an airplane).

To address granularity variations, we use two counting methods: individual and grouped. Individual counting treats each reference item separately, calculating recall as the total score divided by the number of individual items, each contributing up to one point. Grouped counting consolidates functionally related items (e.g., a digital camera's *dials* and *shutter button* as a control unit; a caravan's *kitchen* and *kitchen equipment* as one group), with recall based on item groups rather than individual items, each contributing up to one point.

In Table 6, we report external recall values, where total reference items vary by counting method. Individual counting results in a higher item count, while grouped counting reduces it, but neither method consistently leads to higher or lower recall. A key finding is that zero-shot recall is consistently higher than few-shot in both part and material categories, indicating that the zero-shot approach captures finer-grained details in the reference datasets. For example, the ParRoT dataset includes detailed chair parts such as *apron, cross rail, top rail*, and *stile*, alongside broader elements like *back, legs*, and *seat*. The zero-shot method retrieves all of these fine-grained parts with only minor discrepancies; e.g., *cross rail* and *apron* are partially grouped under the broader label *frame*. In contrast, the few-shot approach generalizes to broader categories such as *legs, seat, backrest*, and *armrests*, often omitting finer distinctions. These findings suggest that zero-shot methods may be preferable for applications requiring fine-grained part recognition, while few-shot methods are better suited for high-level categorization.

We also compute **inverse recall**, measuring how well external datasets capture our part information.[5]

[5]Only part information is evaluated due to significant discrepancies in material counts and granularity between the reference and external datasets.

In this evaluation, subtype-specific part items in our datasets were excluded from evaluation, as external datasets lack subtype information. For example, gas tanks in barbecue grills (specific to gas models) and child seats in shopping carts were omitted as they are not general features. Inverse recall evaluation follows the same object set used for external recall, but additionally includes objects that were previously excluded from the external recall, contributing to the 'Extended' column in Table 7. Unlike other columns, 'Extended' reflects only the lowest inverse recall for each external dataset, rather than a range, as the additional objects are counted individually to provide a conservative underrepresentation estimate.

As shown in Table 7, inverse recall is consistently lower than external recall, except for ParRoT and CSLB when measured on the few-shot items; however, their scores still remain below their external recall scores in the zero-shot setting. This confirms that our dataset contains more detailed part information than external datasets. McRae, WordNet, and ConceptNet show particularly low inverse recall, with McRae dropping to 22.06% in zero-shot. The 'Extended' column highlights major gaps, for ConceptNet in particular, which lacks coverage for many objects. Interestingly, few-shot inverse recall consistently exceeds zero-shot, which suggests that while the zero-shot method excels at capturing compositional details, the few-shot method better aligns with the structured knowledge from existing datasets.

5. Conclusion

Our study developed protocols for using LLMs to extract structured knowledge on artifact subtypes, parts, and materials, integrating computational and cognitive perspectives. Our findings show LLMs' potential for mining everyday knowledge, and reveal subtleties and complexities in the conception of physical objects as reflected in the LLM-derived knowledge and the human evaluation of that knowledge. Specifically, we contribute: (i) empirical evidence that LLMs generate artifact knowledge that aligns with or surpasses layman cognition, while also exhibiting gaps and inconsistencies that reflect broader issues in representations of commonsense knowledge; (ii) insights into how few-shot and zero-shot prompting influences the specificity and breadth of extracted knowledge; and (iii) a structured dataset of extracted knowledge as a resource for the NLP/AI community, benchmarked against human annotations and external datasets to highlight strengths and conceptual ambiguities in knowledge representation.

While our work demonstrates promising results in mining and structuring part and material knowledge from LLMs, it is not without limitations. A key challenge lies in balancing specificity and generality when choosing between few-shot and zero-shot prompting. Hybrid strategies, such as multi-step prompts with in-context learning, could improve coherence by preserving hierarchical consistency of generated knowledge and reducing redundancy, as each step could account for prior outputs.

Another complication arises from the inherent subjectivity in how people conceptualize the structure and attributes of objects, making it difficult to achieve a unified, objective representation. For example, a box kite might be described as consisting of *spars*, *sails*, and *bridle*, or alternatively as *a rigid framework* with *covers*, *bridle*, and *tethering line*. Similarly, an iced tea spoon could be seen as having distinct parts (e.g., a bowl and a handle) or merely as a singular, inseparable object. Likewise, whether a pamphlet should be divided into *front* and *back* (as nearly all physical objects could be) depends on subjective framing. This variability makes it difficult to define a universally "correct" representation and complicates both knowledge generation evaluation.

A promising direction for future work involves evaluating whether specific parts reliably distinguish between subtypes. For instance, if a part appears in one subtype but not another, we might ask whether its presence is a meaningful differentiator. Addressing this gets complicated by semantically similar terms (e.g., holder's name vs. cardholder name in payment cards; dump body vs. dump box in dump trucks), variation in the part granularity of description (e.g., net ties, Velcro strips vs. cable, wire), and ambiguous inclusions where one subtype lists a broad part while another lists a more specific version (e.g., *wheel* vs. *wheel hub*). Notwithstanding these challenges, our study has demonstrated that LLMs, when queried effectively, can provide commonsense object knowledge that matches human accuracy and greatly exceeds the typical scope of individual expertise. We have made this concrete by providing a wide-ranging knowledge set derived from our study, which can serve as a resource for further research.

Declaration on Generative AI

During the preparation of this work, the author(s) used ChatGPT solely to improve the conciseness of the abstract. No content was generated or altered beyond language refinement.

References

[1] B. Bhavya, J. Xiong, C. Zhai, Analogy generation by prompting large language models: A case study of InstructGPT, in: S. Shaikh, T. Ferreira, A. Stent (Eds.), Proceedings of the 15th International Conference on Natural Language Generation, Association for Computational Linguistics, Waterville, Maine, USA and virtual meeting, 2022, pp. 298–312. URL: https://aclanthology.org/2022.inlg-main.25. doi:10.18653/v1/2022.inlg-main.25.

[2] S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg, H. Nori, H. Palangi, M. T. Ribeiro, Y. Zhang, Sparks of artificial general intelligence: Early experiments with GPT-4, Computing Research Repository arXiv:2303.12712 (2023). URL: https://arxiv.org/abs/2303.12712.

[3] S. R. Moghaddam, C. J. Honey, Boosting theory-of-mind performance in large language models via prompting, Computing Research Repository arXiv:2304.11490 (2023). URL: https://arxiv.org/abs/2304.11490.

[4] B. He, M. Xia, X. Yu, P. Jian, H. Meng, Z. Chen, An educational robot system of visual question answering for preschoolers, in: Proceedings of the 2nd International Conference on Robotics and Automation Engineering (ICRAE 2017), 2017, pp. 441–445. URL: https://ieeexplore.ieee.org/document/8291426.

[5] P. Zuidberg Dos Martires, N. Kumar, A. Persson, A. Loutfi, L. De Raedt, Symbolic learning and reasoning with noisy data for probabilistic anchoring, Frontiers in Robotics and AI 7 (2020). URL: https://www.frontiersin.org/journals/robotics-and-ai/articles/10.3389/frobt.2020.00100. doi:10.3389/frobt.2020.00100.

[6] Y. Fang, K. Kuan, J. Lin, C. Tan, V. Chandrasekhar, Object detection meets knowledge graphs, in: Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17, 2017, pp. 1661–1667. URL: https://doi.org/10.24963/ijcai.2017/230. doi:10.24963/ijcai.2017/230.

[7] P. Cui, S. Liu, W. Zhu, General knowledge embedded image representation learning, Trans. Multi. 20 (2018) 198–207. URL: https://doi.org/10.1109/TMM.2017.2724843. doi:10.1109/TMM.2017.2724843.

[8] V. Shevchenko, D. Teney, A. Dick, A. van den Hengel, Reasoning over vision and language: Exploring the benefits of supplemental knowledge, in: Proceedings of the Third Workshop on Beyond Vision and LANguage: inTEgrating Real-world kNowledge (LANTERN), Association for Computational Linguistics, Kyiv, Ukraine, 2021, pp. 1–18. URL: https://aclanthology.org/2021.lantern-1.1.

[9] J. Yuan, A. Le-Tuan, M. Nguyen-Duc, T.-K. Tran, M. Hauswirth, D. Le-Phuoc, VisionKG: Unleashing the power of visual datasets via knowledge graph, Computing Research Repository arXiv:2309.13610 (2024). URL: https://arxiv.org/abs/2309.13610.

[10] E. J. Barezi, P. Kordjamshidi, Find the gap: Knowledge base reasoning for visual question answering, Computing Research Repository arXiv:2404.10226 (2024). URL: https://arxiv.org/abs/2404.10226.

[11] M. Berland, E. Charniak, Finding parts in very large corpora, in: Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, College Park, Maryland, USA, 1999, pp. 57–64. URL: https://aclanthology.org/P99-1008. doi:10.3115/1034678.1034697.

[12] M. Poesio, T. Ishikawa, S. Schulte im Walde, R. Vieira, Acquiring lexical knowledge for anaphora resolution, in: Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02), European Language Resources Association (ELRA), Las Palmas, Canary

Islands - Spain, 2002, pp. 1220–1224. URL: http://www.lrec-conf.org/proceedings/lrec2002/pdf/117.pdf.

[13] R. Girju, A. Badulescu, D. Moldovan, Automatic discovery of part-whole relations, Computational Linguistics 32 (2006) 83–135. URL: https://aclanthology.org/J06-1005. doi:10.1162/coli.2006.32.1.83.

[14] W. R. van Hage, H. Kolb, G. Schreiber, A method for learning part-whole relations, in: I. Cruz, S. Decker, D. Allemang, C. Preist, D. Schwabe, P. Mika, M. Uschold, L. M. Aroyo (Eds.), The Semantic Web - ISWC 2006, Springer Berlin Heidelberg, Berlin, Heidelberg, 2006, pp. 723–735.

[15] M. Poesio, A. Almuhareb, Identifying concept attributes using a classifier, in: Proceedings of the ACL-SIGLEX Workshop on Deep Lexical Acquisition, Association for Computational Linguistics, Ann Arbor, Michigan, 2005, pp. 18–27. URL: https://aclanthology.org/W05-1003.

[16] D. Tesfaye, M. Zock, Automatic extraction of part-whole relations, in: Proceedings of the 9th International Workshop on Natural Language Processing and Cognitive Science (ICEIS 2012) - NLPCS, INSTICC, SciTePress, 2012, pp. 130–139. doi:10.5220/0004113801300139.

[17] R. Speer, J. Chin, C. Havasi, ConceptNet 5.5: an open multilingual graph of general knowledge, in: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, AAAI'17, AAAI Press, 2017, p. 4444–4451.

[18] G. A. Miller, WordNet: a lexical database for english, Commun. ACM 38 (1995) 39–41. URL: https://doi.org/10.1145/219717.219748. doi:10.1145/219717.219748.

[19] Y. Gu, B. Dalvi Mishra, P. Clark, Do language models have coherent mental models of everyday things?, in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 1892–1913. URL: https://aclanthology.org/2023.acl-long.106. doi:10.18653/v1/2023.acl-long.106.

[20] K. McRae, G. S. Cree, M. S. Seidenberg, C. Mcnorgan, Semantic feature production norms for a large set of living and nonliving things, Behavior Research Methods 37 (2005) 547–559. URL: https://doi.org/10.3758/BF03192726. doi:10.3758/BF03192726.

[21] B. J. Devereux, L. K. Tyler, J. Geertzen, B. Randall, The centre for speech, language and the brain (CSLB) concept property norms, Behavior Research Methods 46 (2014) 1119–1127. URL: https://doi.org/10.3758/s13428-013-0420-4. doi:10.3758/s13428-013-0420-4.

[22] H. Zhang, D. Khashabi, Y. Song, D. Roth, TransOMCS: From linguistic graphs to commonsense knowledge, in: C. Bessiere (Ed.), Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20, International Joint Conferences on Artificial Intelligence Organization, 2020, pp. 4004–4010. URL: https://doi.org/10.24963/ijcai.2020/554. doi:10.24963/ijcai.2020/554.

[23] J. D. Hwang, C. Bhagavatula, R. Le Bras, J. Da, K. Sakaguchi, A. Bosselut, Y. Choi, COMET-ATOMIC 2020: On symbolic and neural commonsense knowledge graphs, in: the AAAI Conference on Artificial Intelligence, volume 35, 2021, pp. 6384–6392.

[24] T.-P. Nguyen, S. Razniewski, J. Romero, G. Weikum, Refined commonsense knowledge from large-scale web contents, IEEE Transactions on Knowledge and Data Engineering (2022). doi:10.1109/TKDE.2022.3206505.

[25] H. Hansen, M. N. Hebart, Semantic features of object concepts generated with GPT-3, in: Proceedings of the 44th Annual Conference of the Cognitive Science Society, volume 44, Cognitive Science Society, 2022, pp. 779–786.

[26] D. Vrandečić, M. Krötzsch, Wikidata: A free collaborative knowledgebase, Commun. ACM 57 (2014) 78–85. URL: https://doi.org/10.1145/2629489. doi:10.1145/2629489.

[27] OpenAI, GPT-4 technical report, Computing Research Repository arXiv:2303.08774 (2023). URL: https://arxiv.org/abs/2303.08774.

[28] K. L. Gwet, Computing inter-rater reliability and its variance in the presence of high agreement, British Journal of Mathematical and Statistical Psychology 61 (2008) 29–48. URL: https://bpspsychub.onlinelibrary.wiley.com/doi/abs/10.1348/000711006X126600. doi:10.1348/000711006X126600.