Can GPT-4 Recover Latent Semantic Relational Information from Word Associations? A Detailed Analysis of Agreement with Human-annotated Semantic Ontologies.

Simon De Deyne[1] **Chunhua Liu**[2] **Lea Frermann**[2]
[1]Complex Human Data Hub, [2]School of Computing and Information Systems
The University of Melbourne
{simon.dedeyne, chunhua.liu1, lea.frermann}@unimelb.edu.au

Abstract

Word associations, i.e., spontaneous responses to a cue word, provide not only a window into the human mental lexicon but have also been shown to be a repository of common-sense knowledge and can underpin efforts in lexicography and the construction of dictionaries. Especially the latter tasks require knowledge about the relations underlying the associations (e.g., Taxonomic vs. Situational); however, to date, there is neither an established ontology of relations nor an effective labelling paradigm. Here, we test GPT-4's ability to infer semantic relations for human-produced word associations. We use four human-labelled data sets of word associations and semantic features, with differing relation inventories and various levels of annotator agreement. We directly prompt GPT-4 with detailed relation definitions without further fine-tuning or training. Our results show that while GPT-4 provided a good account of higher-level classifications (e.g., Taxonomic vs Situational), prompting instructions alone cannot obtain similar performance for detailed classifications (e.g., superordinate, subordinate or coordinate relations) despite high agreement among human annotators. This suggests that latent relations can at least be partially recovered from word associations and highlights ways in which LLMs could be improved and human annotation protocols could adapted to reduce coding ambiguity.

Keywords: Large language models, semantic relations, word associations

1. Introduction

The word association test (WAT) provides important information about the organisation of the mental lexicon. In a typical study, participants are presented with a cue word (e.g., *dog*) and produce the first word(s) that come to mind (e.g., *cat* or *bark*). This procedure is often referred to as *free* word association as participants are not restricted in their responses, making it one of the most general methods to obtain subjective behavioural estimates of word meaning (Deese, 1965).

In recent years, online crowd-sourcing approaches such as the Small World of Words project have demonstrated that this approach is highly scaleable, with several datasets including millions of responses published in Dutch, English, Spanish and Chinese (De Deyne et al., 2019). As such, word associations provide valuable resources for the fields of lexicography and semantic typology, which study the availability and organization of senses and meaning within and across languages.

A common type of analysis of these data involves classifying responses according to a semantic ontology that covers taxonomic (*dog* – *cat*), concept properties (*dog* – *tail*), situational properties (*dog* – *park*) or introspective properties (*dog* – *friend*). This is of interest to cognitive science, where these classifications can shed light on the nature of our mental representation and the time course over which this information becomes available (Fitzpatrick and Thwaites, 2020), (Garrard et al., 2001), metaphor comprehension and analogies (Lu et al., 2022).

Word associations have also shown promise as a tool to derive common sense knowledge (Liu et al., 2021). In this respect, recent work suggests that they could fill the gaps in other lexical knowledge graphs. While word associations do not capture the depth of other approaches (e.g., the number of senses of a word), they do capture frequent senses and measure what aspects of meaning are dominant among a community of speakers. Importantly, word associations are informed not only by our linguistic environment but encode extra-linguistic experiential information as well that is difficult to reveal by only studying how words co-occur in language (Fitzpatrick and Thwaites, 2020). Various supervised and unsupervised approaches to predict associations from text had correspondingly mixed success (Griffiths et al., 2007; Cattle and Ma, 2017; Liu et al., 2022).

Large language models (LLMs) like GPT-4 (Achiam et al., 2023) have shown unprecedented abilities not only to generate naturalistic text but also to support complex data annotation (Gilardi et al., 2023) and annotation of single words or word pairs for comparison with human similarity judgments, induction and lexical ratings (e.g., concreteness) (Han et al., 2024; Marjieh et al., 2023; Trott, 2023).

Here we test the ability of GPT-4 (Achiam et al., 2023), a state-of-the-art LLM, to recover semantic relations for human-produced word associations. This is of interest for three reasons. First, this new generation of models, with the capacity to encode long prompts, does not have the same working memory constraints human annotators have when confronted with extensive fine-grained semantic ontologies. Second, we extend a line of work that assesses the utility of LLMs as cognitive models to the task of semantic relation labelling. Third, from a practical perspective, a model that can automatically predict semantic relations can support the construction or augmentation of lexical or commonsense databases.

A highly influential ontology in the cognitive sciences is the Wu-Barsalou (WB) ontology (Wu and Barsalou, 2009). The WB scheme is hierarchically organised and consists of four major relation types, which we will refer to as Level 1: Taxonomic relations, Entity/concept properties, Situational properties and Introspective properties. More detailed Level 2 distinctions are nested within each relation class (e.g., Subordinate, Coordinate are properties nested under Taxonomic relations). While the WB ontology was initially developed to investigate grounding in semantic representations, it has since been applied broadly across many property listing tasks (PLT, see Bolognesi et al., 2017, for an overview) and was recently adapted to the WAT (Liu et al., 2022; Chen et al., 2024). The scheme has been adapted over the years to suit the needs of individual researchers. However, these changes tend to be minor simplifications of Level 2 distinctions such as grouping Buildings and Location or Subordinates and Individuals (see Bolognesi et al., 2017; Liu et al., 2022).

In contrast to the WAT, the PLT is often assumed to be less ambiguous and consequently easier to annotate because the properties can be phrases (e.g., *dog - is a kind of animal*) that can be easily mapped onto the ontology (Superordinate). However, an inspection of existing semantic feature generation studies suggests that features are often coded by the annotators as a single-word response (e.g., *zebra – horse*), similar to a word association. In the PLT of Vivas et al. (2021), for example, 18% of features consisted of a single words, whereas in Bolognesi et al. (2017), 92% of features consisted of a single word. Consequently, presumed ambiguity is not limited to word association per se but is also highly prevalent in semantic feature norms.

1.1. Current work

This study will use previously annotated datasets for word associations and semantic features. The latter are included as contrast cases that allow us to contextualise our findings, as the semantic relation is often included in the participant response. We focus primarily on the Wu-Barsalou semantic ontology (WB), which is widely used in cognitive psychology and GPT-4 as a state-of-the-art (SOTA) LLM. To the best of our knowledge, we are the first to use GPT-4 for the task of relation labelling, despite its remarkable performance for related tasks where limited context is available (e.g., pairwise similarity judgments). Focusing on a single model provides us with an opportunity to analyse (mis)classification and inconsistencies across the different datasets and levels in the label hierarchy. In sum, we address the following research questions:

- To what degree can latent semantic relations be recovered in SOTA LLMs?
- How does performance vary for broad vs fine-grained relation labels?
- How does the nature of the task (WAT vs PLT) affect the results?
- What are the most common confusions, and to what degree do these reflect limitations of the model or inherent ambiguity due to word association data or existing coding schemes?

2. Methods

We introduce the primary relation ontology, which researchers have adopted for classifying word associations, the datasets, and the LLM that will be used in current work.

2.1. Datasets

The current study includes four recent datasets. Studies were included according to the following criteria: 1) the use of the WB scheme (or a close derivative) for the relation annotation; 2) including a large number of concepts; 3) the availability of English translations in the published data for non-English datasets; and 4) the use of multiple annotators with the inter-rater agreement information included in the original study. All data sets share the same four Level 1 relations (Taxonomic, Entity, Situational, and Introspective) but differ in their Level 2 labels. See more details of the labels along with other dataset statistics in Table 1.

2.1.1. Bolognesi-2017

The PLT dataset in Bolognesi et al. (2017) consists of English concept-feature pairs that were carefully annotated through an ontology that resembles a decision tree. This relation ontology has been updated from the WB ontology to accommodate both concrete and abstract concepts effectively. The resulting dataset had a high inter-rater agreement

Bolognesi-2017: #C = 180, #(C,R) = 1919, #L2 = 20

TAXONOMIC RELATION (T): Synonyms, description and linguistic clues (syn), Antonyms (ant), Superordinates (sup), Subordinates and instances (sub), Coordinate (coor)

ENTITY PROPERTY (E): Perceptual properties (perc), Non-perceptual properties (sys), Components, materials and substances (comp), Larger wholes, thematic larger wholes, and disciplines (whol), Entity behaviors (beh)

SITUATIONAL PROPERTY (S): Objects (obj), Participants (par), Actions (act), Properties of contextual entities (other), Function (fun), Locations, containers, and buildings (loc), Time and events (time)

INTROSPECTIVE PROPERTY (I): Evaluations (eval), Emotions (emo), Contingencies and complex cognitive operations (cont)

Vivas-2022 #C= 400, #(C,R) = 2669, #L1 = 33

TAXONOMIC RELATION (T): Synonym (syn), *Ontological category (ont)*, Superordinate (super), Coordinate (coord), Subordinate (subord)

E: *External component (excomp)*, *Internal component (incomp)*, *External surface property (exsurf)*, *Internal surface property (insurf)*, Substance/Material (mat), *Spatial relation (spat)*, Systemic property (sys), Larger whole (whole), Entity behavior (beh), Abstract entity property (abstr)

SITUATIONAL PROPERTY (S): Person (person), Living thing (living), Object (object), Social organization (socorg), *Social artifact (socart)*, *Building (build)*, Location (loc), Spatial relation (spat), Time (time), Action (action), Event (event), Function (func), Physical state (physt), Social state (socst)

INTROSPECTIVE PROPERTY (I): Affect/emotion (emot), Evaluation (eval), *Representational state (rep)*, *Cognitive operation (cogop)*, Contingency (contin), Negation (neg)

Chen-2024 #C = 505, #(C,R) = 2292, #L2 = 21

TAXONOMIC RELATION (T): Synonym (syn), Superordinate (super), Coordinate (coord), Subordinate (sub), Antonym (ant)

ENTITY PROPERTY (E): Components/Material/Substance (comp), Whole (E-whole), Entity property (prop), Entity behavior (beh), Typical state (state)

SITUATIONAL PROPERTY (S): Function (function), Location/Container/Building (loc), Object (obj), Action (action), Agent (agent), Time/Events (time), Contextual entity property (context), Situational state of target (targetstate)

INTROSPECTIVE PROPERTY (I): Evaluation (eval), Emotion (emo), Contingencies and complex cognitive operations (contin)

Liu-2022 #C = 340, #(C,R) = 476, #L2 = 15

TAXONOMIC RELATION (T): Synonym (syn), Antonym (ant), *Category-Exemplar-Pairs (cat)*, Members-of-same-Category (coord)

ENTITY PROPERTY (E): PartOf (part), Material-MadeOf (mat), *HasProperty (prop)*

SITUATIONAL PROPERTY (S): Time (time), Location (loc), Function (func), *Has-Prerequisite (preq)*, *Result-In (result)*, Action (action), *Thematic (them)*

INTROSPECTIVE PROPERTY (I): Emotion-Evaluation (emo)

Table 1: Summary of datasets. #C denotes the number of unique cues, #(C,R) denotes the number of unique cue-response pairs, #L2 denotes the number of Level 2 relations. The dataset-specific L2 labels are in italics.

with Cohen's κ = .886 for the Level 1 distinctions, and κ = .866 for the Level 2 distinctions.

2.1.2. Vivas-2022

The Vivas et al. (2021) Features PLT dataset consisted of noun-feature pairs collected from Spanish speakers across a range of concrete semantic domains. The reported inter-rater agreement measured as Krippendorff's α was high: .78 for novice coders and .86 for trained coders (Vivas et al., 2021). The ontology closely followed the original WB scheme. In the current analyses, we did not include additional quantifier codes and two codes that were not used by any of the annotators (C-INDIV and S-MANNER). A separate set of Meta-codes (e.g., hesitations, repetition, comments) was also not included in the current results.

A second dataset,*Vivas-2022 Asso*, was derived by extracting a key word (e.g., *zebra*, *music instrument*). This way, additional relational cues such as <is a> were removed, allowing us to define a baseline to determine how these relation indicators reduce ambiguity when annotating PLT data.[1]

2.1.3. Chen-2024

The Chen et al. (2024) WAT data consists of a semantic ontology derived from the WB ontology. The cues and responses were derived from the English Small World of Words project (De Deyne et al., 2019). The stimuli comprised 507 nouns (ranging in concreteness) and their top 5 associative responses. All cue-response pairs were coded by two trained coders for broad (Level 1) and fine-grained (Level 2) distinctions. For this study, we only used the Taxonomic, Entity, Situational and Introspective Level 1 properties (see Table 1 for a list of included Level 2 properties). We did not include form position-based properties since these could also be estimated from word co-occurrence data directly and overlap significantly with semantic properties and also omitted meta codes (e.g., erroneous responses) similar to the approach for the Vivas-2022 dataset. The inter-rater agreement, measured as Cohen's κ, was high, .81, for both Level 1 and Level 2 relations.

2.1.4. Liu-2022

The Word Association Explanation database (WAX) (Liu et al., 2022) includes word associations for a total of 15K different English cue-response pairs. A subset of 520 pairs was annotated with semantic relations. Human coders were recruited through Amazon Mechanical Turk. The ontology represents a simplification of the WB ontology, focusing on the main types across all of the four major Level 1 distinctions. The Level 2 properties also included a few additional relations from ConceptNet (Speer et al., 2017) for event-related associations (e.g., Has-Prequisite, Result-In). The pairwise annotator agreement was moderate, Cohen's κ = 0.42.

Like Chen-2024, we did not include linguistic and form-based responses (e.g., Sound Similarity, Common Phrases). An unspecified category (None-of-the-above) was also removed. Finally, note that Emotion-Evaluation were originally grouped under Concept/entity properties. For reasons of comparability, we decided to move this property to a separate Level 1 Introspective properties section consistent with the other datasets.

2.2. SOTA LLM Model

We used GPT-4 (Achiam et al., 2023) through the OpenAI API and specified model version *gpt-4-0613*. Across all studies, the temperature was set to 0, and no optional system prompts were provided. Cue-response pairs were randomized and split into batches of 100 items before being concatenated to the instruction prompt.

2.2.1. Prompting

All prompts followed the same structure at the start and end but differed in terms of the definitions and examples, which were taken from the original articles. All materials and prompts are available in the original articles and online repository.[2] The default prompt was as follows:

> You will be presented with a list of word pairs consisting of an associated cue and an associated target word separated by ' – '.
>
> You are asked to choose a code with square brackets [] that best describes the semantic relation between the cue and the target word. Each code refers to a specific semantic relation that refers to Taxonomic properties, Concept properties, Situation properties, or Introspective properties.
>
> We will now provide you with a definition and examples for each of these, which you will carefully consider when choosing one of the codes.
>
> {Relation taxonomy with definitions and examples.}
>
> Remember to only choose from the above codes between square brackets. Do not further elaborate on your response. Format your response as follows cue — target: code.
>
> List:
>
> {List of 100 cue association pairs: }

For the Bolognesi-2017, Vivas-2022 Features and Vivas-2022 Association, the first sentence was replaced by "You will be presented with a list of word pairs consisting of a cue and a semantic feature separated by '–'. ". Finally, consistent with the instructions in (Vivas et al., 2021), we added "In these examples, the relation signified by the semantic feature is higlighted by using capitalized letters." after the third sentence ("Each code refers...").

3. Results

3.1. Response preprocessing

All responses were provided in the cue — target: code format consistent with the instructions, which

[1] The Bolognesi-2017 dataset consisted mainly of single words, and so this procedure was not applied.

[2] Materials, instruction prompts with definitions and examples as well as the analysis scripts are available at https://github.com/SimonDeDeyne/lrec2024

means no further manual extraction was required. On a very small number of occasions, erroneous codes (i.e., codes not in the instructions were returned). These were subsequently removed.

3.2. Classification

For each of the datasets, we calculated accuracy, precision, recall, macro-F scores, and Kappa interrater reliability at both Level 1 (broad) and Level 2 (detailed). Results are presented for a cue-response type-based classification and a token-based classification, where the latter is weighted by the number of times participants generated a particular response. This provides information that is more useful for real-world settings where only a subset of cue-response pairs might be inspected, which means that accurate relation labels are especially important for the most frequent responses.

Since the response classes (i.e., relation labels) are unbalanced, classification metrics were weighted by prevalence (between 0 and 1) before averaging over classes. With these balanced scores, the role of relatively infrequent classes, such as Introspective properties, which were rare, was proportionate when averaging all four Level 1 classes.

Unlike other datasets, the Chen-2024 included the codes for two individual annotators, A and B. Unless stated otherwise, we also provide the results for the LLM and individual coder agreement.

The results are shown in Table 2 and Table 3. The last three columns show the baseline performance using the majority class (MC), a score where only the majority relation class was considered and which is contrasted with accuracy (see Table 2 and 3). In all cases, the accuracy rate significantly differed from the MC baseline.

3.2.1. Type-based results

The results in Table 2 show high values across all metrics for the feature datasets (Bolognesi et al., 2017; Vivas et al., 2021) at Level 1 and moderate results at Level 2 of the ontology. The results of deriving pseudo-associations after censoring semantic relations from features for the Vivas-2022 Asso dataset had a negligible effect at Level 1 and only a minor drop in performance at Level 2. The results for the two word associations sets (Chen et al., 2024; Liu et al., 2022), were somewhat lower, with good results at Level 1 and moderate to low results at Level 2. The agreement between the LLM predictions and individual coders for Chen-2024 was highly consistent for A and B, with slightly better results for annotator A. However, comparing the scores with those obtained by directly comparing annotators A and B (see Chen AB in Tables2) suggests some room for further improvement when benchmarked against trained human annotators, and this is notably the case for the Level 2 Ontology annotations.

3.2.2. Token-based results

To calculate performance that considers how frequently the responses are generated, weighted results were calculated on the raw data before tabulation. Doing so provides an estimate of classification performance that is more relevant for applications and also allows us to determine whether infrequent responses are inherently more difficult to classify. Consistent with this, Table 3 shows results that are largely consistent with Table 2, albeit slightly higher. The only exception to this pattern was the Liu-2022 dataset, where the difference was less pronounced, which is likely to reflect the relatively small range of frequency given the limited number of cue presentations in this dataset. Similar to the type-based results, LLM prediction performed comparably across annotators in the Chen-2024 dataset but was still lower compared to the results when comparing two trained human annotators. For simplicity, we will only consider the results for Coder A and the remaining analysis.

3.3. Error analysis

Token-based confusion matrices for the Level 1 distinctions are plotted in Figure 1. Each cell encodes the proportion of cross-classifications and supplements Table 3. The main focus is on the entries on off-diagonal elements, which indicate systematic differences between human coders and the model classification. Note that the values do not have to be symmetric. For example, in the Bolognesi dataset, 2% of the responses humans consider introspective were coded taxonomic. Vice versa, only 1% of the responses humans code as taxonomic are labeled as introspective.

Consistent confusion was present in the Bolognesi data for Introspective properties across most other L1 relations. Closer inspection showed that many of the pairs were coded as “Contingencies and complex cognitive operations”. Relative large proportions of these confusions were also found for Taxonomic vs Entity properties (0.06 for Chen-2024). In addition, Taxonomic and Entity properties were also frequently confused in the Liu-2022 dataset (0.09). Insightful examples include *genius – brilliant*, which human annotators code as an entity property, but GPT-4 considers a synonym. This highlights the fact that the model does not capture a human noun-bias, which is typical in association data where words are ambiguous in terms of part of speech. Another example is *lonely – depressed*, which was also considered a synonym but coded as a “Result-In” feature by the annotators. More

	Level 1 Ontology					Level 2 Ontology				
	MC	acc.	κ	prec.	F1	MC	acc.	κ	prec.	F1
Bolognesi	0.388	0.717	0.609	0.765	0.730	0.116	0.542	0.510	0.647	0.564
Vivas Feat	0.438	0.846	0.768	0.867	0.851	0.207	0.614	0.584	0.720	0.618
Vivas Asso	0.401	0.846	0.769	0.852	0.847	0.129	0.599	0.571	0.623	0.577
Chen A	0.419	0.763	0.653	0.769	0.764	0.194	0.523	0.487	0.615	0.536
Chen B	0.419	0.713	0.584	0.738	0.716	0.194	0.492	0.454	0.607	0.499
Liu	0.420	0.718	0.562	0.758	0.727	0.282	0.464	0.399	0.535	0.471
IAA (Chen AB)	0.357	0.880	0.825	0.888	0.879	0.127	0.808	0.794	0.822	0.809

[a] All accuracy vs MC comparisons were significant, $p < .001$.

[b] Recall is identical to accuracy after prevalence weighting.

Table 2: Type-based classification results (acc. = accuracy, MC = Majority Class, κ, prec. = precision, F1) for the Level 1 (left) and 2 (right) ontologies across semantic feature (Bolognesi-2017, Vivas-2022 Feat) and word association (Vivas-2022 Asso, Chen-2024, Liu-2022) datasets. We list agreement with GPT-4 for individual annotators (Chen A and B), alongside inter-annotator scores for annotators A and B of Chen-2024 (IAA ChenAB).

	Level 1 Ontology					Level 2 Ontology				
	MC	acc.	κ	prec.	F1	MC	acc.	κ	prec.	F1
Bolognesi	0.389	0.733	0.630	0.778	0.746	0.126	0.566	0.535	0.664	0.586
Vivas Feat	0.413	0.865	0.797	0.882	0.869	0.225	0.622	0.591	0.726	0.623
Vivas Asso	0.409	0.860	0.791	0.865	0.861	0.138	0.594	0.565	0.606	0.565
Chen A	0.422	0.780	0.676	0.787	0.781	0.194	0.540	0.502	0.624	0.549
Chen B	0.422	0.730	0.606	0.755	0.732	0.194	0.507	0.468	0.614	0.509
Liu	0.407	0.723	0.574	0.762	0.729	0.272	0.489	0.425	0.544	0.491
IAA (Chen AB)	0.344	0.884	0.830	0.893	0.884	0.139	0.816	0.801	0.830	0.817

[a] All accuracy vs MC comparisons were significant, $p < .001$.

[b] Recall is identical to accuracy after prevalence weighting.

Table 3: Token-based classification results (acc. = accuracy, MC = Majority Class, κ, prec. = precision, F1) for the Level 1 (left) and 2 (right) ontologies across semantic feature (Bolognesi-2017, Vivas-2022 Feat) and word association (Vivas-2022 Asso, Chen-2024, Liu-2022) datasets. We list agreement with GPT-4 for individual annotators (Chen A and B), alongside inter-annotator scores for annotators A and B of Chen-2024 (IAA ChenAB).

.

generally, GPT-4 tends to be biased towards taxonomic responding, which is not always incorrect, but highlights the fact that relation types are not mutually exclusive.

The remainder of the error analysis at the detailed Level 2 will primarily focus on the word association datasets (Chen-2024 and Liu-2022). The micro-level confusion matrix for the Chen-2024 dataset shown in Figure 2 indicates a combination of confusion within and between macro-categories. As shown in the upper left corner, the LLM struggles to distinguish between different types within the Level 1 Taxonomy group, favoring Synonymy over Coordinate, Superordinate and Subordinate relations. The LLM also confuses Synonyms with Entity properties and Entity components. Examples of Entity properties include confusion where Large Wholes are confused with Situated-objects and Entity components. Among Situation properties, functions and actions are also frequently confused.

As can be seen from the large proportion of highlighted off-diagonal elements in Figure 3, confusion is spread across all four major semantic relation categories. It is seemingly lower for Taxonomic categories, although it should be noted that the Liu-2022 ontology does not distinguish between Subordinates and Superordinate relations, which might skew the comparison with Chen-2024. Beyond Level 1 confusion in Figure 1, Figure 3 shows that different types of Situation properties are not clearly distinguished.

To illustrate, Figure 3 shows that situational actions (S-act) and thematic relations (S-them) are easily confused. This is also an interesting case. The former is defined in the instructions as "An

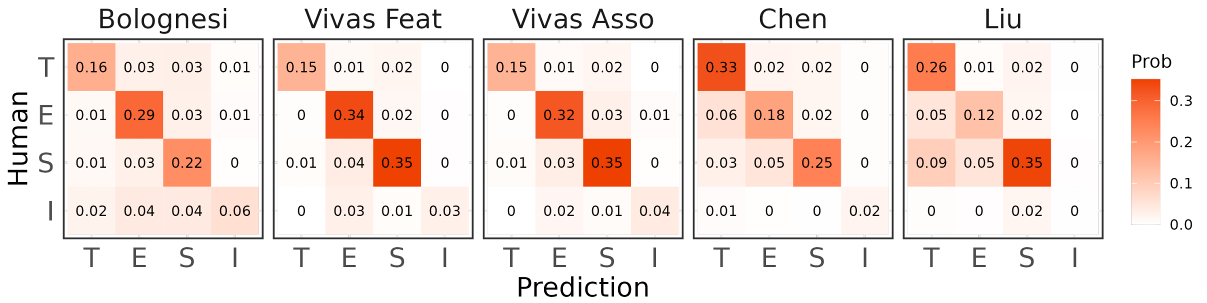

Figure 1: Confusion matrices for macro-level distinctions across five datasets (Properties: T = Taxonomic, E = Entity/Concept, S = Situation, I = Introspective).

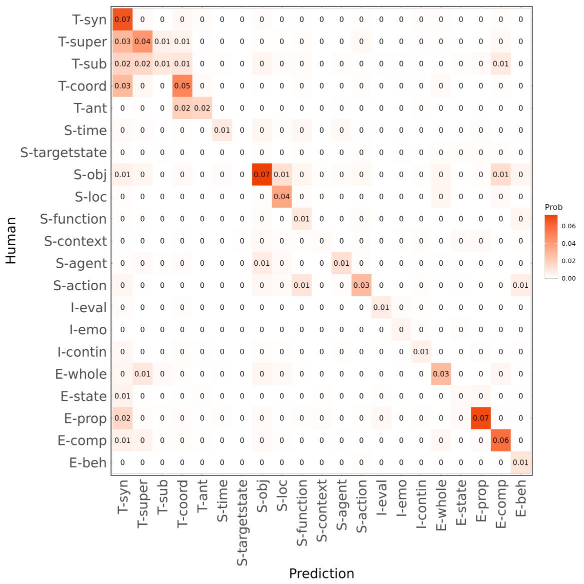

Figure 2: Confusion matrix for the Chen-2024 dataset showing a cross-tabulation of proportions for GPT-4 on the x-axis and human (coder A) reference classification on the y-axis.

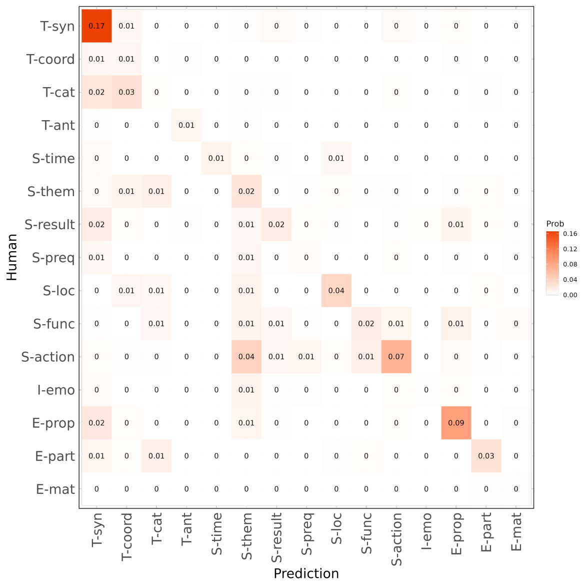

Figure 3: Confusion matrix for the Liu-2022 dataset analogues to Figure 2.

action that a participant (could be the cue, association or others) performs in a situation., whereas thematic relations are defined as “Cue and association participate in a common event or scenario. None of the other situational properties applies.”. Examples of misclassified actions include *dollar – earn* and *running – race*. Still, other pairs like *tactful – conversation* that were labelled S-act by humans but S-them by the LLM illustrate the presence of false negatives (and potential limitations of the original ontology) as well.

4. Discussion

In this study, we investigated to what degree GPT-4 can recover the latent semantic relations in word association tasks. While the findings pertain to datasets with different stimuli and different variants of the WB ontology, the overall pattern of results was consistent. First, our results across two word association datasets using GPT-4 showed good performance in making broad distinctions regarding Taxonomic, Entity, Situation and Introspective properties. More fine-grained distinctions were predicted only partially, despite relatively high levels of human inter-annotator agreement. This suggests room for further improvements, although procedural aspects such as calibration or consensus coding, which are commonly employed in human annotations, make this comparison less straightforward.

Second, a comparison with human data derived from the Property Listing Task showed high performance in capturing broad distinctions and good to moderate performance in making fine-grained distinctions. Moreover, this performance was not entirely driven by the fact that the responses in the property listing task are less ambiguous. Even when disambiguation information in the form of explicit indicators was removed, and only a single word was retained, performance was similar. Furthermore, the performance for word associations was on par when compared to the Bolognesi-2017 dataset that covered a more challenging set of cues by including many abstract concepts and single-word responses.

4.1. Comparison with previous work

As far as we know, previous work that has used LLMs to predict semantic relations using the WB taxonomy is limited. One exception is the work by Liu et al. (2022), in which a subset of training relations was used to fine-tune BERT (Devlin et al., 2019) and BART-Large (Lewis et al., 2020) to predict performance among a test set of the Liu-2022 relations. We investigated how BART, the best-performing model, compared with GPT-4 for 88 unique cue-response pairs shared among both datasets.

Across all analyses, the results showed that the GPT-4 outperformed BART. Illustrating this with the token-based analysis, the results for the L1 level were GPT-4: accuracy = 0.732, κ = 0.593, precision: 0.781, F1 = 0.733; and BART: accuracy = 0.653, κ = 0.479, precision: 0.645, F1 = 0.641. At the more detailed L2 level, we obtained for GPT-4: accuracy = 0.521, κ = 0.455, precision: 0.621, F1 = 0.535; and BART: accuracy = 0.493, κ = 0.431, precision: 0.540, F1 = 0.491. Interestingly, when comparing both types of LLMs, their mutual agreement was higher than that obtained against human annotators. For the L1 level: accuracy = 0.756, κ = 0.623, precision: 0.775, F1 = 0.742 and for the L2 level: accuracy = 0.577, κ = 0.523, precision: 0.616, F1 = 0.562. This suggests that the relations predicted by different types of language models might have more in common with human annotators. That said, given the small number of pairs in this comparison, more work is needed before strong conclusions can be drawn.

4.2. How ambiguous are word associations?

One way of determining to what degree word associations can be annotated is by comparing the relative performance for agreement among human annotators and LLM predictions. To do so, we compared the same set of classification metrics for the responses of two annotators in Chen-2024 against the LLM prediction. This showed that some relations are inherently difficult for human annotators and LLMs (e.g., S-targetstate, S-function). Other relations, like subordinates, have high agreement among annotators but low agreement in LLMs (see Figure 4) To illustrate, a pair like *sister – daughter* is coded as a subordinate relation. At least two factors could potentially explain these findings. First, in most cases, synonyms and antonyms are also coordinates. As such, this suggests a shortcoming of the WB ontology, which could be resolved by adding a third level to the hierarchy where synonyms/antonyms are nested under coordinates. Second, GPT-4 might struggle with directional relations such as superordinates and subordinates, which is supported by the patterns in Figure 2 and Figure 3, showing difficulties distinguishing between synonyms, superordinates and subordinates. To investigate this possibility in more detail, we prompted GPT-4 with explicit propositions such as *daughter is a kind/type/instance of sister*, but this did not dramatically improve results.

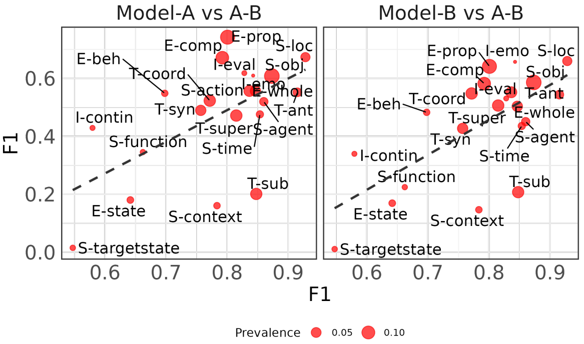

Figure 4: Comparing Coder A and B F1-scores vs Model predictions. Node size indicates prevalence. Observations under the regression line show relations that have higher F1 for human coders compared to LLM classifications against Coder A (left panel) or Coder B (right panel).

While responses in the WAT are ambiguous without further insight from the participant who generated them, it is possible that in cases with ambiguity, associations are prone to several biases that promote certain interpretations over others. Specifically for concrete words, our results contrasting the Vivas-2022 features with association responses that removed relation indicators suggest that for concrete concepts, association-like features can be generated without much loss of information.[3] However, performance also depends on concreteness. The high performance for the Vivas-2022 association dataset might reflect the fact that most of the words were very concrete. However, consistent with previous findings by Liu et al. (2022), the Bolognesi-2017 metrics, which cover both concrete and abstract concepts, were somewhat lower than the primarily concrete data from Vivas-2022.

[3] One caveat is that the PLT is a more restricted form of the WAT because only a subset of semantic relations are highlighted in the participant's instructions (often accompanied by examples), whereas word associations are free.

4.3. Limitations

The use of a closed-sourced model has several inherent limitations. While these have been discussed at length elsewhere (e.g., Frank, 2023), it should be noted that some limitations are practical in nature. One of them is that different prompting regimes cannot be controlled experimentally as the cost to do so would become prohibitively large. In all our analyses, the model was asked to generate responses for 100 item pairs simultaneously, reflecting such constraints.

Second, there are also limitations to the WB ontology. On the one hand, some of the distinctions used in the original work were specific to research questions related to groundedness (e.g., in contrasting internal and external perceptual features, as these were implied in mental simulations) (Wu and Barsalou, 2009). The ontology also needs to be further adapted to work for word associations. While this does not present major difficulties, some details do not translate well (e.g., “Contingencies and complex cognitive operations”). Furthermore, distinctions between entity and situations properties, such as *function* (currently encoded as a Situational property) or *behavior* (currently encoded as an entity property), tend only to be distinguished in terms of how typical they are for an entity or a situation. As a consequence, some of the entity vs situation properties might be conflated with whether they apply in most situations or specific ones.

4.4. Future directions

The current work primarily focused on the WB ontology. Still, other ontologies have taken inspiration from modal-specific neuroscientific models to distinguish different ways in which words could be related (Garrard et al., 2001; Montefinese et al., 2013; Vinson and Vigliocco, 2008). It would be interesting to see how SOTA LLMs would account for these, especially since this would require access to accurate perceptual information (but see Marjieh et al., 2023, for a convincing demonstration of GPT-4 in this area).

An alternative approach could *infer* task-specific relation ontologies from word associations themselves. Liu et al. (2022) collected free-text explanations with word associations and then clustered explanations into data-driven relation types without supervision. LLMs may be prompted with a less constrained framework to allow for the generation of a label inventory from scratch.

While the current work focuses on labelling a single relation, the ontologies allow for multiple relation labels for a specific cue-response pair. A more refined procedure would consider the possibility that multiple labels might apply but vary in degree or prototypicality (Jurgens et al., 2012; Liu et al., 2022). Here one possibility would be to derive classification probabilities from a fine-tuned LLM in combination with either a sparsity constraint or a rule-based approach to ensure the number of relations that can be inferred remains small. Furthermore, much more work is also needed to determine the best way to prompt the model, including which definitions to give and what examples to provide (see Jurgens et al., 2012, for an interesting analogy-based approach). Furthermore, it is likely that different types of LLMs benefit from different prompt types, and further gains could be achieved by, for example, implementing a voting mechanism across multiple LLMs.

More broadly, many questions remain about determining what semantic relations to derive in the first place. While an answer to this depends on the intended use of these relations, LLMs could assist us in iteratively refining existing ontologies by merging or splitting distinctions or refining definitions of relations. This could go in tandem with a data-driven use of LLMs to freely group different types of cue-response pairs or label the relations might prove useful (e.g., Liu et al., 2022).

5. Conclusion

Recent Large Language Models hold considerable promise in annotating semantic relations from human elicitation tasks such as word associations. The current results suggest that broad distinctions are adequately captured by GPT-4, which is considered state-of-the-art at the moment of writing. GPT-4 requires very limited requirements editing of responses, which is important to scale the approach. However, there is sufficient room for improvement, especially for more fine-grained distinctions, such as different types of taxonomic relations. While the recovery of latent semantic relations in word association data will always be subject to some degree of ambiguity, the current results also suggest several ways in which existing coding schemes can be improved to facilitate the annotation process, which ultimately would benefit the automatic labelling of these relations as well.

6. Acknowledgements

We want to thank the Complex Human Data Hub at the University of Melbourne for funding support, Chen Wen, Meredith McKague, and Enzo Capistrano for thoughtful discussions, and the reviewers for their valuable suggestions. LF is supported by the Australian Research Council Discovery Early Career Research Award (Grant No. DE230100761).

7. Bibliographical References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Marianna Bolognesi, Roosmaryn Pilgram, and Romy van den Heerik. 2017. Reliability in content analysis: The case of semantic feature norms classification. *Behavior Research Methods*, 49:1984–2001.

Andrew Cattle and Xiaojuan Ma. 2017. Predicting word association strengths. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 1283–1288.

Wen Chen, Chunhua Liu, Meredith McKague, and Simon De Deyne. 2024. Semantic alignment in Chinese and English: a concept ontology-based approach.

Simon De Deyne, Danielle J Navarro, Andrew Perfors, Marc Brysbaert, and Gert Storms. 2019. The Small World of Words English word association norms for over 12,000 cue words. *Behavior Research Methods*, 51:987–1006.

James Deese. 1965. *The structure of associations in language and thought*. Johns Hopkins Press.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*, pages 4171–4186. Association for Computational Linguistics.

Tess Fitzpatrick and Peter Thwaites. 2020. Word association research and the l2 lexicon. *Language Teaching*, 53(3):237–274.

Michael C Frank. 2023. Openly accessible llms can help us to understand human cognition. *Nature Human Behaviour*, 7(11):1825–1827.

Peter Garrard, Matthew A Lambon Ralph, John R Hodges, and Karalyn Patterson. 2001. Prototypicality, distinctiveness, and intercorrelation: Analyses of the semantic attributes of living and nonliving concepts. *Cognitive neuropsychology*, 18(2):125–174.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd-workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(3).

Thomas L Griffiths, Mark Steyvers, and Joshua B Tenenbaum. 2007. Topics in semantic representation. *Psychological review*, 114(2):211.

Simon Jerome Han, Keith J Ransom, Andrew Perfors, and Charles Kemp. 2024. Inductive reasoning in humans and large language models. *Cognitive Systems Research*, 83:101155.

David Jurgens, Saif Mohammad, Peter Turney, and Keith Holyoak. 2012. Semeval-2012 task 2: Measuring degrees of relational similarity. In ** SEM 2012: The First Joint Conference on Lexical and Computational Semantics–Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 356–364.

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*, pages 7871–7880, Online. Association for Computational Linguistics.

Chunhua Liu, Trevor Cohn, Simon De Deyne, and Lea Frermann. 2022. Wax: A new dataset for word association explanations. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing*, pages 106–120.

Chunhua Liu, Trevor Cohn, and Lea Frermann. 2021. Commonsense knowledge in word associations and ConceptNet. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 481–495.

Hongjing Lu, Nicholas Ichien, and Keith J Holyoak. 2022. Probabilistic analogical mapping with semantic relation networks. *Psychological review*, 5:1078–1103.

Raja Marjieh, Ilia Sucholutsky, Pol van Rijn, Nori Jacoby, and Thomas L Griffiths. 2023. What language reveals about perception: Distilling psychophysical knowledge from large language models. *arXiv preprint arXiv:2302.01308*.

Maria Montefinese, Ettore Ambrosini, Beth Fairfield, and Nicola Mammarella. 2013. Semantic memory: A feature-based analysis and new norms for italian. *Behavior research methods*, 45:440–461.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31.

Massimo Stella, Nicole M Beckage, and Markus Brede. 2017. Multiplex lexical networks reveal patterns in early word acquisition in children. *Scientific reports*, 7(1):46730.

Sean Trott. 2023. Can large language models help augment english psycholinguistic datasets?

David P Vinson and Gabriella Vigliocco. 2008. Semantic feature production norms for a large set of objects and events. *Behavior Research Methods*, 40(1):183–190.

Leticia Vivas, Maria Montefinese, Marianna Bolognesi, and Jorge Vivas. 2020. Core features: measures and characterization for different languages. *Cognitive processing*, 21(4):651–667.

Leticia Vivas, M Yerro, Sofía Romanelli, A García Coni, Ana Comesaña, F Lizarralde, I Passoni, and J Vivas. 2021. New spanish semantic feature production norms for older adults. *Behavior Research Methods*, pages 1–17.

Lingling Wu and Lawrence W Barsalou. 2009. Grounding concepts in perceptual simulation: Evidence from property generation. *Acta Psychologica*.