How Humans and LLMs Organize Conceptual Knowledge: Exploring Subordinate Categories in Italian

Andrea Pedrotti$^{\alpha}$**, Giulia Rambelli**$^{\beta}$**, Caterina Villani**$^{\beta}$**, Marianna Bolognesi**$^{\beta}$
$^{\alpha}$Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo" (ISTI-CNR)
andrea.pedrotti@isti.cnr.it
$^{\beta}$Università di Bologna
{giulia.rambelli4,caterina.villani6,m.bolognesi}@unibo.it

Abstract

People can categorize the same entity at multiple taxonomic levels, such as basic (*bear*), superordinate (*animal*), and subordinate (*grizzly bear*). While prior research has focused on basic-level categories, this study is the first attempt to examine the organization of categories by analyzing exemplars produced at the subordinate level. We present a new Italian psycholinguistic dataset of human-generated exemplars for 187 concrete words. We then use these data to evaluate whether textual and vision LLMs produce meaningful exemplars that align with human category organization across three key tasks: exemplar generation, category induction, and typicality judgment. Our findings show a low alignment between humans and LLMs, consistent with previous studies. However, their performance varies notably across different semantic domains. Ultimately, this study highlights both the promises and the constraints of using AI-generated exemplars to support psychological and linguistic research.[1]

1 Introduction

Concepts are the "building blocks" of human cognition, allowing us to interpret and categorize reality (Murphy, 2002). The same category can be represented at different levels of inclusiveness (*categorical specificity*; Bolognesi et al., 2020). For instance, a two-wheeled object may simultaneously be categorized as an *electric bike*, a *bike*, or a *vehicle*, reflecting a hierarchical taxonomy that ranges from a very specific and not inclusive category that only includes members with many common features (*mountain bikes*, *electric bikes*) to a more general and inclusive category that includes a wide variety of items that do not necessarily share many common features (*bike, cars, bus*).

Most studies on hierarchical organization of categories in the human mind have focused on basic-level categories, showing their advantages in processing and acquisition (Rosch et al., 1976; Hajibayova, 2013 for a review), paying little attention to the more specific *subordinate* categories. Yet, words at the subordinate level are crucial for effective communication in specialized domains, as their lexicon conveys richer and more precisely defined semantic content, often derived through linguistic combinations.

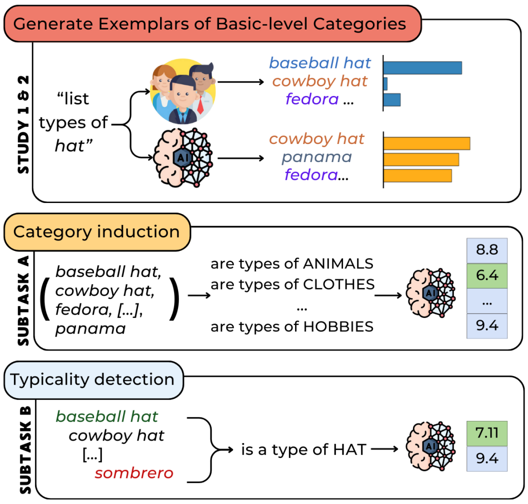

Figure 1: Visual representation of studies' design. English exemplars are used for illustration only.

Current cognitive theories acknowledge that both sensorimotor and linguistic experiences contribute to our conceptual representation (Barsalou et al., 2008; Louwerse, 2018; Davis and Yee, 2021). For instance, one may observe that *apples* can be red, yellow, or green, but learn in a book that the word *Fuji* refers to a specific variety of apples. Although concepts can be represented independently from words, linguistic labels often act as cues (Lupyan, 2012; Lupyan and Lewis, 2019) that help to create and organize our knowledge, grouping items based on perceived similarities, even if we have never encountered a particular instance before. The extent to which the organization of

[1] Data and code is available on GitHub and OSF.

human conceptual categories is influenced by the distributional properties of linguistic input remains a central question in cognitive science, linguistics, and artificial intelligence (van Hoef et al., 2023).

This paper investigates the organization and the contents of conceptual categories produced at a subordinate level by humans and Large Language Models (LLMs). The remarkable success of LLMs raises questions about their plausibility as models of human cognition, as their performance closely resembles human-like language understanding and generation across several tasks (Wang et al., 2018; Brown et al., 2020; Floridi and Chiriatti, 2020; Bommasani et al., 2022; Wei et al., 2022). However, while their functional linguistic competence—reflected in their general knowledge and reasoning skills through language—is undeniable, their parallelism with the human mind remains highly debated (i.a., Bender and Koller, 2020; Marcus, 2020; Mahowald et al., 2024). In contrast to LLMs, human conceptual categories emerge from the integration of linguistic and extra-linguistic (sensory) information. Investigating the structural organization of categories in LLMs may provide insight into the extent to which category formation depends exclusively on linguistic experience; thus, contributing to the larger debate on the role of language in learning semantic knowledge (Lupyan and Lewis, 2019). While previous works have explored the organization of superordinate categories in both humans and LLMs, we are the first to investigate the organization of basic-level categories. Specifically, we present two studies to address the following research questions:

- **RQ1: How do humans create and organize basic-level categories, considering the exemplars produced at a subordinate level?** We introduce a new Italian psycholinguistic dataset, collecting exemplars of 187 basic concrete categories generated by human participants (§3). We explore the variability of exemplars as a function of category types, assuming that this variability reflects the richness of the linguistic vocabulary and linguistic knowledge in semantic domains.
- **RQ2: Do LLMs have the same category structure as humans?** We probe recent LLMs to generate exemplars for the same 187 basic-level categories and compare their predictions with humans (§4), as illustrated in Figure 1. We assess whether LLMs capture human conceptual organization using two classification subtasks: *category induction* (§5.1) and *typicality prediction* (§5.2). Finally, we compare vision LMs (vLMs) to investigate whether pre-training extra-linguistic knowledge enhances overall performance.

2 Background and Related Works

2.1 Categories in the Human Mind

Classical cognitive research showed that categories are organized hierarchically in the human mind: a *bulldog* is a type of *dog*, which is a type of *mammal*, and more broadly an *animal*, with each category including the previous one. In other words, categories vary in level of *specificity*—i.e., how inclusive the category of reference is (Cohen and Lefebvre, 2005; Bolognesi et al., 2020). Superordinate categories (e.g., *furniture, vehicle*) encompass broader classes, while subordinate categories (e.g., *wooden upholstered chairs, red sports cars*) represent more specific instances. The basic level (e.g., *chair, car*), often considered the most informative level, lies between these two extremes, and words that denote basic-level categories are typically easier to understand and process (Murphy, 2002).

A common approach for investigating the structure of categorical knowledge involves analyzing typicality effects, by asking typicality ratings on a Likert scale (i.e., "How typical is a *cat* for the category *mammal*?") or by instructing participants to freely name members of a given category. The ladder, called "semantic fluency" or "category instance generation" (Castro et al., 2021), requires participants to actively retrieve exemplars of a category, which is a more cognitively demanding task than simply judging its typicality within a category. However, typicality ratings can be extracted from category instance generation tasks by aggregating the frequency of exemplar productions. Conversely, words judged as typical for their category are usually more available than words judged to be relatively atypical (Natividad Hernández-Muñoz and Ellis, 2006). In seminal studies, Rosch (1978) observed that some exemplars (e.g., *robin, crow*) are perceived as more representative of a category (e.g., *birds*) than others (e.g., *penguin, ostrich*). This graded structure, as explained by prototype theory (Rosch, 1975), reflects the fact that frequently shared properties among category members tend to be integrated into a central prototype.

While cognitive research has extensively focused on superordinate and basic-level categories, subordinate categories have received less attention. Concepts at the subordinate level have some notable peculiarities. First, their referents share more attributes than those within basic-level categories (Rosch et al., 1976). Additionally, subordinate concepts encode greater perceptual detail, making it more challenging to process individual exemplars. As a result, people tend to name objects at the basic level unless subordinate-level information is particularly relevant. Finally, language plays a crucial role in forming subordinate categories, often created through linguistic compositionality (*electric car, sports car*). To the best of our knowledge, no studies in English or any other language have investigated the organization and contents of basic-level categories (e.g., *dog*, *hammer*), by asking participants to generate concepts at the subordinate level.

2.2 Categories in LLMs

Previous works on predicting category structure in LLMs have primarily focused on the typicality of a category member, yielding mixed results. Heyman and Heyman (2019) predicted human typicality ratings by correlating similarity scores between category exemplars (e.g., *robin, crow*) and prototype vectors (*bird*), finding that static embeddings poorly accounted for human judgments. Renner et al. (2023) improved predictions using BERT and WordNet metrics, showing that their combination aligns best with human judgments. Recently, Heyman and Heyman (2024) found out that ChatGPT produces typicality ratings comparable to human participants (.60-.64). Conversely, Misra et al. (2021) tested LLMs on taxonomic categorization ("football is a sport"), showing modest correlations with human ratings (between 0.24 and 0.41) and weaker distinctions between typical and atypical items (as observed in other experimental settings, i.e., Kauf et al., 2023). Moreover, Misra et al. (2023) highlighted that LLMs struggle with fine-grained property attributions, questioning their plausibility as models of human semantic memory.

Beyond typicality, Nighojkar et al. (2022) used Transformer models (RoBERTa-Large, DistilBERT, and miniBERTa-med-small) to model the semantic fluency task (§2.1). They designed different approaches to predict the next item in a given list ("Examples of fruits are the strawberry and the [MASK]") for five superordinate categories (Fruits, Vegetables, Animal, Supermarket items, Tool, Foods). Among the models, RoBERTa-Large proved to be the best-performing approach, although it still achieved low performance (16% overall accuracy).

Concurrently, researchers have investigated whether vision models align with human conceptual understanding (Peterson et al., 2018; Battleday et al., 2020; Günther et al., 2023; Upadhyay et al., 2022). Regardless of the specific experimental design, these studies correlated vision-based similarity scores between any pair of exemplar and category images and evaluated these similarities against human typicality judgments. Recently, Vemuri et al. (2024) evaluated both language and vision models, comparing their correlation with human typicality ratings, and found that textual models are better than vision models for 27 categories, surpassing prior results from Castro et al. (2021).

Recent works have also tested the abstract reasoning abilities of LLMs. For example, Samadarshi et al. (2024) assessed LLMs performance on the *New York Times* Connections game, finding better performance in Semantic Relations and Encyclopedic Knowledge, which might be due to existing information in pre-training data. However, LLMs accuracy remains below 50%.

All the aforementioned works focused exclusively on English, and primarily explore the internal organization of superordinate categories (e.g., *fruit, tools*). To our knowledge, no research has yet explored this for Italian or investigated the internal organization of basic-level categories (e.g., *dog, hammer*).

3 Study 1: A New Psycholinguistic Dataset of Basic-Level Exemplars

Methods. Stimuli consist of 187 basic-level concrete categories previously produced by Italian native speakers as the most representative concepts for 12 superordinate semantic categories[2] (Montefinese et al., 2012). We administer an exemplar generation task to 365 Italian L1 speakers on Prolific. Each participant is presented with a list of 15-16 categories and asked to produce as many exemplars as possible for each concept (e.g., `List a type of`) at their own pace. The final dataset, after post-processing typos and misspellings, consists of 24.659 exemplars.

[2] ANIMALS, BODY PARTS, CLOTHES, FOODS, FURNISHING, FURNITURE, HOBBIES, HOUSING, KITCHEN, PLANTS, STATIONERY, VEHICLES.

We compute the same measures as Montefinese et al. (2012) to describe the relationship between a given concept and its exemplars, such as the proportion of participants who produce a target exemplar given a category (*dominance*), the mean output position of each exemplar for a category (*mean rank order*), and the proportion of participants who produce a given exemplar as their first response (*first occurrence value*). We primarily focus on *exemplar availability*, which represents how readily an exemplar is produced as a member of a category. This measure is determined by the exemplar's position in a participant's response list, its overall production frequency within the category, the earliest position it appears across participants, and the total number of participants who mention it.

Results and Discussion. In line with Montefinese et al. (2012), we find that dominance, availability, and first occurrence are all strongly and positively correlated (r_s = 0.95, 0.75, 0.89; for dominance vs. availability, dominance vs. first occurrence, and availability vs. first occurrence); whereas mean rank order of production correlates weakly and negatively with the other three measures (r_s = - 0.09; -0.21, -0.15; for dominance, first occurrence, and availability respectively). To identify the most representative exemplars for each concept, we retain only those exemplars with a dominance value higher than or equal to 0.1 (i.e., exemplars produced by at least 10% of participants). This cut-off criterion results in a total of **1696 exemplars** in the final dataset.

Figure 2 shows the numbers of dominant exemplars for the 12 subordinate categories. The highest number of exemplars is produced for the **FOOD** category (**270** exemplars), followed by **CLOTHES** (**206** exemplars), whereas the category of PLANTS has the smallest number of exemplars (77). Indeed, the number of dominant exemplars varies considerably within each basic-level category, spanning from a minimum of 1 exemplar (e.g., *sunflower*, *rubber plant*) to a maximum of 31 exemplars (e.g., *pasta*, *dog*). The fact that the extent of our subordinate lexicon varies in human cognition suggests that **some subordinated categories might pose challenges in terms of accessibility to semantic memory**. This could be due to their low frequency or familiarity, or to a higher degree of individual variability in knowledge within a specific domain compared to others.

Subsequently, for each basic-level category, we compare the top-1 and top-5 exemplars ordered by availability with those ordered by dominance, examining whether both the exemplars and their order align. Overall, 77.0% of the top-1 dominant exemplar is also the top-1 available. This indicates that more frequently produced exemplars tend to be more readily available in participants' responses, reflecting their prominence in the conceptual category. However, only 13.9% of the top-5 dominant exemplars overlap with the ranking of the top-5 most available exemplars. For example, the top-5 dominant exemplars for the basic category "*cereal*" are *oat, spelt, wheat, corn, barley*, while the top-5 available were *wheat, oat, spelt, corn, barley*. This outcome points to potential variability in the frequency of production and availability of exemplars within different conceptual categories. In conclusion, we observe that this task is more challenging than retrieving exemplars of superordinate-level categories and that some categories are more accessible than others.

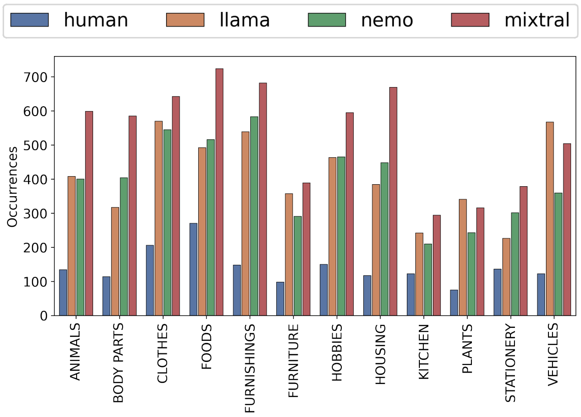

Figure 2: Number of valid exemplars across 12 superordinate categories for humans and textual LLMs.

4 Study 2: LLMs' Exemplars Generation

We probe several LLMs on the task described in §3 to compare their organization of subordinate-level conceptual representations with human subjects. We assess models' performance considering: (*i*) the number of hallucinations generated (i.e., non-existent exemplars created by combining words into *ad hoc* instances); (*ii*) the overlap with human subjects regarding the most available (typical) exemplar, and (*iii*) whether discrepancies between human and LLMs-generated exemplars follow a consistent pattern.

We analyse our data from two complementary

	ANIMALS	BODY PARTS	CLOTHES	FOODS	FURNISHING	FURNITURE	HOBBIES	HOUSING	KITCHEN	PLANTS	STATIONERY	VEHICLES	avg
llama-3.2-3B	0.63	0.74	0.76	0.84	0.61	0.69	0.67	0.72	0.59	0.50	0.63	0.63	0.67
llama-3.1-8B	0.48	0.64	0.64	0.86	0.61	0.69	0.74	0.72	0.49	0.41	0.49	0.63	0.62
llama-3.1-70B	0.81	0.77	0.89	0.98	0.82	0.83	0.85	0.93	0.80	0.68	0.61	0.83	**0.82**
mistral-7B	0.52	0.61	0.43	0.79	0.50	0.41	0.69	0.61	0.46	0.42	0.39	0.57	0.53
nemo-12B	0.71	0.72	0.69	0.90	0.71	0.65	0.79	0.86	0.56	0.47	0.58	0.70	0.69
mixtral-8x7B	0.73	0.76	0.77	0.95	0.74	0.76	0.81	0.86	0.67	0.54	0.53	0.79	0.74
llava-7B	0.52	0.60	0.54	0.67	0.57	0.53	0.70	0.61	0.48	0.48	0.57	0.61	0.57
idefics2-8B	0.64	0.76	0.62	0.80	0.75	0.67	0.82	0.71	0.53	0.67	0.65	0.65	0.69
category avg	0.63	0.70	0.67	**0.85**	0.66	0.66	0.76	0.75	0.57	0.52	0.56	0.68	0.67

Table 1: Percentage of valid exemplars generated by various LLMs.

perspectives. On the one hand, we assess the models' accuracy based on their similarity to human-generated exemplars (our gold standard). On the other, we perform some qualitative analyses to explore whether and how the categorical knowledge encoded by language models differs from that of humans.

Setup. Building upon the methodology described in §3, we task the models with generating exemplars for the same 187 basic-level concepts presented to human subjects. We use two LLMs families: *(i)* **LLaMA family**, including LLaMA-v3.1 in its 8 and 70B versions, and LLaMA-v3.2-3B (LlamaTeam, 2024), and *(ii)* **Mistral family**, comprising Mistral-7B (Jiang et al., 2023), Mixtral-8x7B (Jiang et al., 2024), and NeMO[3]. Furthermore, to investigate the impact of perceptual extra-linguistic stimulus, we also use the vLMs **LLaVA** (Liu et al., 2023) and **Idefics2** (Laurençon et al., 2024) (cf. Appendix B.1 for an in-depth description).

We model the generation process as a few-shot setting (Brown et al., 2020) completion task. The model receives a simplified version of the instructions from §3 to obtain comparable results. The instruction is followed by a question-answer example before generating exemplars for a new concept. We follow the few-shot prompting scenario, as this approach should positively affect the model's performance. We experiment with parameters to obtain an outcome balanced between predictability and creativity. For each model, we perform five runs for each basic-level category (cf. Appendix B.5).

4.1 Analysis 1: LLMs Tends to Generate *ad hoc* Expressions instead of Exemplars

The generated responses consist of a list of exemplars separated by newlines (i.e., '\n'). To ensure data quality, we first clean the outputs by removing duplicate exemplars, keeping only their first occurrences. We then validate the outputs by checking whether each exemplar appears at least once in the Italian corpus ItTenTen (Jakubíček et al., 2013; Suchomel et al., 2012)[4], thereby distinguishing valid exemplars from (possible) hallucinations.

This data-cleaning step allows for an overall evaluation of the quality of the generated exemplars in terms of the percentage of valid (i.e., existing expression) exemplars. Table 1 shows that the performance differs widely across models, with larger and more recent LLMs generating a higher proportion of valid exemplars in comparison to smaller models or vLMs. For instance, LLaMA-v3.1-70B generates 82% valid exemplars, while Mistral-7B generates only 52% valid exemplars. The lowest performance is observed by LLaVa-7B (44%).

Notably, the number of valid exemplars varies depending on the superordinate category. Categories, such as **FOOD** (**85%**), **HOBBIES** (**76%**), and **HOUSING** (**75%**), yield a higher proportion of valid exemplars across models. In contrast, categories like KITCHEN and PLANTS exhibit more noise, with only 57% and 52% of valid exemplars, respectively. This indicates that models acquire a non-uniform knowledge of subordinate-level exemplars, with a broader and more precise coverage of certain basic-level concepts, while showing a more brittle grasp of others. These results partially align with human behaviour: the categories' exemplars that are easiest (FOOD) and those that are most difficult (PLANTS) to recall are the same for both humans and LLMs.

Considering unattested expressions, LLMs often rely on their compositional abilities to generate surface-acceptable expressions. However, this 'creative' process produces invalid multi-word expressions (i.e., hallucinations) that lack validation among human speakers (i.e., their corpus frequency is zero) and/or real-world referents. We conduct a qualitative analysis of zero-frequency items to identify recurring generative tendencies on LLaMA-3.1-70B (the best-performing model

[3] https://mistral.ai/news/mistral-nemo/

[4] We use the SketchEngine API to collect frequencies.

in terms of valid exemplars generated). Among others, we observe that the model tends to replicate the surface-level syntactic or morphological structure of a valid, attested exemplar, leading to the overgeneralization of that structure to produce novel combinations. For instance, the expression *abete rosso* ('red fir') and *abete di Douglas* ('Douglas fir') serve as a template for generating further expressions like *abete bianco di Scozia* ('white Scotch fir') or *abete rosso di California* ('red California fir'), none of which refer to real-world referents. Similarly, the models extract from *candelabro a 5 braccia* ('5-armed candelabrum') the syntactic pattern a N bracci/a to build multiple variants, as *a 13 bracci*. Therefore, models tend to identify productive syntactic patterns and extend them compositionally, rather than drawing on actual distributional evidence or domain knowledge. In essence, **imitation-based** errors are structural extrapolations that mirror known exemplars too closely, prioritizing form over grounded meaning.

Additionally, the generated expressions are grammatically well-formed but semantically incoherent, implausible, or internally contradictory. For example, *geranio a foglie di quercia* ('geranium with oak leaves') or *a foglie di rosmarino* ('with rosemary leaves') attribute biologically implausible features. Similarly, *maglia a punto croce* ('knitwear in cross-stitch') is semantically incoherent, because *punto croce* is a specific embroidery technique used to decorate fabrics—not for constructing knitwear. In these cases, LLMs apply compositional plausibility without conceptual coherence: models generate a surface-acceptable phrase that violates domain-specific knowledge or real-world constraints, thereby rendering the expression **nonsensical**. Finally, some generated outputs are not attested exemplars but rather **novel, *ad hoc* instances** (Barsalou, 1983). For example, the model generates instances of *cassettiera* ('dresser') based on spatial context (e.g., *c. da corridoio* 'hallway dresser', *c. da esterno* 'outdoor dresser') or intended contents (e.g., *c. per giocattoli* 'for toys', *per oggetti di cancelleria* 'for stationery items'). While such expressions might be interpretable and even plausible, they are not attested in usage and do not correspond to established members of the category, i.e., they do not qualify as exemplars stored in long-term memory.

Additional examples of these generative patterns are provided in Tables 7 and 8 (cf. Appendix B.8). Overall, these examples illustrate how hallucinations often arise from systematic, though flawed, generalization strategies, revealing **a gap between surface-level fluency and semantic grounding**.

4.2 Analysis 2: Humans and LLMs Disagree on the Most Available Exemplars

In the second analysis, we compare the valid exemplars generated by the LLMs with human-generated exemplars. Specifically, we sort both human and LLMs exemplars according to their *availability score*, which reflects the ease with which a word can be produced as a category member (§3). Table 2 reports the results of the intersection between the top-n ($n = \{1, 3, 5\}$) most available human-generated and machine-generated exemplars, with overlap computed regardless of the production order. The best results are observed for top-5 matches, with Nemo-12B reaching an overlap of 24% of the generated exemplars. The number of matches varies across categories (cf. Appendix B.7). The most significant overlap is observed within the categories of FOODS (Nemo-12B: 37%, overall: 29%) and ANIMALS (Nemo-12B: 36%, overall: 29%). In contrast, the lowest overlap emerges within the categories BODY PARTS and FURNISHING (Nemo-12B: 16%, overall: 12%).

These lower scores may arise for two reasons. First, the model generates valid exemplars, sometimes even matching those produced by humans, but not the most available ones. For example, the top-5 human-generated exemplars of *cane* 'dog' (*labrador*, *pastore tedesco* ' German shepherd', *bassotto* 'dachshund ', *chihuahua*, *golden retriever*) only partially overlap with those generated by nemo-12B (*pastore tedesco*, *golden retriever*, *beagle*, *labrador*, *husky siberiano* 'siberian husky'). Besides, *bulldog* is in the top-5 most available exemplars in five models, despite having a lower corpus frequency than other words (e.g., *chihuahua*, *dalmatian*). The variation among models suggests that there are **no specific criteria** (e.g., frequency) **that determine the generation of one exemplar over another, implying a category organization that is essentially flat**.

Secondly, some models produce incorrect exemplars: in some cases, meronyms are generated (i.e., *polpaccio* 'calf' as a type of *gamba* 'leg'), in others, the basic-level category is misinterpreted due to polysemy (i.e., the word *braccio* 'arm' refers both to a human body part and to an extension of some-

Model	Top-1	Top-3	Top-5
llama-3.2-3B	0.09	0.13	0.14
llama-3.1-8B	0.14	0.18	0.20
llama-3.1-70B	0.18	0.20	0.21
mistral-7B	0.13	0.12	0.13
nemo-12B	**0.25**	**0.24**	**0.24**
mixtral-8x7B	0.18	0.19	0.19
llava-7B	0.12	0.13	0.15
idefics2-8B	0.08	0.10	0.10

Table 2: Matches among the top-n human and machine-generated most available exemplars.

Model	Basic-level	Superordinate
llama-3.2-3B	0.84	0.52
llama-3.1-8B	0.96	0.63
llama-3.1-70B	0.95	**0.64**
mistral-7B	0.89	0.59
nemo-12B	0.95	0.46
mixtral-8x7B	**0.98**	0.57
llava-7B	0.93	0.59
idefics2-8B	0.94	0.38

Table 3: SUBTASK A–Accuracy for basic-level and superordinate category prediction at the aggregated level.

thing), resulting in nonsensical outputs. Incorrect exemplar generation is especially evident in vLMs. For example, `idefics2-8B` not only relies on compositional operations but also lists other types of trees (e.g., *acacia, eucalyptus, maple* as exemplars of *abete* 'fir'), failing to generate subordinate exemplars and generating basic-level exemplars instead.

5 Are LLMs Sensitive to Human Category Structure?

The comparative analyses of human and LLMs-generated exemplars revealed no significant overlap between these two sets. However, despite some noisy *ad hoc* exemplars, models also produce valid exemplars that humans did not recall. We use human data to build two additional classification tasks:

A. **Category Induction**: Given the 10 most available human-generated exemplars, select their basic/superordinate category;

B. **Typicality Detection**: Given the most and least available human-generated exemplars, identify the typical (i.e., most available) member of the basic category.

These tasks are designed to evaluate the model's consistency in representing categories and their exemplars using close-ended formats. Rather than generating exemplars, the model selects correct answers based on its perplexity score, making evaluation easier and more reliable.

5.1 SUBTASK A: Category Induction

Previous studies revealed that basic-level members of a category can elicit the activation of their corresponding superordinate categories in the mental lexicon (Barsalou, 1982; Ross and Murphy, 1999). While tasks in §4 were focused on exemplar generation, here we explore to what extent LLMs are able to identify the category to which an exemplar belongs to. Specifically, we investigate whether subordinate-level members of a given category can activate their *(i)* basic and *(ii)* superordinate category in LLMs. This allows us to compare recall performances at different levels of taxonomy, from the (more *specific*) basic and (more *general*) superordinate categories, and to better investigate the organization of conceptual categories in the learned latent space of LLMs.

Setup. The task is structured as a classification task. Given an input sentence containing a sequence of subordinate-level exemplars, the model has to select the correct category that has produced the listed exemplars. The category can be: (*i*) one of the 187 basic-level categories (e.g., *abete* 'fir', *aereo* 'plane'), or (*ii*) one of the 12 superordinate categories (e.g., *pianta* 'plant', *veicolo* 'vehicle'). We select up to 10 most available human-generated exemplars for each basic-level concept. Each list is converted into a prompt in the form: "e_1, e_2, . . . , e_{10} `are types of {category}`", where e_n denotes the n-th selected human-produced exemplar and `category` is a category name, either at basic-level or superordinate one. We then compute the model's perplexity for each pair and select the category associated with the sentence that has the lowest perplexity score.

Results. Overall, models obtain higher results when predicting the basic-level concept (e.g., *abete* 'fir') rather than the more abstract superordinate category (e.g., *pianta* 'plant'; cf. Table 3). This result is surprising, considering that the number of superordinate categories is smaller (12 vs 187 concept terms). A possible explanation is that models have seen the occurrence <exemplar, basic-level concept> more frequently than the pair <exemplar, superordinate-level concept>. In addition, most of

Model	Low	Medium	High
llama-3.2-3B	0.65	0.62	0.42
llama-3.1-8B	0.58	0.60	0.42
llama-3.1-70B	0.73	0.68	0.61
mistral-7B	0.50	0.57	0.47
nemo-12B	0.53	0.69	0.52
mixtral-8x7B	0.72	0.55	0.57
llava-7B	0.48	0.62	0.48
idefics2-8B	0.53	0.58	0.45

Table 4: SUBTASK B – Typicality Accuracy for basic-level categories for the three coverage groupings.

Model	Low $\|\Delta\|$	Medium $\|\Delta\|$	High $\|\Delta\|$
llama-3.2-3B	0.47	0.58	0.61
llama-3.1-8B	0.45	0.50	0.57
llama-3.1-70B	0.59	0.61	0.70
mistral-7B	0.41	0.49	0.53
nemo-12B	0.43	0.49	0.69
mixtral-8x7B	0.49	0.55	0.69
llava-7B	0.42	0.47	0.61
idefics2-8B	0.46	0.47	0.49

Table 5: SUBTASK B – Typicality Accuracy for basic-level categories, grouped by the absolute difference in exemplars availability.

the time, the exemplar itself can contain the concept sub-string, e.g., *abete di Natale* ('Christmas tree') vs *?pianta di Natale* ('Christmas plant'). Interestingly, LLM performance varies across semantic domains: models score nearly perfectly on ANIMALS, KITCHEN, and VEHICLES, but perform poorly on FURNISHING, HOBBIES, and STATIONERY (cf. Appendix C). As expected, **LLMs more effectively acquire taxonomic relations for categories shaped by encyclopedic knowledge** (factual information typically learned through education or texts, e.g., "a lion is a mammal") than those grounded in commonsense knowledge (e.g. "domino is a game").

5.2 SUBTASK B: Typicality Prediction

One key aspect of category structure that has been extensively studied with LLMs is *typicality* (§2.2): Some members of a category are considered more representative than others (e.g., *robin* vs. *penguin* as types of birds). Previous studies have found only a moderate correlation between human judgments and LLMs. In addition, their focus was basic-level exemplars of superordinate categories. In this subtask, we investigate whether, despite their misalignment with humans in generating the most available exemplars (§4), LLMs can still recognize that the most available item (e.g., *bicchiere di vetro*, 'glass tumbler') is more typical than the less available one (e.g., *bicchiere da shot* 'shot glass') for a given category (e.g., *bicchiere* 'glass').

Setup. We group the 187 basic-level categories by the number of exemplars produced by humans into three groups: (*i*) **low** (up to 5 exemplars), (*ii*) **medium** (6–10 exemplars), and (*iii*) **high productivity** (more than 10 exemplars). This grouping allows us to test if the internal dimension of the category impacts typicality detection results. For each basic-level concept, we then select the most available and the least available human-generated exemplars and evaluate the models' perplexity on the two sentences: "`{1st exemplar} is a type of {concept}` vs. `{last exemplar} is a type of {concept}`." Similarly to §5.1, a pair is considered a positive prediction if the perplexity for the first sentence is lower than that assigned to the second one.

Results. Overall, `LLaMA-3.1-70B` performs best across the three groupings, reaching 73% accuracy in the low-productivity setting (cf. Table 4), a good score compared to past studies. However, accuracy varies across groupings: **as the number of human exemplars for a category increases, LLMs are less likely to detect the typical item.** This suggests that when humans provide fewer exemplars, the first one is cognitively dominant compared to the other ones, a distinction reflected in the model's perplexity scores. However, in richer categories, the cognitive distinctive attributes among exemplars diminish, thus resulting in LLMs' lower accuracies (cf. Appendix D).

Effect of Availability Differences. Additionally, we assess accuracy across groups defined by the absolute difference in availability ($|\Delta|$) between the most and least available exemplars. We categorize these differences into three levels: **low** Δ for differences less than 0.2, **high** Δ for differences greater than 0.4, and **medium** Δ for all other cases. This grouping results in a balanced distribution of pairs (57, 75, and 55, respectively). Looking at the average results in Table 5, we observe that **pairs with a higher typicality delta are easier to predict**, yielding higher accuracy scores. For example, the best performing model `LLaMA-3.1-70B` achieves almost a 20% increase when moving from the low to the high Δ setting (and a ~30% on average across all the models). This additional analysis reveals that LLMs are sensitive to the internal structure of human basic-level categories: **the smaller the variability in human availability**, **the more difficult it becomes for the model to identify the most typical items**.

6 General Discussion and Conclusions

This study explored basic-level category organization in humans, who integrate linguistic and sensory information, and LLMs, which rely solely on linguistic data. In a generation task, Italian speakers and various LLMs and vLMs produced lists of exemplars for 187 basic-level concrete categories. We hypothesized that the most frequent exemplars generated by models would align with those of humans, as subordinate concepts reflect specialized knowledge and are constrained by language.

Findings in §4 reveal a low alignment between model and human performance. However, comparative analyses show that some models (particularly `LLaMA-3.1-70B`) can still generate meaningful exemplars comparable to those produced by humans across many semantic domains. Interestingly, these models produce more exemplars than humans for technical and specialized categories that require access to encyclopedic knowledge (i.e., PLANTS): e.g., `LLaMA-3.1-70B` generates 26 real exemplars for *orchidea* 'orchid', while humans generated only 5. This ability points to a possible use of LLMs in automatically generating exemplars for large sets of concepts (i.e., for automatic ontology population), in line with similar findings for semantic feature production norms (Hansen and Hebart, 2022). However, our results also call for some caution.

First, the models often generate hallucinations and incorrect exemplars, especially for categories where extralinguistic information plays a more critical role than linguistic data. This is especially evident in the BODY PARTS category, where conceptual confusion (*piede di porco* 'crowbar') or *ad hoc* instances (*testa di cavallo* 'horse head') are common. While frequency analysis can help reduce hallucinations, human annotation is needed to verify accuracy, at least at this taxonomic conceptual level. Secondly, LLMs do not show the same categorical organization of humans. The generated exemplars vary significantly across models, with alignment to human responses below 25% (§4).

Additional subtasks in §5 illustrate that models struggle to build a hierarchical conceptual organization like humans, limiting their ability to reason along the taxonomic axis (§5.1). While they perform well in basic-level category induction, they underperform in the superordinate category setting. Moreover, LLMs often fail to identify the most typical exemplar when a category includes multiple similarly available items (§5.2) but perform better when one exemplar clearly dominates in availability. These results suggest that (proto)typicality effects are harder to detect within basic-level categories, likely due to their relatively flat internal structure and the high number of shared attributes among subordinate exemplars. Finally, we found that vLMs still perform poorly in the exemplar generation task, in line with previous research (Vemuri et al., 2024), showing that text-based models align more closely with human typicality judgments.

Our study has several methodological implications worth mentioning. We provided a dataset of human-generated exemplars for basic-level concrete categories in Italian, along with statistical measures, extending Montefinese et al. (2012). Since existing Italian datasets often lack concepts spanning multiple taxonomic levels, this resource will be useful in cognitive psychology and AI research on semantic category structure. This need for comprehended datasets becomes evident when comparing existing resources in other languages, such as English (e.g., Banks and Connell, 2023). Moreover, our study highlights the potential and limitations of LLMs in capturing human categorical knowledge at the subordinate level, in line with previous literature. Future work should explore how LLMs generate exemplars for superordinate categories (e.g., *animals, plants*) and whether they align more with human behaviour at this level. Additionally, comparing results across languages could also reveal cultural influences on concept representation and potential biases in LLMs.

In conclusion, our results show that the organization of subordinate categories varies as a function of semantic domains in both humans and LLMs. Notably, the more extralinguistic or linguistic information is relevant to a given category, the more the performance of LLMs and humans diverges. These observations have practical implications for NLP systems, such as educational tools (e.g., vocabulary teaching, interactive learning apps), knowledge base population, and generally, to improve category-aware language generation (i.e., chatbots that better interpret user intent by responding with the appropriate level of specificity).

Limitations

1. **Cultural Biases**: Model are trained on English and/or multilingual corpora which may not reflect the lexical preferences of Italian speakers.

2. **Methodology in 4.2**: In the comparison between LLMs and human-generated exemplars, we used a simple string matching, so *abete di Natale* 'Christmas fir' and *abeti di Natale* 'Christmas firs' are considered different strings. While this approach could count good strings as mismatches, the human judgments are manually normalized, and models prefer the singular form consistently. In conclusion, we believe that this approximation does not exclude too many possibly good exemplars.

3. **Exclude GPT from analyses**: We did not use GPT because we cannot access the perplexity values of the model. While some could argue that GPT last models could achieve better performances for the presented tasks, we prefer open models that can be accessed in their internal representations.

Ethical Considerations

- We administrated the exemplars generation task described in §3 to a total of 365 participants (48.5% women; 49.9% man; 1.6% non-binary; M age = 26.3; SD age = 3.76; range age 18-35) on Prolific. All participants were Italian native speakers and reported no language or attentional disorders. Participants were compensated with Euro € 1.80 for generating exemplars in a single list, with an average survey duration of 15 minutes. The data is anonymized to make identification of individuals impossible.

- Since the human data were collected in 2023 and never released, all LLMs have not been exposed to these stimuli, allowing us to test the emerging abilities of these models and their semantic knowledge.

- This research demonstrates the utility of language models as valuable tools in cognitive science and linguistics. However, it is crucial to acknowledge that these models acquire and produce language through mechanisms that differ significantly from human language processing. Consequently, extrapolating these findings directly to human mind organization can lead to potential risks and unintended consequences.

Acknowledgements

We would like to thank the anonymous reviewers for their feedback and comments. AP has been supported by the project "Word Embeddings: From Cognitive Linguistics to Language Engineering, and Back" (WEMB), funded by the Italian Ministry of University and Research (MUR) under the PRIN 2022 funding scheme (CUP B53D23013050006), the PNRR (Prot. IR0000013) "SoBigData.it: Strengthening the Italian RI for Social Mining and Big Data Analytics". GR, CV, and MB have been funded by the European Union (GRANT AGREEMENT: ERC-2021-STG-101039777). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

References

Briony Banks and Louise Connell. 2023. Category production norms for 117 concrete and abstract categories. *Behavior Research Methods*, 55(3):1292–1313.

Lawrence W Barsalou. 1982. Context-independent and context-dependent information in concepts. *Memory & cognition*, 10(1):82–93.

Lawrence W Barsalou. 1983. Ad hoc categories. *Memory & cognition*, 11:211–227.

Lawrence W Barsalou, Ava Santos, W Kyle Simmons, and Christine D Wilson. 2008. Language and simulation in conceptual processing. *Symbols, embodiment, and meaning*, pages 245–283.

Ruairidh M Battleday, Joshua C Peterson, and Thomas L Griffiths. 2020. Capturing human categorization of natural images by combining deep networks and cognitive models. *Nature communications*, 11(1):5418.

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.

Marianna Bolognesi, Christian Burgers, and Tommaso Caselli. 2020. On abstraction: decoupling conceptual concreteness and categorical specificity. *Cognitive Processing*, 21(3):365–381.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. 2022. On the opportunities and risks of foundation models.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Nichol Castro, Taylor Curley, and Christopher Hertzog. 2021. Category norms with a cross-sectional sample of adults in the united states: Consideration of cohort, age, and historical effects on semantic categories. *Behavior research methods*, 53:898–917.

Henri Cohen and Claire Lefebvre. 2005. *Handbook of categorization in cognitive science*. Elsevier.

Charles P Davis and Eiling Yee. 2021. Building semantic memory from embodied and distributional language experience. *Wiley Interdisciplinary Reviews: Cognitive Science*, 12(5):e1555.

Luciano Floridi and Massimo Chiriatti. 2020. Gpt-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30:681–694.

Fritz Günther, Marco Marelli, Sam Tureski, and Marco Alessandro Petilli. 2023. ViSpa (vision spaces): A computer-vision-based representation system for individual images and concept prototypes, with large-scale evaluation. *Psychol. Rev.*, 130(4):896–934.

Lala Hajibayova. 2013. Basic-level categories: A review. *Journal of Information Science*, 39(5):676–687.

Hannes Hansen and Martin N Hebart. 2022. Semantic features of object concepts generated with GPT-3. In *Proceedings of the Annual Meeting of the Cognitive Science Society*.

Tom Heyman and Geert Heyman. 2019. Can prediction-based distributional semantic models predict typicality? *Quarterly Journal of Experimental Psychology*, 72(8):2084–2109. PMID: 30704340.

Tom Heyman and Geert Heyman. 2024. The impact of chatgpt on human data collection: A case study involving typicality norming data. *Behavior Research Methods*, 56(5):4974–4981.

Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. 2021. Perceiver: General perception with iterative attention. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 4651–4664. PMLR.

Miloš Jakubíček, Adam Kilgarriff, Vojtěch Kovář, Pavel Rychlỳ, and Vít Suchomel. 2013. The tenten corpus family. In *7th international corpus linguistics conference CL*, pages 125–127. Valladolid.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, L'elio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts. *ArXiv*, abs/2401.04088.

Carina Kauf, Anna A Ivanova, Giulia Rambelli, Emmanuele Chersoni, Jingyuan Selena She, Zawad Chowdhury, Evelina Fedorenko, and Alessandro Lenci. 2023. Event knowledge in large language models: the gap between the impossible and the unlikely. *Cognitive Science*, 47(11):e13386.

Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. 2024. What matters when building vision-language models?

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, pages 34892–34916.

LlamaTeam. 2024. The llama 3 herd of models.

Max M Louwerse. 2018. Knowing the meaning of a word by the linguistic and perceptual company it keeps. *Topics in cognitive science*, 10(3):573–589.

Gary Lupyan. 2012. Linguistically modulated perception and cognition: The label-feedback hypothesis. *Frontiers in psychology*, 3:54.

Gary Lupyan and Molly Lewis. 2019. From words-as-mappings to words-as-cues: The role of language in semantic knowledge. *Language, Cognition and Neuroscience*, 34(10):1319–1337.

Kyle Mahowald, Anna A. Ivanova, Idan A. Blank, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. 2024. Dissociating language and thought in large language models. *Trends in Cognitive Sciences*, 28(6):517–540.

Gary Marcus. 2020. The next decade in ai: Four steps towards robust artificial intelligence.

Kanishka Misra, Allyson Ettinger, and Julia Rayz. 2021. Do language models learn typicality judgments from text? In *Proceedings of the Annual Meeting of the Cognitive Science Society, 43*, pages 216–222.

Kanishka Misra, Julia Rayz, and Allyson Ettinger. 2023. COMPS: Conceptual minimal pair sentences for testing robust property knowledge and its inheritance in pre-trained language models. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2928–2949, Dubrovnik, Croatia. Association for Computational Linguistics.

Maria Montefinese, Ettore Ambrosini, Beth Fairfield, and Nicola Mammarella. 2012. Semantic memory: A feature-based analysis and new norms for italian. *Behavior Research Methods*, 45:440 – 461.

Gregory Murphy. 2002. *The big book of concepts*. MIT press.

Cristina Izura Natividad Hernández-Muñoz and Andrew W. Ellis. 2006. Cognitive aspects of lexical availability. *European Journal of Cognitive Psychology*, 18(5):730–755.

Animesh Nighojkar, Anna Khlyzova, and John Licato. 2022. Cognitive modeling of semantic fluency using transformers. *arXiv preprint arXiv:2208.09719*.

Joshua C Peterson, Joshua T Abbott, and Thomas L Griffiths. 2018. Evaluating (and improving) the correspondence between deep neural networks and human representations. *Cognitive science*, 42(8):2648–2669.

Joseph Renner, Pascal Denis, Remi Gilleron, and Angèle Brunellière. 2023. Exploring category structure with contextual language models and lexical semantic networks. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2277–2290, Dubrovnik, Croatia. Association for Computational Linguistics.

Eleanor Rosch. 1975. Cognitive representations of semantic categories. *Journal of experimental psychology: General*, 104(3):192.

Eleanor Rosch. 1978. Principles of categorization. *Cognition and categorization/Erlbaum*.

Eleanor Rosch, Carolyn B Mervis, Wayne D Gray, David M Johnson, and Penny Boyes-Braem. 1976. Basic objects in natural categories. *Cognitive psychology*, 8(3):382–439.

Brian H Ross and Gregory L Murphy. 1999. Food for thought: Cross-classification and category organization in a complex real-world domain. *Cognitive psychology*, 38(4):495–553.

Prisha Samadarshi, Mariam Mustafa, Anushka Kulkarni, Raven Rothkopf, Tuhin Chakrabarty, and Smaranda Muresan. 2024. Connecting the dots: Evaluating abstract reasoning capabilities of llms using the new york times connections word game. *arXiv preprint arXiv:2406.11012*.

Vít Suchomel, Jan Pomikálek, et al. 2012. Efficient web crawling for large text corpora. In *Proceedings of the seventh Web as Corpus Workshop (WAC7)*, pages 39–43.

Neha Upadhyay, Kritika Mittal, and Sashank Varma. 2022. Typicality gradients in computer vision models. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 44.

Rens van Hoef, Louise Connell, and Dermot Lynott. 2023. The effects of sensorimotor and linguistic information on the basic-level advantage. *Cognition*, 241:105606.

Siddhartha K Vemuri, Raj Sanjay Shah, and Sashank Varma. 2024. How Well Do Deep Learning Models Capture Human Concepts? The Case of the Typicality Effect. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 46:5160–5167.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11975–11986.

A Study 1

A.1 Metrics

In this section, we define the metrics described in Section 3 used to evaluate the exemplars obtained from human participants.

Exemplar Dominance

$$ED(E) = P(E|C) = \frac{N(E \cap C)}{N(C)} \tag{1}$$

where $N(E \cap C)$ is equal to the number of participants who produced the exemplar E when in response to the concept C, and $N(C)$ is the number of participants elicited by C.

Mean Rank Order

$$MRO(E) = \frac{\sum^{N(C)} r_i(E|C)}{N(C)} \tag{2}$$

First Occurrence Value

$$FOV(E, C) = \frac{N_{first}(E)}{N(C)} \tag{3}$$

Exemplar Availability

$$EA(E,C) = \sum_{p=1}^{n} \frac{f_{pi}}{N} \cdot e^{[-2.3 \cdot (\frac{p-1}{n-1}])} \tag{4}$$

where p is the rank of the produced exemplar E, n is its lowest rank obtained across multiple participants, f_{pi} is the number of participants who produced the exemplar i at the same position p, and N is the total number of participant who have seen the category C.

B Study 2

B.1 Models Description

In this section, we provide the details on the pre-trained language models listed in §4. All models are open-source and available via huggingface[5].

B.2 Unimodal Language Models

LLaMA-3.1 (LlamaTeam, 2024) is a collection of pre-trained auto-regressive large language models openly released by Meta AI. In our experiments, we rely on the instruction-based version, which are fine-tuned for dialogue use case with multilingual input. We assess performance of both the small version (8B parameters[6]) and the larger one (with 70B parameters[7]). We avoid testing the extra-large version (405B parameters) due to computational constraints. All models are first pre-trained (SFT) on a mix of publicly available online data and further aligned with human preferences via RLHF.

LLaMA-3.2 is the next iteration of llama models. With respect to version 3.1, they differ in models sizes (1B, 3B, 11B, and 90B parameters) and multimodal capabilities. However, at the moment of writing, the multimodal version of llama-3.2 is not accessible in the EU, due to European regulations[8]. For this reason, we are not able to provide any insight about the multimodal versions. Concerning the assessed version, we limit ourselves to the small (3B) model.[9]

Mistral (Jiang et al., 2023) is a pre-trained auto-regressive large language model released by Mitral AI[10]. The model leverages Grouped-Query Attention and Sliding Windows Attention to improve inference time and memory requirements, and to enable handling longer input sequences.

Mistral-8x7B is an ensemble mixture of experts model[11] of eight 7B parameter models developed by Mistral AI. The individual models are trained with Grouped-Query Attention (GQA) and Sliding

[5] https://huggingface.co/

[6] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

[7] https://huggingface.co/meta-llama/Llama-3.1-70B-Instruct

[8] https://huggingface.co/meta-llama/Llama-3.2-11B-Vision-Instruct

[9] https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct

[10] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2

[11] https://huggingface.co/mistral/Mistral-7B-Instruct

Window Attention (SWA) mechanisms, enabling efficient handling of long sequences and improving inference speed. A routing system takes care of distributing the input to the appropriate experts. This mechanism increases the number of parameters of a model while controlling cost and latency, as the model only uses a fraction of the total set of parameters per token. For our experiments, we use the standard instruction-tuned version of Mistral-7B, focusing on its capacity for multilingual inputs and dialogue generation.

NeMo is a 12B model[12] designed for multilingual applications. It is trained on function calling, has a large context window, and is particularly strong in English, French, German, Spanish, Italian, Portuguese, Chinese, Japanese, Korean, Arabic, and Hindi. Mistral NeMo uses a new tokenizer, Tekken, based on Tiktoken, that was trained on over more than 100 languages, and compresses natural language text and source code more efficiently than the SentencePiece tokenizer used in previous Mistral models.

B.3 Multimodal Language Models

LLaVA (Liu et al., 2023) is a multimodal model that integrates visual understanding with language capabilities by combining a vision encoder (e.g., CLIP's Vision Transformer) with a large language model (e.g., LLaMA). It is designed for open-ended vision-language tasks, such as image captioning, visual question answering, and reasoning about images. The model is trained following a two-stage training approach: first, the vision and the language encoders are aligned by training a projection layers that map visual features into the LLM's embedding space. Second, the model undergoes an instruction tuning phase, using curated vision-language datasets to improve coherence and accuracy in responses.

Idefics2 (Laurençon et al., 2024) is the result of a throughout ablation of the design choices available for vLMs pre-training. To encode visual features in the LLM's embedding space, Idefics2 leverages a SigLIP's vision encoder (Zhai et al., 2023) followed by a learned Perceiver pooling (Jaegle et al., 2021) and an multi-layer perceptron projection. The pooled sequence is then concatenated with the text embeddings to obtain an interleaved sequence of images and texts. The model is trained according to the usual vLMs pipeline, with a first stage focusing on the alignment of the two modality embedders, followed by a second instruction-tuning stage.

B.4 Perplexity Computation

Perplexity is computed according the following formula:

$$\mathrm{PPL}(X) = \exp\left\{ \frac{1}{t} \sum_{i}^{t} \log p_\theta(x_i \mid x_{<i}) \right\} \quad (5)$$

where x_i is the target expression (i.e., either the basic or superordinate category, in SUBTASK A, or the subordinate level exemplar, in SUBTASK B) and $x_{<i}$ is the fixed prompt. In our settings, this is equivalent to the exponentiation of the cross-entropy loss. We compute the perplexity for the target tokens only (x_i), and mask the non-target tokens ($x_{(<i)}$) accordingly. Notice that in our experiments the perplexity is used to compare output of the same model, therefore normalization is not required to compare the binary-accuracy score (i.e., the evaluation metrics for SUBTASK A and B).

B.5 Prompting Strategy

To obtain a list of exemplars (i.e., basic-level concepts) from a LLMs, we use the following Italian prompt:

```
<s>[INST] Data una parola che
denota una concetto, elenca
tutta i ‘tipi di’ quel concetto.
Elenca solo i nomi delle entità.
Per esempio per il concetto
‘elettrodomestico’ elenca:
frullatore, aspirapolvere,
tostapane, lavatrice. Ora
fai lo stesso per il concetto
‘<CONCEPT>’ [/INST] Questa è una
lista ’tipi di’ che appartengono
al concetto ‘<CONCEPT>‘:
```

where `<CONCEPT>` is replaced with the eliciting concept. For the non-Italian reader, we provide an English translation of previous prompt:

```
<s>[INST] Given a word denoting a
concept, list all of the ‘kinds
of’ of the given concept. List
only words denoting entities.
For example, for the concept
```

[12] https://huggingface.co/mistralai/Mistral-Nemo-Base-2407

```
‘electric appliance‘ list:
‘mixer’, ‘vacuum cleaner’,
‘toaster’, ‘washing machine’.
Now do the same for the concept
‘<CONCEPT>’:
```

B.6 Model-specific sampling parameters

Regarding hyperparameters, we set `top-p` to 0.75 to limit the long tail of low-probability tokens that may be sampled, while `frequency` and `repetition penalty` are set to 0.

B.7 Top-5 Matches

Table 6 shows the percentage of matches among the top-5 human- produced and LLMs-generated exemplars, reporting individual accuracy for each of the 12 superordinate categories.

B.8 Generated Exemplars and Hallucinations

In Table 7 we report the exemplars generated by the LLaMA-3.1-70B, the best-performing model, for the 12 superordinate categories. For each of the 12 superordinate categories, we select the basic-level concept for which humans have generated the greatest amount of exemplars. In Table 8, we report the exemplars generated for the 12 basic-level concepts that produced the greatest amount of unattested occurrences according to the Italian Corpus ItTenTen.

In our study, we automatically identify low-frequency occurring terms via the Italian corpus ItTenTen. By analyzing exemplars with an absolute frequency equal to zero we can gain a deeper insight regarding hallucination generation in the exemplars generation task. We divide unattested exemplars into false negatives (e.g., exemplars for which we retrieved a zero frequency due to misspellings or morphosyntactical issues) and hallucinations. Through qualitative analysis, we observe several recurring patterns and categorize most of the hallucinations into the following groupings: *ad-hoc instances*, *nonsensical*, *foreign-language based*, *conceptual confusion*, and *imitation-based.*

***Ad-hoc* Instances:** These instances reflect the model’s ability to creatively compose category-consistent yet ungrounded expressions, relying on syntactic and semantic cues rather than empirical knowledge. As such, ad hoc constructions are generated “on the fly” to fit perceived communicative goals, but lack the frequency-based support or conventionalization required to qualify as exemplars stored in long-term memory. Some examples are: MAGLIA ‘a punto catenella’ (chain stitched KNITWEAR), ‘a punto scritto a rombi’ (diamond shape stitched KNITWEAR), GALLO ‘della giungla verde’ (COCK of the green jungle), ‘della giungla rosso’ (red COCK of the jungle), CASSETTIERA ‘per giocattoli’ (toy DRAWER), or CASSETTIRA ‘per attrezzi’ (tool DRAWER), ‘da corridoio’ (hallway DRAWER).

Nonsensical: Expressions that are grammatically well-formed but semantically incoherent, implausible, or internally contradictory, often resulting from incongruous or incompatible feature combinations. Some examples are: GERANIO ‘a foglie di quercia’ (GERANIUM with oak leaves), ‘a foglie di rosmarino’ (GERANIUM with rosemary leaves). CRUCIVERBA ‘a parole sovrapposte’ (CROSSWORD with overlapping words), ‘a parole crociate’ (CROSSWORD with word crossed). TRATTORE ‘a cingoli in acciaio’ (TRACTOR with steel tank track). GALLO ‘cedrone giapponese’ (Japanese capercaillie COCK).

Foreign-Language Based: Refers to expressions that denote a real-world referent conceptualized in a foreign language with respect to Italian. For example, GALLO ‘di Crèvecœur’ (Crèvecœur CHICKEN) has no attested translation in Italian.

Conceptual Confusion: Cases in which the model misinterprets the intended sense or category of a lexical item, leading to the generation of exemplars that belong to a different semantic domain. For example, when prompted with *margherita* as a flower (i.e., ‘daisy’), the model generates *d’Austria* (‘of Austria’), referencing Margherita d’Austria (Margaret of Parma, a historical figure[13]), and *d’Ungheria* (‘of Hungary’), referencing Margherita d’Ungheria (Saint Margaret of Hungary[14]).

Imitation Based: In this case, LLMs replicate the surface-level syntactic or morphological structure of a valid, attested exemplar, leading to the overgeneralization of that structure across subsequent, unattested or spurious exemplars. This imitation is often form-driven rather than grounded in semantic plausibility or real-world usage. This phenomenon typically arises when a salient exemplar

[13] https://en.wikipedia.org/wiki/Margaret_of_Parma

[14] https://en.wikipedia.org/wiki/Margaret_of_Hungary_(saint)

	ANIMALS	BODY PARTS	CLOTHES	FOODS	FURNISHING	FURNITURE	HOBBIES	HOUSING	KITCHEN	PLANTS	STATIONERY	VEHICLES	avg
`llama-3.2-3B`	0.24	0.13	0.09	0.33	0.11	0.13	0.08	0.10	0.12	0.06	0.06	0.21	0.14
`llama-3.1-8B`	0.32	0.13	0.09	0.36	0.20	0.24	0.18	0.19	0.17	0.15	0.13	0.25	0.20
`llama-3.1-70B`	0.35	0.12	0.15	0.29	0.19	0.28	0.23	0.19	0.15	0.18	0.18	0.18	0.21
`mistral-7B`	0.25	0.14	0.07	0.24	0.03	0.09	0.14	0.13	0.13	0.16	0.08	0.13	0.13
`nemo-12B`	0.36	0.16	0.26	0.37	0.16	0.31	0.26	0.18	0.23	0.24	0.25	0.15	**0.24**
`mixtral-8x7B`	0.27	0.11	0.18	0.25	0.19	0.20	0.23	0.19	0.18	0.18	0.21	0.14	0.19
`llava-7B`	0.32	0.09	0.11	0.28	0.06	0.11	0.15	0.10	0.10	0.22	0.10	0.15	0.15
`idefics2-8B`	0.25	0.09	0.06	0.25	0.03	0.04	0.11	0.03	0.03	0.10	0.03	0.21	0.10
category avg	**0.29**	0.12	0.12	**0.29**	0.12	0.17	0.17	0.13	0.13	0.16	0.13	0.17	0.17

Table 6: Percentage of matches among **top five** most available exemplars.

introduces a productive or familiar template, which the model then extends combinatorially without regard for corpus evidence or conceptual appropriateness. For instance, the attested exemplar TERRAZZO ‘alla veneziana‘ (Venetian PAVEMENT) serves as a template `NOUN + ADJECTIVE (ITALIAN LOCATION)` for generating further expressions like *terrazzo genovese, milanese, bergamasca, pavese, fiorentina*, none of which are attested or conventional within the category. Similarly, for the concept CANDELABRO ‘a 5 bracci/a‘, the syntactical structure ‘a N bracci/a’ is reiterated multiple times with increasing numbers of arms.

C Subtask A

In this section, we report the in-depth results for the experiment described in Section 5.1. Tables 9a and 9b report individual accuracy for each of the 12 superordinate categories for basic-level and superordinate-level category prediction, respectively.

D Subtask B

In the following tables, we report individual accuracy for each of the 12 superordinate categories for Subtask B. Results are grouped into three blocks according to the number of exemplars generated by the human subjects: (i) low coverage (up to 5 exemplars; Table 10a), (ii) medium coverage (6–10 exemplars; Table 10b), and (iii) high coverage (more than 10 exemplars; Table 10c). Note that the columns containing ‘na’ values are the results of the frequency-based grouping. For example, we do not have any basic-level concept belonging to the super-ordinate category of `plants` that elicited a **high** number of exemplars in the human experimental phase. Hence, the empty column in Tables 10a and 10c.

D.1 SUBTASK B: Typicality Variation by Availability Score

In this Section, we report the results for the typicality prediction experiments described in Section 5.2 by aggregating the results along the availability score. Specifically, we group results according to the absolute difference between the availability score of the most-available exemplars and the availability score of the least-available one. The availability score is computed on the human experiment’s results.

	ANIMALS **Cane** (dog)	BODY PARTS **Capelli** (hair)	CLOTHES **Scarpa** (shoe)	FOODS **Pasta** (pasta)	FURNISHING **Vaso** (vase)	FURNITURE **Sedia** (chair)
1	pastore tedesco	riccio	stivale	spaghetti	di fiori	poltrona
2	segugio	ricci	sandalo	fettuccine	di rame	a dondolo
3	rottweiler	afro	ciabatta	penne	di cristallo	a rotelle
4	alano	ondulato	da ballo	farfalle	di ceramica	sgabello
5	dobermann	crespo	anfibio	tortellini	di porcellana	pieghevole
6	levriero	riccio afro	stivaletto	rigatoni	di terracotta	sdraio
7	corso	liscio lucido	da trekking	cannelloni	di vetro	da giardino
8	pinscher	liscio	da ginnastica	ravioli	di metallo	da ufficio
9	boxer	ondulati	da calcio	gnocchi	urna	a sdraio
10	beagle	crespi	da tennis	maccheroni	di legno	da bar
11	poodle	ricciolino	zoccolo	lasagne	di marmo	da ristorante
12	pug	liscio opaco	da sci	vermicelli	di plastica	da spiaggia
13	dalmata	mosso	mocassino	tagliatelle	di argento	reclinabile
14	bulldog	mossi	da ciclismo	fusilli	di oro	per bambini
15	lupo	lisci	da danza classica	linguine	di ottone	a sacco
16	shih tzu	crespo lucido	da basket	ditalini	di pietra	a schienale alto
17	pitbull	ondulato lucido	da neve	pappardelle	di argilla	a schienale basso
18	basset hound	mosso lucido	da danza	orecchiette	di notte	pouf
19	chihuahua	riccio afro lucido	da calcetto	conchiglie	di bronzo	panchina
20	collie	riccio lucido	da equitazione	lasagna	greco	a braccioli

	HOBBIES **Libro** (book)	HOUSING **Camera** (room)	KITCHEN **Pentola** (pan)	PLANTS **Margherita** (daisy)	STATIONERY **Foglio** (sheet)	VEHICLES **Automobile** (car)
1	romanzo	da letto	a pressione	comune	di carta	berlina
2	saggio	oscura	casseruola	dei prati	di via	autocarro
3	dizionario	di equilibrio	di ghisa	di savoia	di alluminio	autobus
4	atlante	d'albergo	padella	pizza	di rame	camion
5	enciclopedia	iperbarica	marmitta	di lorena	di calcolo	minivan
6	agenda	di sicurezza	wok	di angoulême	di stile	monovolume
7	manuale	di combustione	di acciaio	di borgogna	elettronico	spider
8	almanacco	a gas	a vapore	di fiandra	di ruta	suv
9	biografia	di commercio	di rame	di scozia	di plastica	pick-up
10	fumetto	di decompressione	di terracotta	di provenza	di piombo	cabriolet
11	trattato	di scoppio	di ceramica	di parma	di registro	station wagon
12	diario	di refrigerazione	di alluminio	di valois	di viti	fuoristrada
13	catalogo	di consiglio	antiaderente	tudor	di stagno	furgone
14	novella	di compensazione	di pietra ollare	d'ungheria	di ottone	citycar
15	autobiografia	di controllo	coccotte	a fiori doppi	di rame berillifero	furgoncino
16	compendio	di manovra	elettrica	d'austria	di rame fosforoso	coupé
17	raccolta	di decantazione	in rame	a fiori giganti	di rame arsenicale	hatchback
18	per bambini	di carico	in pietra ollare	a fiori piccoli	di stagnola	pulmino
19	racconto	di regia	paiolo	a fiori colorati	di alluminio stagnato	autovettura
20	monografia	mortuaria	calderone	a fiori bianchi	di lavoro	roadster

Table 7: Up to 20 exemplars generated by LLaMA-3.1-70B (the best-performing model in terms of valid exemplars generated), sorted by availability score. For each of the 12 superordinate categories (in UPPERCASE), we select the basic-level category (in **bold**) for which humans have generated the greatest amount of exemplars. Cells with a **light-blue background** indicate exemplars not produced by the human study group but still considered valid, with more than 15 occurrences in the ItTenTen corpus. Exemplars with lower frequency are denoted by a **light-yellow background**. A **light-red background** indicates unattested exemplars, which are regarded as hallucinations.

	ANIMALS	BODY PARTS	CLOTHES	FOODS	FURNISHING	FURNITURE
	Gallo (cock)	**Spalla** (shoulder)	**Maglia** (sweater)	**Latte** (milk)	**Candelabro** (candelabra)	**Cassettiera** (dresser)
1	cedrone	a sbalzo	a coste	di cocco	da tavolo	da ufficio
2	bankiva	a volant	a punto croce	di soia	a sospensione	da cucina
3	silvestre	a bretella	a righe	di capra	da terra	da bagno
4	nero	a bretelle	a losanghe	di mucca	da parete	da notte
5	di banca	a botte	rasata	di pecora	a 5 bracci	da camera da letto
6	di wallich	a spigolo	a uncinetto	di bufala	a 3 braccia	per giocattoli
7	da combattimento	a pizzo	a punto	di avena	a stelo	da ingresso
8	di sonnerat	a cuscino	a tubolare	di arachidi	a 5 braccia	da comodino
9	della giungla	all’americana	a rombi	di mandorla	a 7 braccia	per attrezzi
10	cedrone giapponese	a kimono	a cavi	di riso	a 9 braccia	da scrivania
11	di faverolles	a punta	a fantasia	di cammello	a 7 bracci	per oggetti di cancelleria
12	di houdan	a sbuffo	a doppia punta	di nocciole	a 9 bracci	da corridoio
13	della malesia	a frangia	a punto catenella	di anacardi	a 11 bracci	da esterno
14	della giungla grigio	a pizzo di sanok	a punto lino	di quinoa	a 13 bracci	da soggiorno
15	di crèvecoeur	a pizzo di lefkara	a punto scritto	di mandorle	da mensola	
16	della giungla verde	a latticciolo	a punto raso	di orzo	da camino	
17	della giungla rosso	a piquet	a punto legaccio	di semi di lino	da altare	
18	di jungla	a pizzo di burano	a punto scritto a rombi	di grano	tripode	
19	di borneo	a pizzo di gorizia	a punto reale	di semi di sesamo	da chiesa	
20	di delacour	a pizzo ricamato	a punto rovescio	di semi di girasole	a più bracci	

	HOBBIES	HOUSING	KITCHEN	PLANTS	STATIONERY	VEHICLES
	Cruciverba (crossword)	**Terrazzo** (terrace)	**Mestolo** (ladle)	**Gernaio** (geranium)	**Colla** (glue)	**Trattore** (tractor)
1	classico	alla veneziana	forato	a foglie di quercia	a caldo	agricolo
2	enigmistica	a sbalzo	da minestra	geranium maculatum	vinilica	cingolato
3	a schema variabile	alla romana	lungo	a fioritura continua	a freddo	a ruote
4	a schema fisso	pensile	a buco	a foglie di vite	a base di lattice	stradale
5	a schema libero	alla genovese	da zuppa	a foglie di betulla	a base di gomma	a cingoli in gomma
6	per bambini	alla milanese	da cucina	a foglie di rosmarino	a base di resina	articolato
7	tematico	alla bergamasca	da gelato	a foglie di alloro	a base di silicio	a cingoli
8	per adulti	a livello	per mescolare	a foglie di felce	epossidica	telescopico
9	a schema personalizzato	alla pavese	per servire	a foglie di platano	per legno	a quattro ruote motrici
10	a schema geometrico	panoramico	per gelato	a fioritura estiva	per carta	a ruote motrici
11	a schema logico	coperto	da legno	a fioritura primaverile	per plastica	a cingoli motrici
12	a schema numerico	fiorito	da salsa	d’appartamento	acrilica	agricolo cingolato
13	a parole sovrapposte	scoperto	da metallo	d’altura	a base di silano	a cingoli motrici 4x2
14	a parole nascoste	giardino	da risotto	cespuglioso	a base di silice	a cingoli motrici 4x4
15	a parole crociate	solarium	da silicone	bicolore	per tessuti	a ruote motrici 4x2
16	a definizioni consecutive	adiacente	per impastare	aquilegifolium	a base di silicato	a ruote motrici 4x4
17	a definizioni incrociate	alla fiorentina	per dosare	annuale	per metalli	a ruote anteriori sterzanti
18	a tema libero		per condire	alpino	per vetro	a ruote posteriori sterzanti
19	a figure		cucchiaio	a foglia rossa	a base di solvente	a due ruote motrici
20	con immagini		a nido d’ape	geranium phaeum	a base d’acqua	a cingoli in acciaio

Table 8: Up to 20 exemplars generated by LLaMA-3.1-70B (the best-performing model in terms of valid exemplars generated), sorted by availability score. We select the basic-level categories that produced the highest number of hallucinations, i.e., expressions unattested in the ItTenTen corpus. For the colouring rationale, see Table 7.

	ANIMALS	BODY PARTS	CLOTHES	FOODS	FURNISHING	FURNITURE	HOBBIES	HOUSING	KITCHEN	PLANTS	STATIONERY	VEHICLES	avg
llama-3.2-3B	0.76	0.81	0.82	0.67	0.95	0.92	0.87	0.73	0.83	0.94	0.94	0.8	0.84
llama-3.1-8B	1.0	0.94	1.0	0.93	1.0	0.92	0.93	0.93	0.92	1.0	1.0	1.0	0.96
llama-3.1-70B	1.0	0.94	0.94	1.0	1.0	0.92	0.93	0.93	0.92	0.94	1.0	0.93	0.95
mistral-7B	0.94	1.0	0.76	0.87	1.0	0.92	0.93	0.8	0.75	0.94	0.88	0.93	0.89
nemo-12B	0.94	1.0	1.0	1.0	1.0	0.83	1.0	1.0	0.92	0.88	0.94	0.93	0.95
mixtral-8x7B	0.94	1.0	0.94	1.0	1.0	1.0	1.0	1.0	1.0	0.94	1.0	1.0	**0.98**
llava-7B	0.94	1.0	0.82	0.67	1.0	0.92	0.93	0.93	1.0	0.94	1.0	1.0	0.93
idefics2-8B	0.88	1.0	0.88	0.8	1.0	0.92	1.0	0.93	1.0	0.94	1.0	0.93	0.94
category avg	0.93	0.96	0.90	0.87	**0.99**	0.92	0.95	0.91	0.92	0.94	0.97	0.94	0.93

(a) Accuracy for **basic-level category** prediction.

	ANIMALS	BODY PARTS	CLOTHES	FOODS	FURNISHING	FURNITURE	HOBBIES	HOUSING	KITCHEN	PLANTS	STATIONERY	VEHICLES	avg
llama-3.2-3B	0.94	0.12	0.71	0.07	0.0	0.75	0.07	0.8	1.0	0.81	0.0	0.93	0.52
llama-3.1-8B	1.0	0.81	0.76	0.2	0.0	0.92	0.13	0.8	1.0	0.94	0.0	1.0	0.63
llama-3.1-70B	1.0	0.69	0.35	0.4	0.0	1.0	0.07	0.93	1.0	0.88	0.44	0.93	**0.64**
mistral-7B	0.94	0.62	0.94	0.33	0.32	0.92	0.0	0.4	1.0	0.56	0.0	1.0	0.59
nemo-12B	0.06	0.81	0.12	0.0	0.0	1.0	0.07	0.2	1.0	0.75	0.5	1.0	0.46
mixtral-8x7B	1.0	0.94	0.06	0.47	0.0	0.83	0.13	0.6	1.0	0.75	0.06	1.0	0.57
llava-7B	0.88	0.88	0.76	0.33	0.11	0.83	0.13	0.67	1.0	0.5	0.0	1.0	0.59
idefics2-8B	0.88	0.0	0.12	0.6	0.0	0.67	0.0	0.53	1.0	0.06	0.0	0.67	0.38
category avg	0.84	0.61	0.48	0.30	0.05	0.86	0.08	0.62	**1.00**	0.66	0.12	0.94	0.53

(b) Accuracy for **superordinate category** prediction.

Table 9: SUBTASK A–Accuracy for category prediction at basic and super-ordinate category level.

	ANIMALS	BODY PARTS	CLOTHES	FOODS	FURNISHING	FURNITURE	HOBBIES	HOUSING	KITCHEN	PLANTS	STATIONERY	VEHICLES	avg
llama-3.2-3B	0.38	0.60	na	na	0.60	0.50	0.67	0.75	na	0.77	1.00	0.60	0.65
llama-3.1-8B	0.38	0.80	na	na	0.60	0.25	0.33	0.75	na	0.54	1.00	0.60	0.58
llama-3.1-70B	0.38	1.00	na	na	0.80	0.75	0.67	0.75	na	0.85	1.00	0.40	**0.73**
mistral-7B	0.62	0.60	na	na	0.80	0.50	0.33	0.00	na	0.54	0.50	0.60	0.50
nemo-12B	0.25	0.40	na	na	0.60	0.50	0.67	0.75	na	0.69	0.50	0.40	0.53
mixtral-8x7B	0.50	1.00	na	na	0.80	0.50	0.67	1.00	na	0.69	0.50	0.80	**0.72**
llava-7B	0.75	0.40	na	na	0.80	0.25	0.33	0.25	na	0.46	0.50	0.60	0.48
idefics2-8B	0.62	0.40	na	na	0.80	0.50	0.33	0.00	na	0.54	1.00	0.60	0.53
category avg	0.48	0.65	na	na	0.72	0.47	0.50	0.53	na	0.63	**0.75**	0.57	0.59

(a) **Low coverage** basic-level categories.

	ANIMALS	BODY PARTS	CLOTHES	FOODS	FURNISHING	FURNITURE	HOBBIES	HOUSING	KITCHEN	PLANTS	STATIONERY	VEHICLES	avg
llama-3.2-3B	0.50	0.33	1.00	0.00	0.67	1.00	0.75	0.67	0.67	0.80	0.67	0.33	0.62
llama-3.1-8B	0.75	0.44	0.80	0.50	0.44	0.60	0.75	0.83	0.33	0.60	0.78	0.33	0.60
llama-3.1-70B	0.75	0.33	1.00	1.00	0.67	0.60	0.75	0.67	0.33	0.80	0.78	0.50	**0.68**
mistral-7B	0.50	0.44	0.80	1.00	0.56	0.60	0.75	0.67	0.50	0.20	0.33	0.50	0.57
nemo-12B	0.75	0.56	0.80	1.00	0.67	1.00	0.75	0.50	0.67	0.80	0.44	0.33	**0.69**
mixtral-8x7B	0.75	0.56	1.00	0.50	0.67	0.40	0.75	0.33	0.67	0.20	0.33	0.50	0.55
llava-7B	0.25	0.33	0.80	1.00	0.56	0.80	0.75	0.67	0.33	0.80	0.44	0.67	0.62
idefics2-8B	0.25	0.22	0.80	1.00	0.67	0.80	0.75	0.50	0.17	0.80	0.44	0.50	0.58
category avg	0.56	0.40	**0.88**	0.75	0.61	0.72	0.75	0.60	0.46	0.62	0.53	0.46	0.61

(b) **Medium coverage** basic-level categories.

	ANIMALS	BODY PARTS	CLOTHES	FOODS	FURNISHING	FURNITURE	HOBBIES	HOUSING	KITCHEN	PLANTS	STATIONERY	VEHICLES	avg
llama-3.2-3B	0.60	0.50	0.58	0.62	0.20	0.33	0.38	0.40	0.17	na	0.40	0.50	0.42
llama-3.1-8B	0.60	0.00	0.42	0.54	0.60	0.00	0.50	0.40	0.50	na	0.60	0.50	0.42
llama-3.1-70B	0.40	1.00	0.83	0.77	0.60	0.67	0.50	0.40	0.67	na	0.40	0.50	**0.61**
mistral-7B	0.20	0.50	0.50	0.62	0.40	0.33	0.38	0.40	0.67	na	0.40	0.75	0.47
nemo-12B	0.60	1.00	0.33	0.69	0.80	0.33	0.12	0.40	0.50	na	0.40	0.50	0.52
mixtral-8x7B	0.80	0.00	0.50	0.62	0.80	0.67	0.25	0.20	0.67	na	0.80	1.00	0.57
llava-7B	0.60	0.50	0.42	0.77	0.40	0.33	0.25	0.40	0.50	na	0.40	0.75	0.48
idefics2-8B	0.80	0.50	0.42	0.69	0.40	0.00	0.25	0.40	0.33	na	0.40	0.75	0.45
category avg	0.57	0.50	0.50	**0.66**	0.52	0.33	0.33	0.38	0.50	na	0.48	**0.66**	0.49

(c) **High coverage** basic-level categories.

Table 10: SUBTASK B–Typicality Accuracy at **different coverage** of basic-level categories.

	ANIMALS	BODY PARTS	CLOTHES	FOODS	FURNISHING	FURNITURE	HOBBIES	HOUSING	KITCHEN	PLANTS	STATIONERY	VEHICLES	avg
llama-3.2-3B	0.50	0.33	0.75	0.71	0.80	1.00	0.57	0.67	0.38	1.00	0.60	0.00	0.61
llama-3.1-8B	0.75	0.33	0.75	0.71	0.80	0.50	0.57	0.67	0.50	0.50	0.80	0.00	0.57
llama-3.1-70B	0.25	0.67	1.00	1.00	1.00	1.00	0.71	0.67	0.50	1.00	0.60	0.00	**0.70**
mistral-7B	0.25	0.67	0.50	0.86	0.60	1.00	0.71	0.50	0.62	0.00	0.60	0.00	0.53
nemo-12B	0.50	1.00	0.75	0.71	1.00	1.00	0.71	0.67	0.75	1.00	0.20	0.00	0.69
mixtral-8x7B	0.75	0.33	0.75	0.86	1.00	1.00	0.71	0.50	0.62	0.50	0.80	0.50	0.69
llava-7B	0.50	0.33	0.50	0.86	0.80	1.00	0.57	0.67	0.50	0.50	0.60	0.50	0.61
idefics2-8B	0.50	0.67	0.50	0.71	0.80	0.50	0.57	0.33	0.25	0.50	0.60	0.00	0.49
category avg	0.50	0.54	0.69	0.80	0.85	**0.88**	0.64	0.58	0.52	0.62	0.60	0.12	0.61

(a) **High absolute difference** in availability score ($|\Delta| > 0.4$).

	ANIMALS	BODY PARTS	CLOTHES	FOODS	FURNISHING	FURNITURE	HOBBIES	HOUSING	KITCHEN	PLANTS	STATIONERY	VEHICLES	avg
llama-3.2-3B	0.50	0.60	0.86	0.38	0.38	0.60	0.40	0.80	0.50	0.67	0.71	0.57	0.58
llama-3.1-8B	0.50	0.60	0.57	0.38	0.50	0.20	0.60	0.80	0.25	0.33	0.71	0.57	0.50
llama-3.1-70B	0.50	0.70	0.86	0.62	0.62	0.40	0.60	0.60	0.50	0.67	0.71	0.57	**0.61**
mistral-7B	0.50	0.50	0.57	0.50	0.62	0.40	0.40	0.40	0.50	0.33	0.29	0.86	0.49
nemo-12B	0.50	0.50	0.29	0.75	0.62	0.60	0.20	0.40	0.25	0.67	0.57	0.57	0.49
mixtral-8x7B	0.50	0.70	0.71	0.38	0.88	0.40	0.40	0.40	0.75	0.33	0.29	0.86	0.55
llava-7B	0.50	0.50	0.29	0.75	0.50	0.60	0.20	0.40	0.25	0.33	0.43	0.86	0.47
idefics2-8B	0.50	0.30	0.43	0.75	0.75	0.40	0.20	0.40	0.25	0.33	0.43	0.86	0.47
category avg	0.50	0.55	0.57	0.56	0.61	0.45	0.38	0.52	0.41	0.46	0.52	**0.71**	0.52

(b) **Medium absolute difference** in availability score ($0.2 \leq |\Delta| \leq 0.4$).

	ANIMALS	BODY PARTS	CLOTHES	FOODS	FURNISHING	FURNITURE	HOBBIES	HOUSING	KITCHEN	PLANTS	STATIONERY	VEHICLES	avg
llama-3.2-3B	0.43	0.00	0.50	na	0.50	0.60	0.67	0.25	na	0.73	0.50	0.50	0.47
llama-3.1-8B	0.43	0.33	0.33	na	0.33	0.40	0.33	0.50	na	0.64	0.75	0.50	0.45
llama-3.1-70B	0.57	0.33	0.83	na	0.50	0.80	0.33	0.50	na	0.82	0.75	0.50	**0.59**
mistral-7B	0.57	0.33	0.67	na	0.50	0.40	0.00	0.25	na	0.64	0.25	0.50	0.41
nemo-12B	0.43	0.33	0.50	na	0.50	0.60	0.00	0.50	na	0.64	0.50	0.33	0.43
mixtral-8x7B	0.71	0.67	0.50	na	0.33	0.40	0.00	0.50	na	0.64	0.50	0.67	0.49
llava-7B	0.71	0.00	0.83	na	0.50	0.20	0.33	0.25	na	0.64	0.25	0.50	0.42
idefics2-8B	0.71	0.00	0.67	na	0.33	0.60	0.33	0.25	na	0.73	0.50	0.50	0.46
category avg	**0.57**	0.25	0.60	na	0.44	0.50	0.25	0.38	na	0.68	0.50	0.50	0.47

(c) **Low absolute difference** in availability score ($|\Delta| < 0.2$).

Table 11: SUBTASK B–Typicality Accuracy at **different availability score** of exemplars.