Finding the Conceptual Glue: Ad-hoc Categorization in People and LLMs

Aja Altenhof*
altenhof@wisc.edu
Linguistics, University of Wisconsin-Madison

Kira Breeden*
kbreeden@wisc.edu
Psychology, University of Wisconsin-Madison

Gary Lupyan
glupyan@wisc.edu
Psychology, University of Wisconsin-Madison

Abstract

What do "signal," "heat," "contact," and "director" have in common? Terms in electrical engineering? Things needed on a movie set? Appreciating what these concepts have in common requires coming up with a situation that provides a coherent grouping for them. These "ad-hoc" categories are constructed on the spot, but what makes some better than others? Humans and large language models (LLMs) generated categories (e.g. "Things with corners") for randomly grouped sets of nouns (e.g., "eye," "road," "table," "paper") and rated how difficult it was to link them. We then asked additional participants to rate applicability and specificity of these categories. LLMs produced more specific and applicable categories and the model and human performance gap increased for less semantically similar categories. Models, without human cognitive constraints, may be able to generate more candidates before selecting the best one, while people settle on what's "good enough."

Keywords: category generation, alignment, Large Language Models

Introduction

What do "security," "system," "information," and "order" have in common? Things relevant to an HR person? Buzz words for government propaganda? The answer isn't obvious. It's difficult to come up with a "good" option to link these words. What makes a connection "good" in the first place?

Categorization begins with a basic problem: why do some things go together? One influential view emphasizes similarity: category members share features, and new items are categorized based on how similar they are to known exemplars (Shepard, Hovland, & Jenkins, 1961). But computing similarity requires deciding what features or dimensions are relevant—zebras, tigers and referee shirts all have stripes, but that doesn't make them meaningfully similar. Background knowledge solves this issue, supplying the conceptual "glue": categories function as learned "theories" about the world, providing the causal, functional and explanatory frameworks that determine which features matter and why they co-occur (Murphy & Medin, 1985). For example, birds have wings, hollow bones, and are able to fly, but just knowing that these features cluster together misses the justifying explanatory structure: birds can fly because they have wings, and they have hollow bones because lighter bodies are better for flight (Murphy & Medin, 1985). These internal theories transform clusters of associated properties into a meaningful whole.

Coherence, however, isn't reducible to semantic relatedness alone. Consider a chain of related objects: cat, dog, bone, and broth. While each individual pair shares an associative link, no single thread unites all four. The same is true of radial categories, where members connect to one central concept in different ways. For example, cat, log and leash all connect to dog—one shares a superordinate category, one rhymes, and one is functionally linked—but they don't cohere with each other. Here, finding a word is not the same as finding a category: a category requires a theory that applies to all its members for the same reason.

This kind of categorization draws on different levels of prior knowledge and experience. Conventional categories have the most scaffolding: both a stored representation and a readily available explanatory framework. For these categories, people are able to easily generalize in a theory-driven way. For example, we can identify a pair of scissors based on our experiences with them: movable blades are essential to its function, but color can vary; they can come in different sizes, but should be no bigger than what can fit in a person's hand.

But what do people do when they encounter a category where there is no existing internal theory? Ad-hoc categories, which are formed by the learner on the spot, require integrating seemingly unrelated ideas around a novel goal or context (Barsalou, 1983). Some ad-hoc categories are schema-based: novel as categories, but still organized around familiar goals or situations. A birth certificate, a pet, and a family photo have little in common at first glance, but should all be rescued in a house fire. Even someone who hasn't previously grouped these items can identify what dimensions matter: irreplaceability, portability, and sentimental value. Schema-based ad-hoc categories are constructed on the fly, but the instructions to do so already exist. Even so, these categories are less stable than conventional ones (Barsalou, 1983). Without the relevant context, people struggle to identify the category at all (Barsalou, 1983). With it, exemplar generation still produces a wider range of items, agreement is low, and typicality judgments are fluid (Barsalou, 1983, 1985, 1987). Beyond this, people must consider whether others share the prior knowledge needed to correctly interpret the category.

Truly novel ad-hoc categories lack both a stored representation and a guiding schema. Learners must discover the conceptual "glue" themselves. This is exactly the task

of the popular *New York Times* game, *Connections*. Players see sixteen items that belong to four categories, none of them entirely straightforward. Four items may form a category based on a shared cultural referent, a phonological pattern, or be related to a common saying. For example, "Second Words in Tarantino Movies" (Brown, Dogs, Fiction, Unchained),"Rhymes" (Choir, Fire, Liar, Fryer) and "_____ Candy" (Cotton, Eye, John, Rock).

In *Connections*, players have to find the categories intended by the game-creator. Even when multiple categories seem possible, there is only one correct answer. Our experiment inverts the logic of *Connections*. Participants are given sets of four words and have to come up with the organizing framework themselves. Because the word sets are random (e.g, "mouth," "floor," "plane," "officer"), there is no one correct answer—no singular ground truth category. As a result, accuracy isn't a meaningful measure of performance. Instead, we can ask what do people do when creating coherence for novel ad-hoc categories? How hard is it to come up with a coherent or "good" category? And what makes a category, "good" in the first place?

A good category should be both *specific* (exclusive to all the members) and *applicable* (encompass all the members). For example, given "submarine," "giraffe," "cowboy," and "outer space" the category "things you can photograph" is applicable, but not specific—you can photograph many things outside the category. In contrast, "things in a Western set in space" describes "cowboy" and "outer space," but excludes "giraffe" and "submarine." This category is more specific, but not applicable. A good category, like "things a child might be obsessed with," captures both. Crucially, coherent categories can take many forms: single words ("adventure"), predicates ("things with an immediately recognizable silhouette") and narratives or scenarios ("a nautical cowboy lost in space befriends a giraffe").

Creating a good category requires searching for candidate relationships, testing how they hold across members, and revising if they fail to cohere. Creating the *best* category requires repeating this process to produce, compare and choose from many possible categories. This requires holding multiple options in working memory, which limits how many categories a person can realistically entertain. If ad-hoc category generation is already hard (Barsalou, 1983), comprehensive ad-hoc category generation for random items is likely to be even more so. People could generate a few categories, pick one that's simply "good enough," and move on. Ad-hoc category generation, similar to other processes with high cognitive load, may favor adequate solutions over optimal ones. Indeed, during online language comprehension, listeners and readers regularly rely on shallow or incomplete representations when full computation is too costly (Ferreira & Patson, 2007). During category generation, constrained by working memory capacity and the effort of extensive search, people may settle early on a just "okay" option.

Large Language Models (LLMs) are not subject to the same constraints, and therefore provide a useful point of comparison. With greater memory capacity, LLMs can generate more categories before selecting the best one. LLMs context windows far exceed human working memory (e.g., 400,000 tokens for GPT-5, OpenAI, 2025).

Reasoning models, such as GPT-5 and Deepseek, are optimized to entertain multiple possible solutions and iteratively revise them (Guo et al., 2025; de Varda, D'Elia, Kean, Lampinen, & Fedorenko, 2025). More categories, in turn, increase the likelihood of finding a specific and applicable one. This leads to two predictions. Given that this type of constraint-satisfaction problem is part of what LLMs are optimized for, we predict that they will produce more specific and applicable categories than people when faced with a truly novel ad-hoc categorization task, *if they are able to select the best option from the candidates they generate*. Conversely, people may generate categories that are less specific, but optimized for different pressures, like communicative efficiency and alignment with a conversational partner (Grice, 1975). Speakers often draw on contextual information and inferred prior knowledge about their interlocutor to balance informativity against plausibility (Goodman & Frank, 2016). This trade-off could lead speakers to generate vague categories that are easier to interpret, prioritizing shared understanding over specificity.

First, the *category generators* (human participants, Open AI's GPT-5, and DeepSeek's deepseek-reasoner) created novel category descriptions for four randomly paired nouns (e.g., "horn,""chick," "flower," "baseball") and indicated how difficult it would be for people to generate the description for them. Afterwards, a second set of human participants, the *category raters*, evaluated category "goodness" on two dimensions: *specificity* (how exclusive the category is to all the members) and *applicability* (how well the category works for all the members.)

Methods

Participants

Category Generators We recruited 55 human participants from the University of Wisconsin - Madison participant pool, who received course credit for their participation. We also included two LLMs: GPT-5.1 Thinking, with reasoning effort set to medium, and Deepseek-chat, with full reasoning trace reporting. Our analyses investigate LLM performance in general (averaged across both models) but future work will compare performance between them.

Category Raters We recruited an additional sample of 533 people from the same participant pool. After excluding 118 participants for failing attention checks (mean response time for both applicability and specificity ratings of < 2.5 s), the final sample included 415 raters.

Stimuli

We began by generating a large pool of nouns from Wordnet (Miller, 1995) that met the following criteria: single words

Lexical Norms	M	Mdn	SD	Min	Max
Frequency	76.098	53.373	80.647	11.333	524.314
Concreteness	4.20	4.60	0.80	2.11	4.98
Similarity[a]	0.34	0.34	0.06	0.23	0.58

Table 1: Word group characteristics ($N = 90$). [a]Pairwise semantic similarities using FastText Common Crawl embeddings. Word frequency counts represent words per million for the dominant part of speech.

(no compounds or hyphenates) that were most frequent as nouns and included in the top 30% most frequent words (based on the SUBTLEX corpus, Brysbaert & New, 2009). One word was randomly selected to be a seed word for each group. The remaining three words were selected randomly within similar frequency and concreteness ranges (Brysbaert, Warriner, & Kuperman, 2014). This ensured that there were no frequency or concreteness outliers in a group. This resulted in 90 4-word groups that served as stimuli for our main study.

Procedure

Category Generators Before beginning the experiment, participants were provided with four example categories to help calibrate responses. On each trial, they saw a group of four words and were asked to generate a category that connects the words in a meaningful way. Though human participants only generated categories for a subset of the word groups (22 or 23 trials), models completed all of them. Models were prompted using separate API calls for each word group. This was completed three times for each model to ensure a better balance between human- and model-generated categories. The order of trials and words within a given trial were randomized. Humans and models received identical prompts. After every trial, participants rated how difficult it was to categorize the items on a likert scale (from "Very Easy" to "Very Difficult") and judged how likely, on a scale from 0-100, they think others would be to generate the same category.

Category Raters The responses produced by the category generators were used as the stimuli for the category raters. Nonsense and "I don't know" answers were removed, resulting in 1,962 candidate categories of approximately 18–22 categories per word group.

Participants were shown four words and their category description. They then rated, on a continuous scale from 0-100, the specificity of the category (with markers from "Very general" to "Very specific") and the applicability of the category (with markers from "Not at all applicable" to "Perfectly applicable"). Each participant only rated one category for each word group to prevent comparison across responses. Likewise, groups were shuffled so raters didn't see categories all produced by the same category generator. Raters were provided with examples of applicable and specific categories to guide their ratings.

Results

Our analysis included 1,121 human-generated categories and 412 model-generated categories for 90 different word sets (selected examples in Table 2. Each category received between 6 and 58 ratings ($M = 10.80$ ratings). For each word group, we computed a semantic similarity measure by calculating all pairwise similarities between words using FastText Common Crawl vector embeddings. Mean characteristics for the word groups are reported in Table 1.

In the following analyses, we investigate whether LLMs and people differ in the goodness of the categories they generate, defined by applicability and specificity. We then ask whether the same word groups give rise to better category generation in both models and people. Finally, we examine which properties of the word groups themselves predict category goodness, and what makes it difficult to generate a category in the first place. Future analyses will evaluate goodness comparisons between different LLMs.

Who generates "better" categories?

We first ask whether LLMs and people differ in the overall applicability and specificity of the categories they generate. To address this question, we fit two linear mixed-effects models. The first regressed applicability ratings on mean word frequency, mean word concreteness, the interaction between mean word-group semantic similarity and subject type, and the interaction between difficulty ratings and subject type, with random intercepts by word group. The second model regressed specificity ratings on mean word-group concreteness, mean word-group frequency, mean word-group semantic similarity, and the interaction between difficulty ratings and subject type, again with random intercepts by word group.

Applicability. LLMs generated categories that were rated as more applicable than those generated by people when controlling for concreteness, frequency, mean pairwise semantic similarity, and difficult ($b = 5.010$, $SE = 1.045$, $t = 4.793$, $p < .001$). Despite this overall difference, applicability ratings for model- and human-generated categories were positively correlated across word groups ($r = 0.559$, $p < .001$; Figure 1). That is, although LLMs tend to generate more applicable categories overall, models and people generate more applicable and less applicable categories for the same word groups. This suggests that word groups that promote more applicable category-generation in people, also promote more applicable category-generation in LLMs.

Specificity. Models also generated more specific categories than people when controlling for concreteness, frequency, mean pairwise semantic similarity, and difficulty ($b = 21.114$, $SE = 0.845$, $t = 24.988$, $p < .001$). In contrast to applicability, however, specificity ratings for human- and model-generated categories were slightly negatively correlated across word groups ($r = -0.210$, $p = .050$; Figure 1). That is, although models and people tend to agree on which

Generator	Word Set	Category
Human	case, death, party, couple	*"All things that tend to appear in murder mysteries involving rich families"*
Human	doctor, phone, hand, town	*"Things that would be helpful in an apocalypse where the phone lines still work"*
Human	murder, dude, afternoon, message	*"Surfing at 1 pm when someone is killed and made to look like a suicide to send a message"*
Model	heavens, stress, punishment, option	*"Concepts often discussed in religious sermons about moral choices and their consequences"*
Model	killer, surprise, space, idiot	*"Sensational buzzwords often used in clickbait-style online titles (crime, shock, awe, and stupidity)"*
Model	neck, arm, hat, computer	*"Things you might adjust for comfort when settling in to use a computer (your neck position, arm position, hat, and the computer itself)"*

Table 2: Examples of categories generated by humans and models.

word groups yield more applicable categories, they respond differently to those same word groups in terms of specificity. The word groups that lead models to generate more specific categories are not the same word groups that lead people to generate more specific categories.

How does semantic distance affect category goodness?

We next examined whether LLMs and people generate similarly applicable categories under the strain of coercing semantically distant concepts into a single category. As mean semantic similarity among the words in a group decreases, applicability declines for both models and people ($b = 86.613$, $SE = 17.803$, $t = 4.865$, $p < .001$). Word groups that are closer together in semantic space tend to result in more applicable categories.

Mean semantic similarity also interacted with subject type such that low semantic similarity resulted in a larger decrease in applicability for people than for models ($b = -45.959$, $SE = 18.938$, $t = -2.427$, $p = .017$; see Figure 2).

We also explored whether applicability depends on the presence of semantic outliers within a word group. Replacing mean similarity with the variance or coefficient of variation of pairwise similarities did not yield significant effects or interactions. However, we computed an outlier-based measure of semantic similarity by first computing the pairwise similarities for all words in a group and then calculating the mean of the pairwise similarities for each word to the other three words. Finally, we selected the word with the minimum mean pairwise similarity in the group and recorded this measure as the outlier distance. This measure did significantly predict applicability ($b = 4.142$, $SE = 0.997$, $t = 4.157$, $p < .001$) and marginally interacted with subject type ($b = -1.976$, $SE = 1.120$, $t = -1.764$, $p = .081$). This pattern suggests that category applicability is influenced not only by average semantic distance, but by the presence of particularly distant concepts within a group. For example, it may be harder to generate an applicable category for "exit," "hug," "warden," and "chopper," not only because the mean semantic similarity is low, but because one word ("hug") is harder to coerce into categories that encompass the others.

Mean semantic similarity did not predict specificity ratings ($b = 1.368$, $SE = 7.906$, $t = 0.173$, $p = .863$), indicating that semantic distance affects whether a category applies to a word group, but possibly not how narrowly that category is defined.

Other predictors of category goodness

We also examined whether other lexical properties of the word groups predicted applicability or specificity. As the mean concreteness of the words in the group increased, applicability decreased ($b = -4.450$, $SE = 0.931$, $t = -4.781$, $p < .001$): more abstract word groups yielded more applicable categories. A plausible explanation is that abstract words are more polysemous and therefore easier to semantically coerce into a shared category (Barsalou, 1983; Gentner & Markman, 1997; Hoffman, Lambon Ralph, & Rogers, 2013). We did not find any significant effects of mean word frequency on applicability or specificity, nor any significant interactions involving these predictors.

What makes it difficult to generate a category in the first place?

Although difficulty ratings were not a primary measure of category goodness, we examined their relationship to semantic similarity and specificity. Difficulty ratings were strongly predicted by mean pairwise semantic similarity within the word group ($b = -5.838$, $SE = 0.989$, $t = -5.905$, $p < .001$), indicating that semantically distant word groups were perceived as more difficult to categorize. The *variance* of the pairwise semantic similarities was not associated with difficulty ratings. Model and human difficulty ratings were moderately correlated ($r = .348$, $p < .001$; computed using Pearson correlation), suggesting partial alignment in "perceived" task difficulty.

Difficulty ratings also interacted with subject type (human vs. LLM) in predicting specificity ($b = 4.846$, $SE = 0.545$, $t = 9.411$, $p < .001$). As human-rated difficulty increased, people generated less specific categories (presumably because they could not think of a more specific one and this is what made the trial difficult). LLMs, however, generated *more* specific categories as LLM-rated difficulty increased (Figure 3).

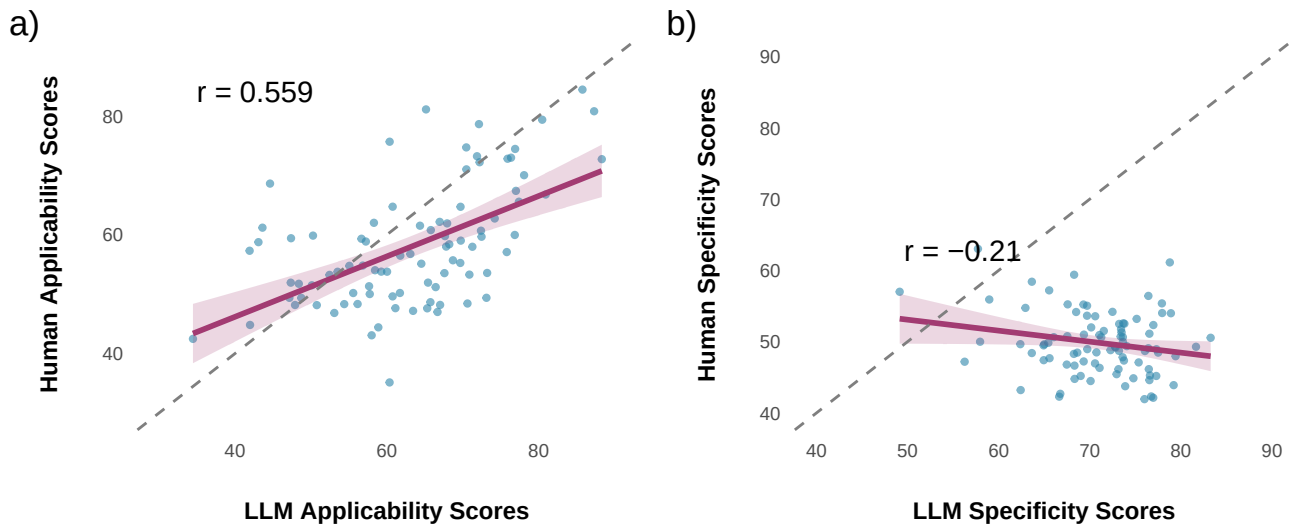

Figure 1: a. The relationship between LLM category applicability scores (x-axis) and human category applicability scores (y-axis). Dotted line indicates where applicability scores of LLMs equal those of people. LLM applicability scores and human applicability scores are positively correlated. b. The relationship between LLM category specificity scores (x-axis) and human category specificity scores (y-axis). Dotted line indicates where specificity scores of LLMs equal those of people. In contrast to applicability scores, specificity scores from LLM-generated versus human-generated categories are not correlated. The line of fit in both plots is based on a linear-model prediction and its standard error.

Discussion

Given groups of four random nouns (e.g., "birthday," "state," "pain," "meeting"), both people and models were able to create some coherent category that linked these concepts ("words that could be used in an angsty teenage girl's diary on her birthday"). LLMs generated categories that were rated by people as more specific and applicable to the word sets than categories generated by other people. This difference between people and LLMs may be caused by one or more of the following factors: (1) *A difference in category generation.* People and LLMs may differ in the number of candidate categories under consideration. People may generate fewer categories than LLMs, deeming the generated candidates "good-enough" (Ferreira, Bailey, & Ferraro, 2002) while LLMs may be able to generate more candidate categories. If this is what is responsible for the advantage shown by LLMs, it should disappear if LLMs and people were given the same candidates and simply asked to choose among them. (2) *A difference in category selection.* It is also possible that LLMs and people differ in the selection process itself. If so, we should find that category choices made by people and LLMs would continue to differ even when provided with all the same candidates.

The current data suggest the human-LLM gap has more to do with differences in generating candidate categories. For example, the LLM advantage in generating more specific categories was most pronounced for the most difficult word sets which tended to be more semantically distant, perhaps because it was most difficult to generate candidate categories. When faced with these semantically distant word groups, participants often generated categories that were only applicable to a subset of the terms. For example, "brain," "beer," "fish," and "ice," yielded "words that can go with bucket," which does not apply to the first item as well. It seems likely that participants settled on this "good-enough" response because they could not think of one that that applied to all four words. Future work will incorporate word groups spanning a wider range of similarity, including highly similar sets.

We do not find that semantic similarity predicts specificity, yet it is most predictive of applicability. Aligned with this, model and human evaluations of specificity were not positively correlated: the groups that prompt specific categories from people are not the same ones that prompt them from models. Though both humans and models can produce highly applicable categories, they arrive at differently specific ones. For instance, though "Things that can be related to an old western stop-motion movie" (from humans) and "things you can take" (from a model) for "picture," "drink," "shot," and "movie" received similar applicability ratings, their specificity ratings differ sharply. What characterizes these word groups where people and models diverge on specificity? One possibility is that the *specific context* of a given word group evokes different representations in models versus people. We plan to evaluate more descriptive features of both the word groups and the generated categories. People may prioritize creating categories that are likely to be *interpretable* by others. Though a model may generate a very specific category, we are left with the question: would someone be able to suc-

Word Set: heavens, stress, punishment, option				
Goodness Measure	**Humans**		**Models**	
	Most	**Least**	**Most**	**Least**
Applicability	*"topics that may be talked about in church"*	*"College"*	*"Things a person might worry about when deciding whether to confess to a serious sin or crime"*	*"Things that can be described as 'heavenly' or 'hellish' depending on perspective"*
Specificity	*"words you might hear in a prison"*	*"anxiety"*	*"Concepts often discussed in religious sermons about moral choices and their consequences"*	*"Things that can be described as 'heavenly' or 'hellish' depending on perspective"*

Table 3: Examples of categories generated by people and LLMs, rated as the most and least applicable and specific for the word-set *heavens, stress, punishment, option*.

Figure 2: Applicability ratings of human and model-generated categories. Lines of fit represent model predictions and error bands represent 95% CIs. Each point represents one word group.

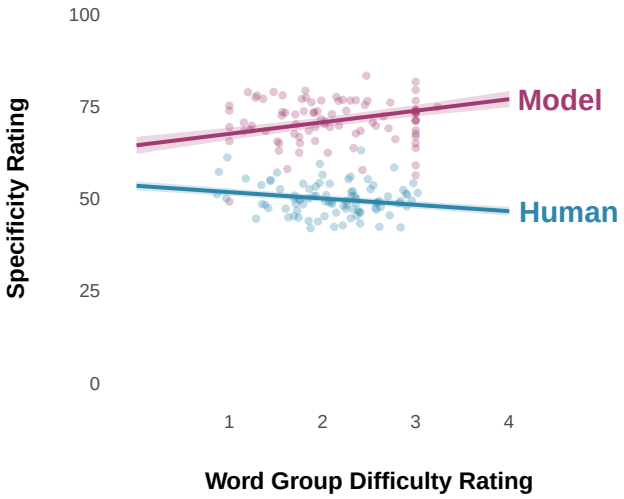

Figure 3: As the difficulty rating provided for a individual trial increases, LLMs generate more specific categories while people generate less specific ones. Lines of fit represent model predictions and error bands represent 95% CIs. Each point represents one word group.

cessfully generate applicable exemplars given that category description? More specific categories may demand more specific exemplars, taking into account the background knowledge of an interlocutor. In a follow-up experiment, we plan to address the question of alignment more explicitly. We will present the generated categories to a new sample of human and model participants tasked with producing exemplars. This will help us determine whether the specificity and applicability of ad-hoc categories is influenced by a bias toward representational alignment between communicators.

Taken together, this work provides evidence that generating a "good" category for semantically distinct groups is difficult, though LLMs outperform people in terms of both category applicability and specificity. While both humans and models are sensitive to semantic distance, they respond to it in systematically different ways. As semantic similarity decreases, models generate more applicable categories and as difficulty of generating the category increases, models generate more specific categories, while people generate less specific ones. Unconstrained by the pressure to respond quickly, models may be able to generate many more candidate categories. Humans, in contrast, rely on good-enough processing that prioritizes satisfactory solutions over optimal ones. At the same time, preliminary evidence suggests humans produce more applicable categories than models for highly similar word groups, where flexible, generalizable representations may be better suited. The gap between model and human performance may be about category *generation* rather than *evaluation*; finding the right conceptual glue to bind categories may depend less on how we choose among candidates, and more on how many we can generate in the first place.

References

Barsalou, L. W. (1983). Ad hoc categories. *Memory & Cog-*

nition, *11*(3), 211–227.

Barsalou, L. W. (1985). Ideals, central tendency, and frequency of instantiation as determinants of graded structure in categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *11*(4), 629–654. doi: 10.1037/0278-7393.11.1-4.629

Barsalou, L. W. (1987). The instability of graded structure: Implications for the nature of concepts. In *Concepts and conceptual development: Ecological and intellectual factors in categorization* (pp. 101–140). Cambridge University Press.

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, *41*(4), 977–990.

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, *46*(3), 904–911.

de Varda, A. G., D'Elia, F. P., Kean, H., Lampinen, A., & Fedorenko, E. (2025). The cost of thinking is similar between large reasoning models and humans. *Proceedings of the National Academy of Sciences*, *122*(47), e2520077122.

Ferreira, F., Bailey, K. G. D., & Ferraro, V. (2002). Good-enough representations in language comprehension. *Current Directions in Psychological Science*, *11*(1), 11–15. doi: 10.1111/1467-8721.00158

Ferreira, F., & Patson, N. D. (2007). The 'good enough' approach to language comprehension. *Language and Linguistics Compass*, *1*(1–2), 71–83. doi: 10.1111/j.1749-818X.2007.00007.x

Gentner, D., & Markman, A. B. (1997). Structure mapping in analogy and similarity. *American Psychologist*, *52*(1), 45–56. doi: 10.1037/0003-066X.52.1.45

Goodman, N. D., & Frank, M. C. (2016). Pragmatic Language Interpretation as Probabilistic Inference. *Trends in Cognitive Sciences*, *20*(11), 818–829. doi: 10.1016/j.tics.2016.08.005

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Syntax and semantics, vol. 3, speech acts* (pp. 41–58). New York: Academic Press.

Guo, D., Yang, D., Zhang, H., Song, J., Wang, P., Zhu, Q., ... Zhang, Z. (2025). Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, *645*(8081), 633–638. doi: 10.1038/s41586-025-09422-z

Hoffman, P., Lambon Ralph, M. A., & Rogers, T. T. (2013). Semantic diversity: A measure of semantic ambiguity based on variability in the contextual usage of words. *Behavior Research Methods*, *45*(3), 718–730. doi: 10.3758/s13428-012-0278-x

Miller, G. A. (1995). Wordnet: a lexical database for english. *Communications of the ACM*, *38*(11), 39–41.

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review*, *92*(3), 289–316. doi: 10.1037/0033-295X.92.3.289

OpenAI. (2025). *Gpt-5.* `https://platform.openai.com/docs/models/gpt-5`.

Shepard, R. N., Hovland, C. I., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological Monographs: General and Applied*, *75*(13), 1.