The Grounding Gap: How LLMs Anchor the Meaning of Abstract Concepts Differently from Humans

Odysseas S. Chlapanis
Department of Informatics
Athens University of Economics and Business
Archimedes, Athena Research Center
Greece

Orfeas Menis Mastromichalakis
Instituto de Telecomunicações
Portugal

Christos H. Papadimitriou
Department of Computer Science
Columbia University
United States

Abstract

Abstract concepts — justice, theory, availability — have no single perceivable referent; in the human brain, their meaning emerges from a web of experiences, affect, and social context. Do large language models (LLMs) ground abstract concepts in a similar way? We study this by replicating property-generation experiments from cognitive science on 21 frontier and open-weight LLMs. Across models and experiments, we find a consistent pattern: when compared to humans, models rely too heavily on word associations, and underproduce properties tied to emotion and internal states. This yields a large and consistent grounding gap: no model exceeds a Pearson correlation $r = 0.38$ with human responses, compared to a human-to-human ceiling above $r = 0.9$. To better interpret this gap, we also replicate a rating experiment on grounding categories and find that here LLMs align more closely with human judgment, and alignment improves as models get larger. We then use sparse autoencoders (SAEs) to inspect whether this information is also reflected in the models' internal features, and we do identify features connected to grounding dimensions such as "sensorimotor" and "social". These findings suggest that current LLMs can recover grounding dimensions when explicitly queried, but do not recruit them in a human-like way when words are generated freely.[1]

1 Introduction

Concrete concepts such as *cat* and *table* are anchored in shared experience through identifiable referents that can be perceived through the senses [Borghi et al., 2017]. In contrast, abstract concepts such as *art*, *adventure*, or *justice*, lack such referents. When humans process abstract words, they reconstruct their meaning from grounded experiences. Consider how a person makes sense of *art*: a colorful *painting*, a feeling of *awe*, a favorite *concert*, as well as more abstract ideas such as *creativity*. Grounded-cognition theories argue that such sensorimotor, emotional, and social anchors are not incidental but constitutive of meaning [Lakoff and Johnson, 2003, Barsalou, 1999, 2026]. Whether LLMs represent abstract concepts in a similarly grounded way, or rely primarily on patterns of linguistic association [Bender and Koller, 2020, Bender et al., 2021], remains an open question.

To measure how closely LLMs align with humans in the grounding of abstract concepts, we replicate on LLMs two property-generation experiments from cognitive science [Harpaintner et al., 2018,

[1] Data and code are publicly available at `https://github.com/odychlapanis/grounding-gap/`.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Kelly et al., 2024]. In these experiments, participants — in these references humans, in this paper models — are given a concept and asked to list the properties, situations, or associations that come to mind. Although individual responses vary, the coded distributions across human participants are highly consistent: people anchor abstract concepts along grounding dimensions such as sensorimotor experience, emotion, and social context in a similar and consistent way. We then compare the distributions produced by models with those produced by humans. Across 21 frontier and open models, we find a large and consistent grounding gap. In both experiments, no instruction-tuned model reaches a human–model correlation above $r \approx 0.38$, while the human–human ceilings exceed $r = 0.9$. Models are also much closer to one another than to humans: model–model correlations are systematically higher than model–human correlations. Taken together, these results show that current LLMs exhibit a shared mode of grounding abstract concepts that is systematically different from the human one.

But does the gap arise because models lack an understanding of the categories that structure human grounding, or because this understanding is not recruited in a human-like way when concepts are generated freely? To answer, we turn to a rating experiment from cognitive science, in which words are evaluated directly along grounding-related dimensions [Troche et al., 2017]. There, recent LLMs align much more closely with human judgments, and this alignment improves with model scale, leaving a gap of only about .1 in Pearson correlation with humans. This contrast suggests that the relevant dimensions are at least partly available to the models when they are explicitly queried, but they are not spontaneously recruited in a human-like way during property generation.

This raises a mechanistic question: Can we find in LLMs internal representations that align with these grounding-related dimensions? To investigate this, we analyze models with sparse autoencoders (SAEs), a technique that helps identify interpretable internal features [Bricken et al., 2023, Huben et al., 2024, Templeton et al., 2024]. We search for features whose activations track grounding-related dimensions, and in two Gemma 3 models we find evidence of such structure: some internal features correlate with human grounding dimensions or sub-dimensions. We then validate these features by showing that steering the model along them increases generation of properties from the corresponding target categories. Taken together, our findings suggest that current LLMs understand the basic grounding categories nearly–but not quite–as well as humans do, and yet they do not recruit them in a human-like way when generating freely.

Contributions.

- We replicate two property-generation experiments from cognitive science on 21 frontier and open LLMs, and show a large and consistent gap between humans and models in the grounding of abstract concepts.
- We replicate a rating experiment on grounding-related dimensions and show that LLMs align much more closely – but not fully – with human judgments there, indicating that the gap is not simply due to failure to recognize the relevant dimensions.
- We use sparse autoencoders to probe two Gemma 3 models for internal features related to grounding dimensions, and find evidence of such structure.

2 Related work

Grounded cognition and grounding norms. Traditional theories of *grounded cognition* in language posit that abstract categories are not represented as arbitrary symbols but are grounded in sensorimotor and affective experience [Barsalou, 1999, Barsalou and Wiemer-Hastings, 2005]. Neurocognitive evidence supports that action words engage motor and premotor cortex [Hauk et al., 2004, Pulvermüller, 2005], and emotional and social dimensions of abstract concepts have been linked to interoceptive processing [Critchley et al., 2004, Mancano and Papagno, 2026]. Behavioral norms operationalize these dimensions at scale: datasets in which human participants rate words on Likert scales for dimensions such as sensorimotor strength, arousal, valence, and socialness. Lynott et al. [2020] provide modality-specific strength ratings for 40,000 English words, while Scott et al. [2019] and Diveica et al. [2023] cover arousal, valence, and socialness dimensions. These corpora enable grounding research to move beyond binary concrete/abstract distinctions [Brysbaert et al., 2014]. The property-listing paradigm [McRae et al., 2005, Barsalou and Wiemer-Hastings, 2005] provides a richer signal: rather than scalar ratings, participants freely generate properties of

concepts, yielding a distributional profile over experiential dimensions. Harpaintner et al. [2018] applied this paradigm specifically to abstract nouns, showing that they elicit rich sensorimotor, social, and affective features. Kelly et al. [2024] extended the paradigm to test whether emotion concepts form a distinct subcategory, finding that they rely more strongly on interoceptive features than other abstract concepts. These two studies form the empirical basis of our behavioral experiments.

LLM cognitive experiments. Early work showed that language models often fail psycholinguistic diagnostics and do not reliably ground semantic properties [Ettinger, 2020]. Pezzelle et al. [2021] compared pre-trained transformer representations against human concept norms and found that visual co-training improves alignment for concrete word pairs, but offers little benefit for abstract ones. Recent work extends this question to frontier and open LLMs by prompting them for human-style ratings and comparing them to established norms. Xu et al. [2025] report strong alignment with Glasgow and Lancaster norms on non-sensorimotor dimensions, but weaker alignment on sensory and motor domains. Wang et al. [2025b] similarly find persistent gaps for abstract concepts, embodied modalities, and emotion or social dimensions across nine LLMs, with alignment improving with scale. Related studies also show that LLMs can approximate affective and sensorimotor ratings [Trott, 2024] and recover aspects of perceptual color structure from text alone [Abdou et al., 2021]. Closest to our setup, Suresh et al. [2023] use property listing rather than ratings, but focus on concrete concepts and feature-overlap similarity against existing norms. In contrast, we target abstract concepts and classify generated properties into experiential categories.

Mechanistic interpretability of semantic concepts. A prominent line of mechanistic interpretability work uses linear probes to localize non-linguistic latent structure in LLM residual streams, recovering concept axes for board game states, spatiotemporal coordinates, and sentiment [Li et al., 2023, Nanda et al., 2023, Gurnee and Tegmark, 2024, Tigges et al., 2024]. Some works extend this paradigm by using these latent vectors to steer models towards the detected concepts [Turner et al., 2024, Panickssery et al., 2024]. Recent studies have identified circuit-level emotion components to control emotional expression [Wang et al., 2025a] and recovered emotion directions in models like Claude Sonnet 4.5 [Sofroniew et al., 2026]. In this work, we analyze features learned by sparse autoencoders (SAEs) to probe whether grounding-related information is reflected in model representations. SAEs decompose polysemantic LLM representations into monosemantic, interpretable features [Bricken et al., 2023, Huben et al., 2024, Templeton et al., 2024]. Recently, SAE features have been applied for analyzing emotion [Wu et al., 2025]. In our work, we leverage the Gemma Scope 2 SAE models [McDougall et al., 2025], the successor to Gemma Scope [Lieberum et al., 2024], and we probe sensorimotor, internal-state, and social concepts.

3 The grounding gap

We study LLM grounding of abstract concepts through *property generation*, a behavioral paradigm from cognitive science in which participants are given a concept and asked to list the features, situations, or associations that come to mind for it. Unlike scalar rating tasks [Troche et al., 2017], which require participants to map conceptual content onto a fixed set of predefined dimensions, property generation captures the content that people spontaneously recruit when representing a concept. For this reason, it has been used as a richer and more faithful way to study grounding, especially for abstract concepts [Barsalou and Wiemer-Hastings, 2005, McRae et al., 2005].

To measure how closely LLMs align with humans on conceptual grounding, we replicate two human property-generation studies on abstract concepts: Harpaintner et al. [2018] and Kelly et al. [2024]. We selected these studies because both provide publicly available data and explicit coding taxonomies, enabling controlled human–model comparison under complementary experimental designs. Harpaintner et al. [2018] (*Experiment 1*) organizes properties of abstract concepts into *sensorimotor*, *internal-state*, *social*, and *verbal-association* content, reflecting dimensions of abstract-concept representation that are well established in the broader literature [Troche et al., 2017, Villani et al., 2019]. Kelly et al. [2024] (*Experiment 2*) provides a more specialized setting, with a different stimulus design and coding taxonomy with a particular focus on emotion concepts. Using both experiments allows us to test whether human–model grounding gaps persist across distinct property-generation paradigms rather than arising from a single dataset or coding scheme.

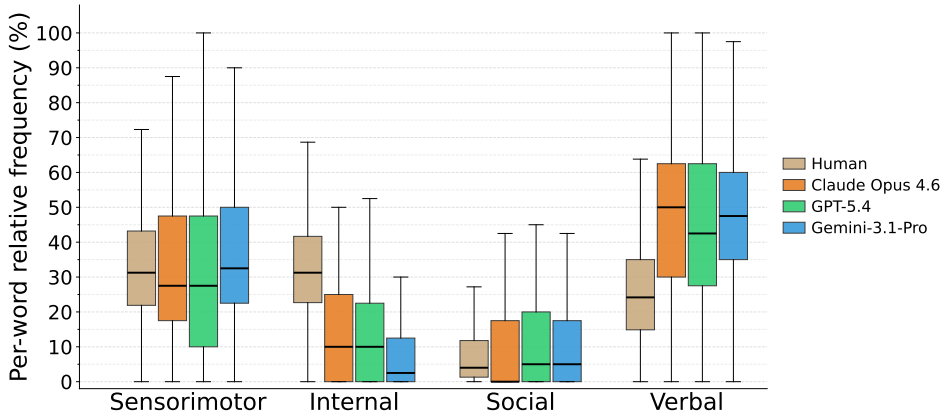


Figure 1: Per-word property-category frequency distributions for three frontier LLMs vs. the human baseline on the 293-abstract-noun benchmark [Harpaintner et al., 2018]. Each box spans the 25th–75th percentile of word-level frequencies; median annotated.

3.1 Experimental setup

We apply the same procedure in both experiments, following their original designs. For each stimulus word, we prompt the model to generate the properties, situations, or associations that come to mind (four properties for Experiment 1 and five for Experiment 2), using the original human instructions (prompts in Appendix G). We then map the generated properties to the taxonomy of the corresponding human study and compare the resulting per-word category generations against the human ones. All scores are averaged over 10 runs per model; see Appendix E for convergence analysis.

Stimulus sets. The stimulus set consists of the target words presented to participants—or, in our case, to the models—for property generation. For the first experiment [Harpaintner et al., 2018], we derive a 293-word English stimulus set from the original 296 German abstract concepts. Two items were excluded as translation duplicates (the original paper provided English translations for all words), and one additional item was excluded because it is used as a one-shot example in the original prompt. In the second experiment [Kelly et al., 2024], the stimulus set contains 354 words split into three subsets: 118 abstract emotion words, 117 abstract non-emotion words, and 119 concrete words. We use the 235 abstract words for our analysis; see Appendix C for findings for concrete words.

Property coding. The generated properties are open-ended, so they must be assigned to the grounding categories defined by each study. Following the original coding schemes, generated properties for Harpaintner et al. [2018] are assigned to: *Sensorimotor*, *Internal State & Emotion*, *Social*, and *Verbal Association*, while for Kelly et al. [2024] to *Taxonomic*, *Entity*, *Situation*, and *Introspective*. We use LLMs-as-coders, after validating candidate models on human-annotated word-to-category pairs. The selected coders achieve agreement with the ground truth that is comparable to, or higher than, human coders, while remaining practical to use at scale. We use *Gemini-2.5-Flash-Lite* as the primary coder for Experiment 1 and *Gemini-2.5-Flash* for Experiment 2. Detailed coder comparisons, annotation protocols, and reliability analyses are reported in Appendix A.

Alignment metric and human ceiling. After coding, each stimulus word is represented by the relative frequencies of its responses across the four target categories. We compare human and model frequency profiles using per-category Pearson r, and report *Mean* r, the average across the four categories, as our main alignment metric. Concretely, for each category we correlate the human and model frequencies across all stimulus words (e.g., the Sensorimotor frequencies of the 293 concepts in Experiment 1) and then average the four per-category correlations. Our conclusions are unchanged under Spearman correlation and representational similarity analysis (Appendix J). The corresponding human ceiling is the Pearson r that two independent splits of raters would obtain on the same words: for Experiment 2 we compute it directly from Kelly et al. [2024]'s participant responses, and for Experiment 1 we estimate it analytically from Harpaintner et al. [2018]'s per-word aggregates using

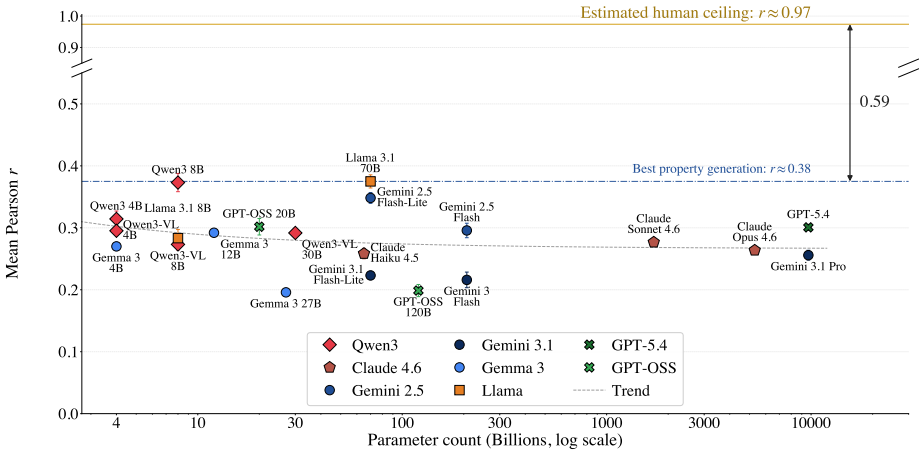


Figure 2: Mean r for all LLMs vs. the human ceiling on Experiment 1 [Harpaintner et al., 2018]. Parameter count for closed models is based on estimations from Li [2026]. The dashed curve shows the saturating fit to the 21 models (trend). The dash-dotted line marks the best model, and the arrow its distance to the human ceiling.

standard inter-rater reliability formulas; the two methods agree to within 0.02 on Experiment 2, validating our estimator (Appendix F).

Models and settings. We evaluated a diverse set of 21 frontier and open-weight LLMs spanning several families; we refer to Appendix B.1 for the full list of models. We used the default temperature for all generations and to ensure coding reliability LLM-coders used a temperature of $T = 0.0$.

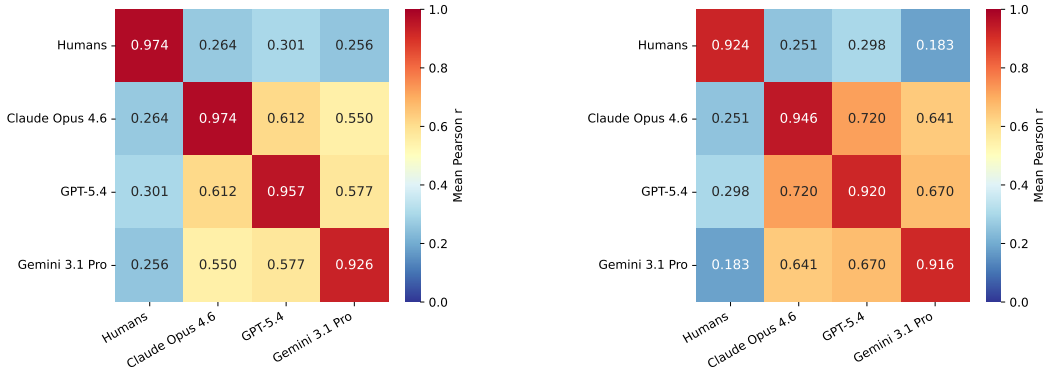
3.2 Results

Experiment 1. Figure 1 illustrates the per-word and per-category distributions for human responses and for three representative frontier models: Claude Opus 4.6, Gemini 3.1 Pro Preview, and GPT-5.4. Relative to humans, all three models generate substantially more *Verbal Association* properties and substantially fewer *Internal State & Emotion* properties. The same overall pattern holds across all models (see Appendix B for the full set of models). The similarity between the frequency of the different models also indicates that this is not a coincidental property of a single model family, but a consistent behavior across current architectures. By contrast, *Sensorimotor* frequencies remain closer to the human range in the aggregate, and *Social* frequencies are also broadly comparable, although this category is relatively sparse for both humans and models and is therefore less informative at the frequency level alone.

The frequency distributions indicate a systematic mismatch between human and model grounding, but aggregate frequency alone does not show whether models ground the *same* concepts in the *same* way as humans. For example, a model may produce a human-like number of *Sensorimotor* or *Social* properties overall, while assigning those properties to different words. We therefore measure alignment directly by computing Pearson correlations between human and model category-frequency generations across the stimulus set, as described in Section 3.1.

Table 1: Per-category Pearson correlations on Experiment 1 for three representative frontier models.

Model	Mean r	Sensorimotor	Internal	Social	Verbal
Estimated human ceiling	*0.974*	*0.966*	*0.964*	*0.971*	*0.992*
GPT-5.4	0.301 ± 0.005	0.332	0.258	0.400	0.213
Claude Opus 4.6	0.264 ± 0.004	0.203	0.205	0.362	0.284
Gemini 3.1 Pro Preview	0.256 ± 0.008	0.307	0.163	0.383	0.171



(a) Experiment 1 (Harpaintner et al. [2018])

(b) Experiment 2 (Kelly et al. [2024])

Figure 3: Pearson-r correlation heatmaps for three representative frontier models and humans.

As shown in Figure 2, human-model alignment remains far below the estimated human–human ceiling across all 21 models. The best model reaches only a mean $r \approx 0.38$, compared with an estimated human ceiling of about 0.97, with the same pattern across individual categories as shown in Table 1. Thus, even when aggregate frequencies appear relatively human-like, models often associate grounding dimensions with different concepts than humans do. This gap is not explained by scale: larger models and stronger general-purpose systems do not show reliably higher human alignment, suggesting that the mismatch is not simply a capability bottleneck that shrinks with model size.

Inter-model correlations reveal a complementary pattern. As shown in Figure 3a, models correlate substantially more with one another than with humans, despite spanning different providers, data mixtures, and architectural choices. This suggests that current LLMs share a rather common mode of abstract-concept grounding, but one that is systematically distinct from the human pattern. The grounding gap is therefore not idiosyncratic to a single model family, but shared across the instruction-tuned LLMs we evaluated. This shared pattern may partly reflect overlapping pretraining corpora and post-training pipelines rather than a property of language modeling as such.

Experiment 2. Experiment 2 confirms the findings of Experiment 1 under a different experimental setup. Human-model alignment remains low in the combined abstract condition, which pools the *abstract emotion* and *abstract non-emotion* items: even the strongest models reach only a mean $r \approx 0.37$, far below the human-to-human correlation of $r \approx 0.92$. Full results for all models are reported in Appendix C.

Inter-model correlations also follow the same pattern as in Experiment 1. As shown in Figure 3b, models correlate substantially more with one another than with humans. Together, the two experiments reveal a robust behavioral grounding gap: current LLMs share a common model-like pattern of abstract-concept representation, but one that remains systematically distinct from humans.

3.3 LLM-coder error margins

The Experiment 1 Pearson correlations and category frequencies reported above are based on labels produced by Gemini 2.5 Flash-Lite. To estimate whether this affects our conclusions, we manually re-coded the outputs of three frontier models (two of the authors served as annotators, coding each property independently; agreement 68.9%, Cohen's $\kappa = 0.59$; Appendix A.1) and compared the resulting metrics with those obtained from Flash-Lite labels on the same property pairs. The differences are minor: replacing Flash-Lite with human coders changes Mean r by only $\Delta = +0.014 \pm 0.019$ on average, with no consistent per-model shift (per-model 95% confidence intervals contain zero), and changes per-category mean frequencies by roughly $\pm 2\%$. These margins are far smaller than the human–model gap reported in Section 3.2, indicating that our main results are not driven by the use of LLMs as coders. Detailed analyses are reported in Appendix A.

Appendices K to O further examine the property budget, prompt framing, instruction tuning, statistical units and multimodality; none of these closes the gap.

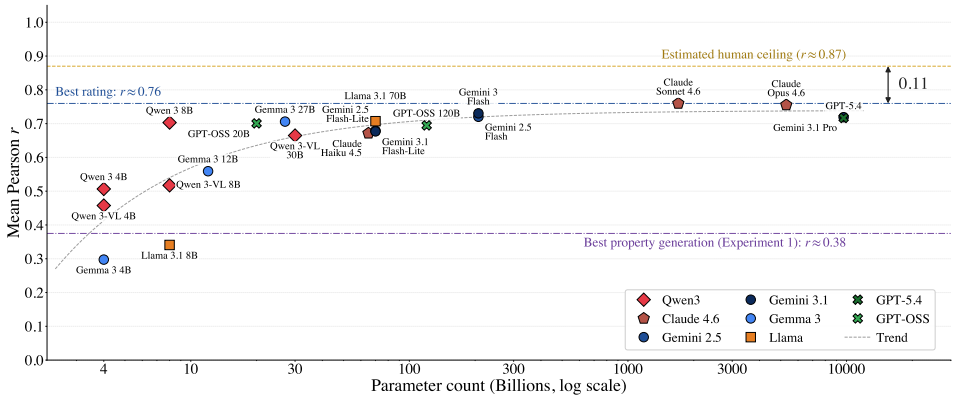


Figure 4: Mean r for all LLMs vs. the human ceiling on the rating experiment [Troche et al., 2017]. Parameter count for closed models is based on estimations from Li [2026]. The dashed curve shows the saturating fit to the 21 models (trend). The dash-dotted lines mark the best model on this task (blue) and on Experiment 1 (purple), and the arrow the distance from the best model to the human ceiling.

4 Rating analysis of grounding dimensions

To better understand the grounding gap identified through the previous experiments, we study whether this gap arises because current LLMs do not properly recover the grounding dimensions that structure human concept representations, or because these dimensions are not recruited in a human-like way when concepts are generated freely. To distinguish between these possibilities, we replicate a rating experiment from cognitive science [Troche et al., 2017]. Unlike property generation, rating experiments do not test the spontaneous associations for a concept; instead, they test whether the model can score a word on semantic dimensions once these dimensions are explicitly specified.

Setup. We follow the setup of Troche et al. [2017], in which human participants rated 751 English nouns on 14 dimensions using a 1–7 Likert scale. The dimensions cover sensory content (*Color*, *Taste/Smell*, *Tactile*, *Visual Form*, *Auditory*), motor content (*Self-Motion*), internal and evaluative content (*Emotion*, *Polarity*, *Morality*, *Thought*), social content (*Social*), and magnitude-related content (*Space*, *Quantity*, *Time*). Using the original prompt "I relate this word to [*X*]," we evaluated the same 21 models from previous experiments over 10 shuffled runs each. Performance is reported as Mean r, the Pearson correlation between model and human ratings averaged across all 14 dimensions.

Results. The results differ sharply from those of property generation (compare Figures 2 and 4). On this task, recent frontier LLMs align much more closely with human judgments: the strongest model reaches Mean $r \approx 0.76$ against an estimated human ceiling of about 0.87, and alignment improves substantially with model scale. This contrast holds when controlling for the word pool, the number of dimensions and the human ceilings: on the 93 words shared by both tasks, models reach 78.4% of the human ceiling in rating but only 31.8% in property generation (Appendix I). It suggests that the grounding gap is not due to an inability to recognize dimensions such as sensation, emotion, or sociality: current LLMs can recover these dimensions when they are made explicit in the task, but do not recruit them in a human-like way when asked to generate the content of abstract concepts freely. Extended experimental details and results are provided in Appendix D.

5 Mechanistic analysis

Having established that LLMs can predict grounding dimension ratings, we investigate whether they internally encode features tracking these dimensions by analyzing the 4B and 12B Gemma 3 instruction-tuned models using Gemma Scope 2 sparse autoencoders (SAEs) [McDougall et al., 2025]. We isolate latent features correlated with grounding dimensions, focusing on those that align strongly with human experiential categories. For some dimensions, the best-matching feature correlates with the human ratings more strongly than the same model's explicit ratings do. Because

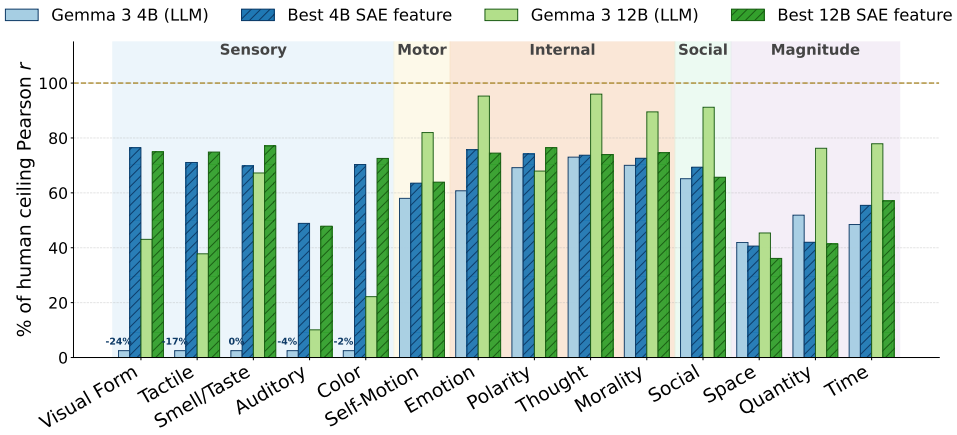


Figure 5: Percentage of mean r for all LLMs and SAE features vs. the estimated human ceiling mean r on the rating experiment dimensions.

the analysis covers two models from a single family, we treat these results as exploratory. While steering these representations predictably increases the generation of target-category properties, we view this merely as a methodological sanity check; artificially amplifying feature activations does not close the grounding gap, as it overrides rather than reflects the model's natural internal structure.

5.1 Methodology

Feature identification. We ask whether the model contains SAE features that correspond to grounding categories or sub-groups beyond individual words. First, we collect a pool of 425 labeled nouns from four psycholinguistic norm databases [Lynott et al., 2020, Diveica et al., 2023, Scott et al., 2019, Brysbaert et al., 2014], each associated with one of the 4 grounding dimensions of interest (sensorimotor, internal-state, social, and abstract content). For each noun w, we generate 10 English sentences ending in w using Gemini-3.1-pro-preview, and then remove w, creating a form of cloze (fill-the-blank) questions to study model activations when generating w. We record the SAE activation at the final token of these cloze questions, i.e. the position where the model is about to generate w, and take the per-feature median across the ten sentences as the noun's activation signature. A feature enters category c's candidate pool if it has non-zero median activation on at least five distinct nouns from c. The approach is similar to the one used by [Sofroniew et al., 2026] for emotion vectors. Construction details are reported in Appendix H.1. This approach gives us a set of candidate features for all 4 grounding dimensions.

Tracking grounding dimension norms with SAE features. The previous stage is designed to select features that their activations are correlated with generation of nouns related to their category, but their activations are not necessarily aligned with humans. Our goal is to find features that not only respond to a grounding dimension, but do so in a way that is aligned with human norms in a grounding dimension. To achieve this we evaluate these candidate features on the same dataset we used in our rating experiment in Section 4. For the *feature* measurement, the original rating prompts are unsuitable: they target a specific dimension and elicit a numeric answer, whereas our candidate features were identified in the regime of free noun generation. We instead use two simple and generic property-generation prompts and the original complex property-generation prompt from Section 3 and average the results, record each candidate feature's activation at the final prompt position, and correlate it with the human ratings on each dimension. This measures how well the feature itself tracks the human grounding signal while simulating the property-generation task.

5.2 Results

Overall, both models show a rather strong behavioral alignment with human ratings across most grounding dimensions (Figure 5). Sensory dimensions are the main exception: model ratings remain far below the human ceiling, which is consistent with the findings of Xu et al. [2025]. Nevertheless,

the SAE analysis identifies features that correlate with all evaluated dimensions, including the behaviorally weaker sensory ones. For most dimensions, the best-matching SAE features reach Pearson correlations in the range of $r = 0.6$ to 0.8, suggesting that these grounding-related dimensions are at least partially reflected in the models' internal representations. The main exception is the magnitude category, which includes space, quantity, and time and shows weaker feature alignment, possibly reflecting the more abstract nature of these dimensions. Overall, these results suggest that the models contain internal features that track several sub-dimensions of grounding, with these features being aligned with human judgments to some extent, and should not be interpreted as complete or fully human-like representations.

5.3 Feature steering

To interpret and validate these features, we conduct a steering experiment on Gemma 3 4B using the highest-correlation SAE feature for each dimension: sensory, motor, internal, social, and magnitude. We replicate Experiment 1 (Section 3.2) under feature steering, using feature clamping [Templeton et al., 2024] to a fixed high activation value. Steering increases target-category property generation in all cases: sensory (+5.3% *Sensorimotor*), motor (+4.7% *Sensorimotor*), internal (+9.4% *Internal*), social (+14.7% *Social*), and magnitude (+1.7% *Verbal*). These effects support the feature–category associations identified in the correlation analysis. We further inspect the highest-activating examples for each feature using Neuronpedia,[2] and find that the features generally activate on words related to their grounding dimensions. However, they are not perfectly category-specific: some also activate on words outside the narrow target dimension, suggesting that they capture broader semantic patterns rather than clean, isolated grounding categories or sub-dimensions. For details, see Appendix H.2.

6 Discussion

Grounding is important for humans because it connects language to lived experience and the physical world. As LLMs take roles originally meant for humans and move closer to agentic autonomy, their apparent differences from humans in grounding acquire some significance. The grounding gap we identified may not go away through scaling and text-based training alone, since (a) among the instruction-tuned models we tested, scaling [Hoffmann et al., 2022, Wei et al., 2022] does not appear to help, with state-of-the-art models remaining remarkably far from human ceilings, and (b) the excessive reliance on verbal associations we observe is consistent with training that is focused on prediction from linguistic context. Our base-model analysis (Appendix M) shows that post-training also contributes: pretrained-only checkpoints are moderately more aligned, although they retain the same qualitative profile. We cannot cleanly separate the contributions of pretraining and post-training, and our evidence directly supports a grounding gap in current post-trained models. Genuine progress on this front may require fundamental shifts, such as novel training techniques or alternative architectures [Fung et al., 2025, Brooks, 1991].

In the two Gemma 3 models we analyzed, some SAE features track grounding dimensions, in some cases more closely than the model's own explicit ratings. However, these features do not exclusively encode sensory, internal, or social experiences (see Appendix H.2 for details). According to one mainstream theory of grounded cognition [Barsalou, 1999], true conceptual understanding requires the capacity to internally simulate physical experiences. Whether LLMs possess, or can acquire through training, internal circuits capable of such simulation remains an open challenge for future research.

Limitations. Our behavioral analysis is limited by the generality of the property-generation experiments we replicate, although both are well established in the literature. To our knowledge, these are the only extant property-generation studies on abstract-concept grounding with both openly available stimulus sets and coding taxonomies suitable for direct replication and comparison. The fact that we observe highly consistent results across these two independently designed experiments, with different stimuli, taxonomies, and participant samples, suggests that the observed gap is not specific to a single experimental setup.

A second limitation is the use of LLMs as coders, which could introduce model-specific biases into the estimated category frequencies and correlations. We mitigate this by validating candidate coders

[2] Neuronpedia is an open-source platform for SAE interpretability and visualization: `https://www.neuronpedia.org`.

against human-annotated word-to-category pairs and selecting models with human-level or stronger agreement. We also show in Section 3.3 that replacing the LLM coder with human coders changes the main metrics only marginally, with differences far smaller than the human–model gaps reported above. This makes it highly unlikely that our conclusions are driven by the coding procedure.

Third, our base-versus-instruct comparison is not a clean causal ablation of post-training, because base models need few-shot demonstrations to follow the task, and these demonstrations can themselves shift the category distribution (Appendix M). The post-training recipe is also undisclosed for most models we study, and varies widely across those that document it, so we cannot attribute behavioral shifts to specific post-training steps.

Finally, our SAE analysis is limited to Gemma models due to the limited availability of high-quality SAEs for recent LLMs and the cost of scaling the analysis. Both models belong to the same family and share data, architecture, and training pipeline, so the features we find may reflect family-specific structure rather than a pattern that generalizes across LLMs. More generally, SAE features are interpretable approximations of model representations, not complete decompositions. The absence of a feature in an SAE therefore does not guarantee that the corresponding information is absent from the underlying model. We therefore treat the SAE results as exploratory evidence that complements the behavioral findings, rather than as the basis for strong mechanistic claims.

7 Conclusion

We investigate how well frontier LLMs are cognitively aligned with humans in their grounding of abstract concepts. We replicate two property-generation experiments on 21 LLMs, and find strong evidence that **LLMs are not grounded in the same way humans are** — but of course they were not trained to be. In particular, LLMs exhibit systematic under-production of internal-state properties and over-production of verbal associations, and overall very poor Pearson correlation with humans, while they correlate well with each other. These failures are uniform across the instruction-tuned LLM families we tested and do not improve with scale. In a replicated rating experiment, LLMs exhibit a reasonable understanding of concept categories, and in Gemma-3-4B and Gemma-3-12B we find SAE features whose activations track these categories, suggesting that grounding-related information can be present internally even when it is not expressed in free generation. Going forward, our results suggest that grounding alignment with humans may be an interesting and novel quantity that is invisible to current testing benchmarks, and may be worth watching in the future. And finally, our results suggest that further scaling or better text-based training alone may not close the gap, so we may need to rethink the way we design and train LLMs.

Acknowledgments and Disclosure of Funding

O. S. Chlapanis was supported by project MIS 5154714 of the National Recovery and Resilience Plan Greece 2.0, funded by the European Union under NextGenerationEU. O. Menis Mastromichalakis was supported by the project DECOLLAGE (ERC-2022-CoG 101088763) and through national funds and co-funded EU funds under UID/50008: Instituto de Telecomunicações. C. H. Papadimitriou was supported by NSF grants CCF-2212233 and CCF-2332922, by ONR MURI grant N00014-25-1-2116, and by the ERC grant NEAGT (101267783).

References

M. Abdou, A. Kulmizev, D. Hershcovich, S. Frank, E. Pavlick, and A. Søgaard. Can language models encode perceptual structure without grounding? a case study in color. In *Proceedings of the 25th Conference on Computational Natural Language Learning (CoNLL)*, 2021. URL https://aclanthology.org/2021.conll-1.9/.

L. W. Barsalou. Perceptual symbol systems. *Behavioral and Brain Sciences*, 22(4):577–660, 1999. doi: 10.1017/S0140525X99002149.

L. W. Barsalou. Grounded cognition. In M. C. Frank and A. Majid, editors, *Open Encyclopedia of Cognitive Science*. MIT Press, 2026. URL https://oecs.mit.edu/pub/9iq4376o.

L. W. Barsalou and K. Wiemer-Hastings. Situating abstract concepts. In D. Pecher and R. A. Zwaan, editors, *Grounding Cognition: The Role of Perception and Action in Memory, Language, and Thought*, pages 129–163. Cambridge University Press, 2005. doi: 10.1017/CBO9780511499968.007.

E. M. Bender and A. Koller. Climbing towards NLU: On meaning, form, and understanding in the age of data. In D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.463. URL https://aclanthology.org/2020.acl-main.463/.

E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097. doi: 10.1145/3442188.3445922. URL https://doi.org/10.1145/3442188.3445922.

A. M. Borghi, F. Binkofski, C. Castelfranchi, F. Cimatti, C. Scorolli, and L. Tummolini. The challenge of abstract concepts. *Psychological Bulletin*, 143(3):263–292, 2017. doi: 10.1037/bul0000089.

T. Bricken, A. Templeton, J. Batson, B. Chen, A. Jermyn, T. Conerly, N. Turner, C. Anil, C. Denison, A. Askell, R. Lasenby, Y. Wu, S. Kravec, N. Schiefer, T. Maxwell, N. Joseph, Z. Hatfield-Dodds, A. Tamkin, K. Nguyen, B. McLean, J. E. Burke, T. Hume, S. Carter, T. Henighan, and C. Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. https://transformer-circuits.pub/2023/monosemantic-features/index.html.

R. A. Brooks. Intelligence without representation. *Artificial Intelligence*, 47(1):139–159, 1991. ISSN 0004-3702. doi: https://doi.org/10.1016/0004-3702(91)90053-M. URL https://www.sciencedirect.com/science/article/pii/000437029190053M.

M. Brysbaert, A. B. Warriner, and V. Kuperman. Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, 46(3):904–911, 2014. doi: 10.3758/s13428-013-0403-5.

H. D. Critchley, S. Wiens, P. Rotshtein, A. Öhman, and R. J. Dolan. Neural systems supporting interoceptive awareness. *Nature Neuroscience*, 7(2):189–195, 2004. doi: 10.1038/nn1176.

V. Diveica, P. M. Pexman, and R. J. Binney. Quantifying social semantics: An inclusive definition of socialness and ratings for 8,388 English words. *Behavior Research Methods*, 55:461–473, 2023. doi: 10.3758/s13428-022-01810-x.

A. Ettinger. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48, 2020. doi: 10.1162/tacl_a_00298.

P. Fung, Y. Bachrach, A. Celikyilmaz, K. Chaudhuri, D. Chen, W. Chung, E. Dupoux, H. Gong, H. Jégou, A. Lazaric, et al. Embodied ai agents: Modeling the world. *arXiv preprint arXiv:2506.22355*, 2025.

W. Gurnee and M. Tegmark. Language models represent space and time. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun, editors, *International Conference on Learning Representations*, volume 2024, pages 2483–2503, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/0a6059857ae5c82ea9726ee9282a7145-Paper-Conference.pdf.

M. Harpaintner, N. M. Trumpp, and M. Kiefer. The semantic content of abstract concepts: A property listing study of 296 abstract words. *Frontiers in Psychology*, 9:1748, 2018. doi: 10.3389/fpsyg.2018.01748.

O. Hauk, I. Johnsrude, and F. Pulvermüller. Somatotopic representation of action words in human motor and premotor cortex. *Neuron*, 41(2):301–307, 2004. doi: 10.1016/S0896-6273(03)00838-9.

J. Hoffmann, S. Borgeaud, A. Mensch, E. Buchatskaya, T. Cai, E. Rutherford, D. de Las Casas, L. A. Hendricks, J. Welbl, A. Clark, T. Hennigan, E. Noland, K. Millican, G. van den Driessche, B. Damoc, A. Guy, S. Osindero, K. Simonyan, E. Elsen, O. Vinyals, J. W. Rae, and L. Sifre. Training compute-optimal large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.

R. Huben, H. Cunningham, L. R. Smith, A. Ewart, and L. Sharkey. Sparse autoencoders find highly interpretable features in language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=F76bwRSLeK.

A. E. Kelly, Y. N. Kenett, J. D. Medaglia, J. J. Reilly, P. Dudhat, and E. G. Chrysikou. Conceptual structure of emotions. *Emotion*, 24(6):1550–1561, 2024. doi: 10.1037/emo0001327.

G. Lakoff and M. Johnson. *Metaphors We Live By*. University of Chicago Press, 2003. ISBN 9780226470993. doi: 10.7208/chicago/9780226470993.001.0001. URL http://dx.doi.org/10.7208/chicago/9780226470993.001.0001.

B. Li. Incompressible knowledge probes: Estimating black-box llm parameter counts via factual capacity. *arXiv preprint arXiv:2604.24827*, 2026.

K. Li, A. K. Hopkins, D. Bau, F. Viégas, H. Pfister, and M. Wattenberg. Emergent world representations: Exploring a sequence model trained on a synthetic task. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=DeG07_TcZvT.

T. Lieberum, S. Rajamanoharan, A. Conmy, L. Smith, N. Sonnerat, V. Varma, J. Kramar, A. Dragan, R. Shah, and N. Nanda. Gemma scope: Open sparse autoencoders everywhere all at once on gemma 2. In Y. Belinkov, N. Kim, J. Jumelet, H. Mohebbi, A. Mueller, and H. Chen, editors, *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 278–300, Miami, Florida, US, Nov. 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.blackboxnlp-1.19. URL https://aclanthology.org/2024.blackboxnlp-1.19/.

D. Lynott, L. Connell, M. Brysbaert, J. Brand, and J. Carney. The Lancaster sensorimotor norms: multidimensional measures of perceptual and action strength for 40,000 English words. *Behavior Research Methods*, 52:1271–1291, 2020. doi: 10.3758/s13428-019-01316-z.

M. Mancano and C. Papagno. Emotional and social dimension of abstract concepts meet with interoception in right anterior insula. *Journal of Neuroscience*, 46(2):e0238252025, 2026. doi: 10.1523/JNEUROSCI.0238-25.2025.

C. McDougall, A. Conmy, J. Kramár, T. Lieberum, S. Rajamanoharan, and N. Nanda. Gemma scope 2 - technical paper. Technical report, Google DeepMind, Sept. 2025. https://storage.googleapis.com/deepmind-media/DeepMind.com/Blog/gemma-scope-2-helping-the-ai-safety-community-deepen-understanding-of-complex-language-model-behavior/Gemma_Scope_2_Technical_Paper.pdf.

K. McRae, G. S. Cree, M. S. Seidenberg, and C. McNorgan. Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods*, 37(4):547–559, 2005. doi: 10.3758/BF03192726.

N. Nanda, A. Lee, and M. Wattenberg. Emergent linear representations in world models of self-supervised sequence models. In Y. Belinkov, S. Hao, J. Jumelet, N. Kim, A. McCarthy, and H. Mohebbi, editors, *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 16–30, Singapore, Dec. 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.blackboxnlp-1.2. URL https://aclanthology.org/2023.blackboxnlp-1.2/.

N. Panickssery, N. Gabrieli, J. Schulz, M. Tong, E. Hubinger, and A. Turner. Steering LLaMA 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15504–15522. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.828.

S. Pezzelle, E. Takmaz, and R. Fernández. Word representation learning in multimodal pre-trained transformers: An intrinsic evaluation. *Transactions of the Association for Computational Linguistics (TACL)*, 9:1563–1579, 2021. URL https://aclanthology.org/2021.tacl-1.93/.

F. Pulvermüller. Brain mechanisms linking language and action. *Nature Reviews Neuroscience*, 6(7): 576–582, 2005. doi: 10.1038/nrn1706.

G. G. Scott, A. Keitel, M. Becirspahic, B. Yao, and S. C. Sereno. The Glasgow norms: ratings of 5,500 words on nine scales. *Behavior Research Methods*, 51:1258–1270, 2019. doi: 10.3758/s13428-018-1099-3.

N. Sofroniew, I. Kauvar, W. Saunders, R. Chen, T. Henighan, S. Hydrie, C. Citro, A. Pearce, J. Tarng, W. Gurnee, J. Batson, S. Zimmerman, K. Rivoire, K. Fish, C. Olah, and J. Lindsey. Emotion concepts and their function in a large language model, 2026. URL https://arxiv.org/abs/2604.07729.

S. Suresh, K. Mukherjee, X. Yu, W.-C. Huang, L. Padua, and T. T. Rogers. Conceptual structure coheres in human cognition but not in large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 722–738, 2023. URL https://aclanthology.org/2023.emnlp-main.47/.

A. Templeton, T. Conerly, J. Marcus, J. Lindsey, T. Bricken, B. Chen, A. Pearce, C. Citro, E. Ameisen, A. Jones, H. Cunningham, N. L. Turner, C. McDougall, M. MacDiarmid, C. D. Freeman, T. R. Sumers, E. Rees, J. Batson, A. Jermyn, S. Carter, C. Olah, and T. Henighan. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Transformer Circuits Thread*, 2024. URL https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html.

C. Tigges, O. J. Hollinsworth, A. Geiger, and N. Nanda. Language models linearly represent sentiment. In Y. Belinkov, N. Kim, J. Jumelet, H. Mohebbi, A. Mueller, and H. Chen, editors, *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 58–87, Miami, Florida, US, Nov. 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.blackboxnlp-1.5. URL https://aclanthology.org/2024.blackboxnlp-1.5/.

J. Troche, S. J. Crutch, and J. Reilly. Defining a conceptual topography of word concreteness: Clustering properties of emotion, sensation, and magnitude among 750 English words. *Frontiers in Psychology*, 8:1787, 2017. doi: 10.3389/fpsyg.2017.01787.

S. Trott. Can large language models help augment english psycholinguistic datasets? *Behavior Research Methods*, 56(6):6082–6100, Jan. 2024. doi: 10.3758/s13428-024-02337-z. URL http://dx.doi.org/10.3758/s13428-024-02337-z.

A. M. Turner, L. Thiergart, G. Leech, D. Udell, J. J. Vazquez, U. Mini, and M. MacDiarmid. Steering language models with activation engineering, 2024. URL https://arxiv.org/abs/2308.10248.

C. Villani, L. Lugli, M. T. Liuzza, and A. M. Borghi. Varieties of abstract concepts and their multiple dimensions. *Language and Cognition*, 11(3):403–430, 2019. doi: 10.1017/langcog.2019.23.

C. Wang, Y. Zhang, R. Yu, Y. Zheng, L. Gao, Z. Song, Z. Xu, G. Xia, H. Zhang, D. Zhao, and X. Chen. Do LLMs "feel"? emotion circuits discovery and control, Oct. 2025a. URL https://arxiv.org/abs/2510.11328.

Y. Wang, D. Liang, and Y. Zeng. Cognitive alignment between humans and llms across multimodal domains, Jan. 2025b. URL https://www.researchsquare.com/article/rs-5736241/v1. Research Square preprint.

J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, E. H. Chi, T. Hashimoto, O. Vinyals, P. Liang, J. Dean, and W. Fedus. Emergent abilities of large language models. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856. URL https://openreview.net/forum?id=yzkSU5zdwD. Survey Certification.

X. Wu, H. Wang, Z. Yan, X. Tang, P. Xu, W.-T. Siok, P. Li, J.-H. Gao, B. Lyu, and L. Qin. Ai shares emotion with humans across languages and cultures, 2025. URL https://arxiv.org/abs/2506.13978.

Q. Xu, Y. Peng, S. A. Nastase, M. Chodorow, M. Wu, and P. Li. Large language models without grounding recover non-sensorimotor but not sensorimotor features of human concepts. *Nature Human Behaviour*, 9:1871–1886, 2025. doi: 10.1038/s41562-025-02203-8.

L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. Xing, H. Zhang, J. Gonzalez, and I. Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/91f18a1287b398d378ef22505bf41832-Paper-Datasets_and_Benchmarks.pdf.

A LLMs-as-coders for property-generation experiments

A.1 Benchmarking LLMs-as-coders

In this section, we evaluate the performance of LLMs on the downstream task of automated property coding. Our methodology builds upon the established "LLM-as-a-judge" paradigm [Zheng et al., 2023]. However, deploying LLMs as coders presents a more straightforward, and arguably more reliable, framework than using them as evaluative judges. A typical LLM-judge must typically surpass the candidate model in reasoning capabilities to accurately score its output, while a LLM-coder merely has to be able to classify the properties properly. This coding task is fundamentally a simplified variation of the rating task detailed in Section 4, a domain where we have already demonstrated robust LLM performance. Furthermore, because this task involves classification rather than quality evaluation, it avoids the self-enhancement bias typically observed in LLM-as-judge setups [Zheng et al., 2023]. This lack of self-bias makes it methodologically sound to use the same model to both generate and subsequently code its own responses.

Data collection and benchmark creation. To rigorously benchmark the models as coders, we needed a substantial dataset of coded properties. Fortunately, Kelly et al. [2024] provides a large-scale public dataset: the complete $49{,}942$ word-property pairs produced in their experiment. For experiment 1, no such word-property pairs were provided, so two of the authors independently labeled 2,077 pairs (agreement 68.9%, Cohen's $\kappa = 0.59$), which were generated by the three frontier LLMs (Claude-Opus 4.6, GPT-5.4, Gemini-3.1-Pro). We utilized these two datasets as our primary benchmarks to identify the optimal LLM coder for each experiment.

Validation metrics and human reference. For each candidate coder we report percentage agreement with the ground-truth label and Cohen's κ. The human reference for Experiment 1 is the 76.8% two-coder joint agreement reported by Harpaintner et al. [2018] on their original validation subset, and our two annotators agreed on 68.9%; Harpaintner et al. [2018] do not report a κ ceiling, so we leave that cell blank. For Experiment 2, the human reference is the 64.25% two-coder agreement and $\kappa = 0.52$ reported by Kelly et al. [2024].

To ensure deterministic coding results, we use temperature $T = 0.0$ for all LLM-coders.

Table 2: Experiment 1 coder leaderboard. Six LLM coders predict the 4-category property label on the 2077-pair ground truth. % Agreement is overall percentage agreement, κ is Cohen's kappa. Per-category columns are recall, i.e. the percentage of pairs labeled by the human as that category that the LLM also labeled as that category. †Human ceiling is the 76.8% two-coder joint agreement reported by Harpaintner et al. [2018]. **Gemini-2.5-Flash-Lite** was selected as the coder for all experiments with this dataset, unless otherwise specified.

Coder	% Agreement	κ	Sensorimotor	Emotion	Social	Association
Human ceiling†	*76.8*	–	–	–	–	–
Gemini 3.1 Pro	68.6	0.510	69.6	62.6	55.5	72.1
Claude Sonnet 4.6	67.6	0.492	55.6	68.5	57.3	76.9
Gemini 2.5 Flash-Lite	67.2	0.505	66.2	75.9	62.1	67.1
Gemini 2.5 Flash	65.6	0.484	77.0	60.6	63.0	60.3
GPT-OSS 120B	64.4	0.477	76.0	66.5	68.6	55.9
GPT-5 mini	64.2	0.486	87.5	68.0	67.8	48.6

Error margins for Gemini-2.5-Flash-Lite as a coder. The Experiment 1 coder benchmark in Table 2 measures property-level agreement, but the Experiment 1 leaderboard actually consumes the per-word category-frequency vector and its Pearson r against the human norms. To estimate the error coder substitution introduces at that leaderboard level, we re-coded the single hand-coded run of three reference generation models (Claude Opus 4.6, Gemini 3.1 Pro, GPT-5.4) with both Flash-Lite and the human ground truth on the exact same (word, property) pairs and bootstrapped the difference over words (1000 paired iterations, same resampled words for both coders within an iteration so that $\Delta = (\text{Flash-Lite} - \text{Human})$ is a paired statistic). The Mean r a human coder would have produced has a cross-model average difference from the LLM-coder values of $\Delta = +0.014 \pm 0.019$, with

Table 3: Experiment 2 coder leaderboard. % Agreement and κ are computed against the 47,123 gold human-coded labels. †Human ceiling is the 64.25% two-coder agreement and $\kappa = 0.52$ reported by Kelly et al. [2024]. **Gemini-2.5-Flash** was selected as the coder for all experiments with this dataset, unless otherwise specified.

Coder	% Agreement	κ
Human ceiling†	*64.25*	*0.52*
Gemini 2.5 Flash	69.2	0.577
GPT-OSS 120B	64.1	0.516
Gemma 3 12B	60.9	0.476
Gemini 2.5 Flash-Lite	60.0	0.482

every model 95% CI containing zero. On the same property pairs the cross-model average frequency Δ's are $+2.3\%$ Sensorimotor, -0.1% Internal, -0.5% Social, -1.7% Verbal Association, with every per-model 95% CI containing zero. So the per-category attribution rates a human coder would have produced are within roughly ± 2 percentage points of the Flash-Lite values.

B Extended results for Experiment 1

Here we present extended quantitative results for the first experiment (Section 3.2) based on the Harpaintner et al. [2018] dataset.

B.1 Detailed model setup

We evaluate 21 LLMs spanning closed-weight frontier and open-weight families. All generations use each provider's default sampling temperature. Coding is performed at $T = 0.0$ for determinism (Section A.1).

The closed-weight models are accessed through their official APIs: **Anthropic** (Claude Opus 4.6, Claude Sonnet 4.6, Claude Haiku 4.5), **OpenAI** (GPT-5.4), and **Google** (Gemini 3.1 Pro, Gemini 3 Flash, Gemini 3.1 Flash-Lite, Gemini 2.5 Flash, Gemini 2.5 Flash-Lite). The open-weight models are run on 8x Quadro RTX 6000 GPUs, with model-parallel sharding for the 70B and 120B systems: **Meta** (Llama 3.1 70B Instruct, Llama 3.1 8B Instruct), **OpenAI open-weights** (GPT-OSS 120B, GPT-OSS 20B), **Google open-weights** (Gemma-3-27B-IT, Gemma-3-12B-IT, Gemma-3-4B-IT), and **Alibaba Qwen** (Qwen3 8B, Qwen3 4B, Qwen3-VL 30B, Qwen3-VL 8B, Qwen3-VL 4B).

Each model produces 10 independent runs over the full 293-word Harpaintner stimulus set. Per-word category-frequency vectors are averaged across runs before correlating with the human norms; $\pm$ values reported in the leaderboard are bootstrap standard deviations over 1000 resamples of the runs (Section E.2).

B.2 Detailed results for all 21 models

Figure 6 presents the frequency profiles for all models. Table 4 details the performance of all 21 models evaluated in our panel. Smaller models occasionally outperform substantially larger ones, suggesting that simply increasing parameter count does not organically close the grounding gap.

C Extended results for Experiment 2

Here we present extended quantitative results for the second experiment (Section 3.2) based on the Kelly et al. [2024] dataset. The Kelly stimulus set partitions the 354 target words into three matched conditions: 118 abstract emotion words, 117 abstract non-emotion words, and 119 concrete words. The main text reports alignment on the combined 235-word abstract subset (emotion plus non-emotion). For completeness, in this appendix we also report the leaderboard restricted to the 119-word concrete subset, which serves as a control condition: concrete nouns are the easy case for property generation, since they admit clearly perceptual properties that all candidate models recover well.

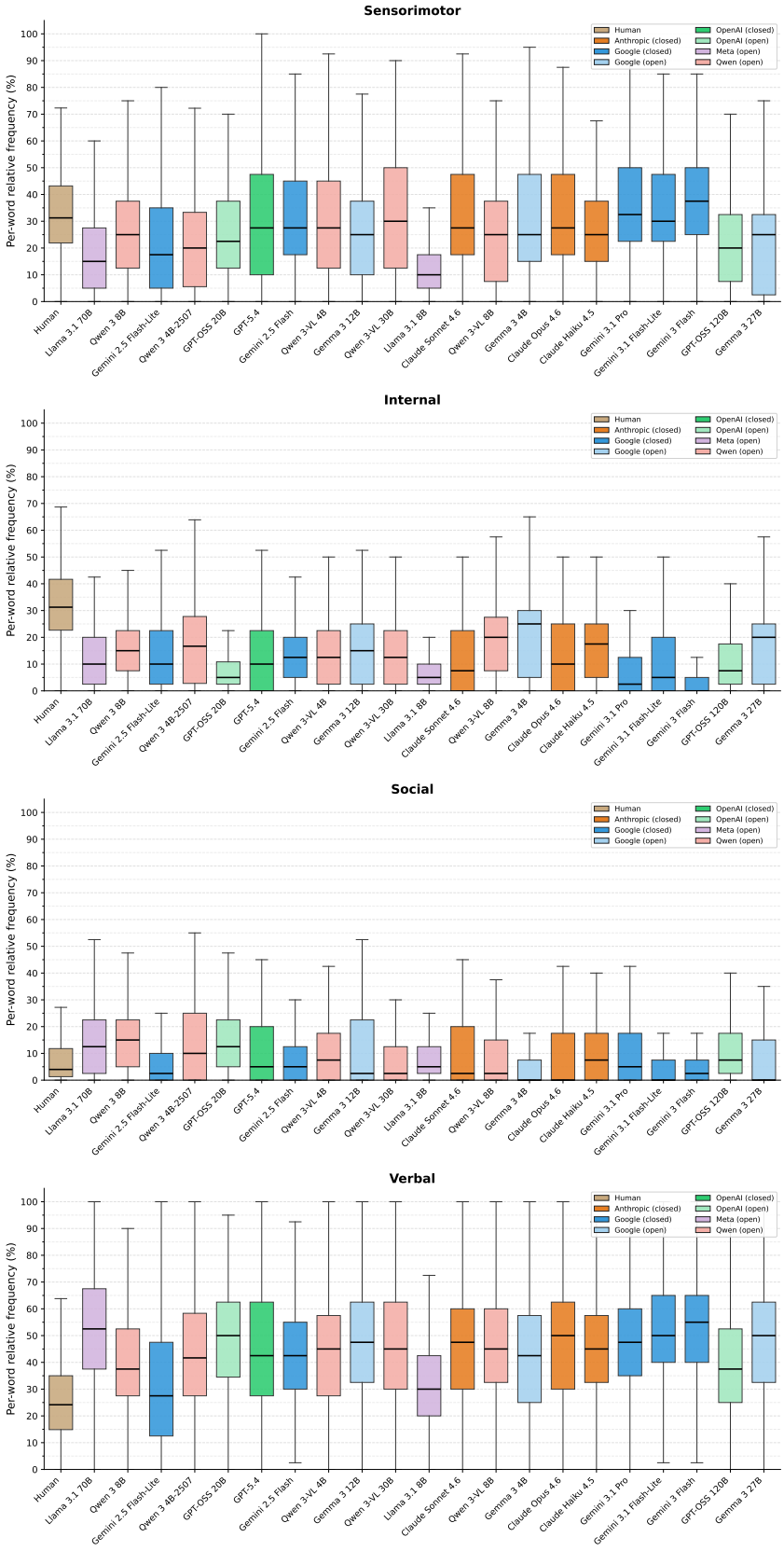


Figure 6: Property-frequency boxplots for all 21 LLMs, sorted by Mean r.

Table 4: Full experiment 1 leaderboard, all 21 models in the panel, sorted by Mean r. Ten generation runs per model. Bold marks the column max.

Rank	Model	Mean r	Sensorimotor	Internal	Social	Verbal
	Estimated human ceiling	*0.974*	*0.966*	*0.964*	*0.971*	*0.992*
1	Llama 3.1 70B	**0.375** $\pm$ 0.011	0.344	**0.464**	0.424	0.267
2	Qwen3 8B	0.373 $\pm$ 0.015	**0.351**	0.390	**0.462**	**0.288**
3	Gemini 2.5 Flash-Lite	0.348 $\pm$ 0.009	0.276	0.382	0.456	0.278
4	Qwen3 4B	0.314 $\pm$ 0.013	0.335	0.405	0.318	0.200
5	GPT-OSS 20B	0.302 $\pm$ 0.013	0.326	0.255	0.429	0.196
6	GPT-5.4	0.301 $\pm$ 0.005	0.332	0.258	0.400	0.213
7	Gemini 2.5 Flash	0.296 $\pm$ 0.012	0.291	0.310	0.364	0.218
8	Qwen3-VL 4B	0.295 $\pm$ 0.007	0.344	0.251	0.326	0.260
9	Gemma-3-12B	0.292 $\pm$ 0.005	0.277	0.283	0.357	0.251
10	Qwen3-VL 30B	0.292 $\pm$ 0.007	0.278	0.240	0.412	0.237
11	Llama 3.1 8B	0.283 $\pm$ 0.014	0.186	0.356	0.406	0.185
12	Claude Sonnet 4.6	0.277 $\pm$ 0.006	0.287	0.242	0.352	0.225
13	Qwen3-VL 8B	0.273 $\pm$ 0.009	0.313	0.256	0.325	0.199
14	Gemma-3-4B	0.270 $\pm$ 0.006	0.279	0.266	0.291	0.244
15	Claude Opus 4.6	0.264 $\pm$ 0.004	0.203	0.205	0.362	0.284
16	Claude Haiku 4.5	0.258 $\pm$ 0.008	0.250	0.196	0.310	0.277
17	Gemini 3.1 Pro	0.256 $\pm$ 0.008	0.307	0.163	0.383	0.171
18	Gemini 3.1 Flash-Lite	0.223 $\pm$ 0.007	0.264	0.134	0.279	0.216
19	Gemini 3 Flash	0.216 $\pm$ 0.013	0.251	0.182	0.294	0.137
20	GPT-OSS 120B	0.199 $\pm$ 0.010	0.210	0.238	0.230	0.116
21	Gemma-3-27B	0.196 $\pm$ 0.005	0.262	0.115	0.262	0.144

Properties are classified by Gemini-2.5-Flash into the four superordinate Kelly categories: **Taxonomic**, **Entity**, **Situation**, and **Introspective**. Each model produces 10 runs over the full 354-word stimulus set; per-concept category-frequency vectors are averaged across runs before correlating with the human reference distribution. The primary metric is Mean r between the per-concept distributions of the model and the human aggregate, computed separately on the abstract and concrete subsets.

Figure 7 replicates the scaling plot for the Experiment 2 setup: even the strongest models reach only Mean $r \approx 0.37$ on the abstract subset, far below the human-to-human ceiling of $r \approx 0.92$, and there is no monotonic trend in parameter count. Table 5 lists every model on both subsets sorted by abstract Mean r. The pattern mirrors Experiment 1: the abstract-subset gap remains roughly flat across model sizes and providers, while the same models reach much higher alignment ($r \in [0.36, 0.53]$) on the concrete subset. This contrast is consistent with the central claim of the paper: the gap is specific to abstract concepts, not a general property-generation deficit. Figure 8 confirms this from a different angle: on the concrete subset the human-to-model and model-to-model correlations sit in similar ranges, whereas on the abstract subset (Fig. 3 in the main text) models are much more correlated with each other than with humans.

D Rating experiment additional results

Task. We replicate the rating study of Troche et al. [2017] using 751 English nouns rated on 14 cognitive dimensions under the uniform framing *"I relate this word to [X]"*, with responses on a 1 to 7 Likert scale (1 = Strongly Disagree, 7 = Strongly Agree). Each model produces ten rating seeds; per-word ratings are averaged across seeds before correlating with the human norms. The primary metric is the Pearson r averaged over the 14 dimensions. The 14 dimensions are organised into five components: **Sensory** (Color, Taste/Smell, Tactile, Visual Form, Auditory), **Motor** (Self-Motion), **Internal** (Emotion, Polarity, Morality, Thought), **Social** (Social), and **Magnitude** (Space, Quantity, Time). Table 6 reports the per-component Pearson correlations alongside the overall Mean r for every model out of the 21.

Results. Three patterns stand out in Table 6, and each one provides a different angle on the grounding gap reported in the main text.

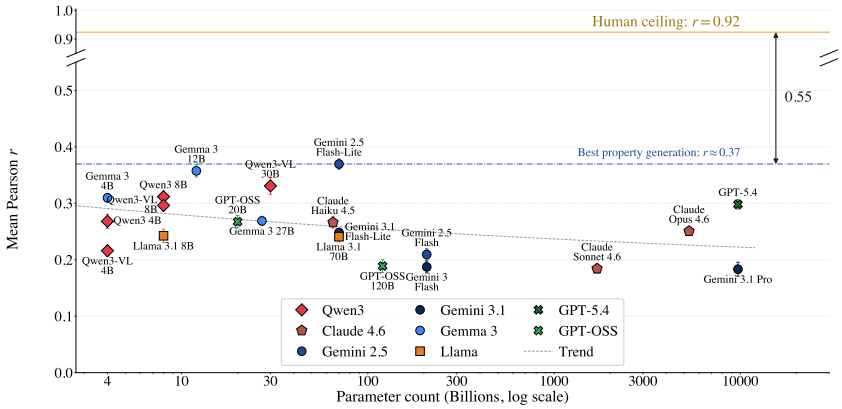


Figure 7: Mean r for all LLMs vs. the human ceiling on Experiment 2 [Kelly et al., 2024]. Parameter count for closed models is based on estimations from Li [2026]. The dashed curve shows the saturating fit to the 21 models (trend). The dash-dotted line marks the best model, and the arrow its distance to the human ceiling.

Table 5: Extended leaderboard on experiment 2 Kelly et al. [2024], for both the abstract and concrete subsets. Models are sorted by Mean r on the abstract subset.

Model	Abstract	Concrete
Human ceiling	*0.924 ± 0.006*	*0.920 ± 0.008*
Gemini 2.5 Flash-Lite	**0.370 ± 0.009**	0.488 ± 0.014
Gemma 3 12B	0.357 ± 0.011	0.393 ± 0.010
Qwen 3-VL 30B	0.331 ± 0.014	0.480 ± 0.015
Qwen 3 8B	0.312 ± 0.008	0.417 ± 0.016
Gemma 3 4B	0.310 ± 0.008	0.423 ± 0.015
GPT-5.4	0.298 ± 0.009	0.482 ± 0.012
Qwen 3-VL 8B	0.296 ± 0.010	0.393 ± 0.015
Gemma 3 27B	0.269 ± 0.008	0.447 ± 0.016
Qwen 3 4B	0.268 ± 0.013	0.360 ± 0.025
GPT-OSS 20B	0.267 ± 0.012	0.422 ± 0.028
Claude Haiku 4.5	0.266 ± 0.010	0.389 ± 0.017
Claude Opus 4.6	0.251 ± 0.007	0.467 ± 0.013
Gemini 3.1 Flash-Lite	0.248 ± 0.007	0.465 ± 0.012
Llama 3.1 8B	0.243 ± 0.012	0.449 ± 0.022
Llama 3.1 70B	0.241 ± 0.009	0.469 ± 0.017
Qwen 3-VL 4B	0.216 ± 0.010	0.391 ± 0.017
Gemini 2.5 Flash	0.209 ± 0.009	**0.531 ± 0.018**
GPT-OSS 120B	0.189 ± 0.012	0.401 ± 0.021
Gemini 3 Flash	0.187 ± 0.010	0.463 ± 0.022
Claude Sonnet 4.6	0.184 ± 0.007	0.436 ± 0.013
Gemini 3.1 Pro	0.183 ± 0.012	0.526 ± 0.018

The first observation is that the strongest models are close to the human ceiling on this task. Claude Sonnet 4.6 reaches Mean $r = 0.760$ against a cross-dimension human ceiling of approximately 0.87 (Appendix F). The same models top out at $r \approx 0.38$ on the property generation task of Experiment 1 against a ceiling of 0.97 (Table 4), a gap about five times larger to the human reference. We take the contrast as direct evidence for the central claim of the paper: when a grounding dimension is named explicitly in the prompt, current LLMs can recover it; the gap is therefore specific to the spontaneous property generation setting and is not a general inability to register sensorimotor, internal, or social content.

The second observation is that scaling helps recognition, but only on a subset of the components. Within most open-weight families we tested, the Sensory component improves dramatically with model size: Gemma scales from -0.086 at 4B to 0.704 at 27B, Llama scales from -0.161 at 8B

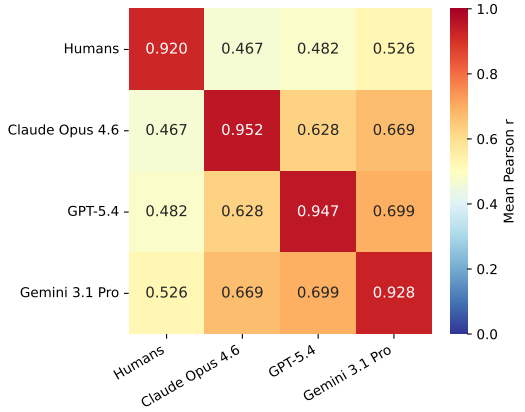


Figure 8: Experiment 2 properties correlation heatmap for the concrete word subset. The correlation between LLMs–humans is closer to LLMs-LLMs in this case, compared to Fig. 3.

Table 6: Rating experiment results for all 21 models. Mean r is the average of the 14 per-dimension Pearson correlations against the human ratings.

Model	Mean r	Sensory	Motor	Internal	Social	Magnitude
Claude Sonnet 4.6	0.760	0.839	0.583	0.787	0.827	0.628
Claude Opus 4.6	0.755	0.868	0.393	0.779	0.812	0.637
Gemini 3 Flash	0.730	0.889	0.189	0.750	0.793	0.600
Gemini 2.5 Flash	0.720	0.872	0.229	0.740	0.813	0.574
Gemini 3.1 Pro	0.719	0.883	0.119	0.753	0.790	0.576
GPT-5.4	0.717	0.846	0.432	0.733	0.737	0.567
Llama 3.1 70B	0.707	0.729	0.495	0.754	0.832	0.637
Gemma 3 27B	0.706	0.704	0.614	0.775	0.829	0.606
Qwen 3 8B	0.703	0.807	0.375	0.717	0.828	0.580
GPT-OSS 20B	0.701	0.757	0.513	0.712	0.825	0.613
GPT-OSS 120B	0.695	0.709	0.512	0.745	0.837	0.617
Gemini 2.5 Flash-Lite	0.679	0.719	0.434	0.743	0.787	0.572
Gemini 3.1 Flash-Lite	0.677	0.829	-0.058	0.686	0.794	0.618
Claude Haiku 4.5	0.672	0.801	0.071	0.711	0.768	0.571
Qwen 3-VL 30B	0.665	0.641	0.604	0.712	0.827	0.607
Gemma 3 12B	0.559	0.326	0.693	0.758	0.811	0.554
Qwen 3-VL 8B	0.517	0.300	0.557	0.693	0.764	0.551
Qwen 3 4B-2507	0.507	0.321	0.540	0.687	0.703	0.499
Qwen 3-VL 4B	0.457	0.192	0.539	0.682	0.740	0.479
Llama 3.1 8B	0.341	-0.161	0.528	0.736	0.710	0.464
Gemma 3 4B	0.297	-0.086	0.491	0.588	0.580	0.391

to 0.729 at 70B, and Qwen3 scales from 0.321 at 4B to 0.807 at 8B (Table 6). The Magnitude component also increases within each family. The Internal and Social components, by contrast, are essentially flat across model sizes: every model in Table 6 sits in the range $[0.58, 0.84]$ on these two components, regardless of parameter count or provider. Recognition of affective and social content is therefore not a capability bottleneck at any scale we tested. Perceptual grounding, on the other hand, is recoverable only at sufficient scale, and once that scale is reached the model can rate perceptual content as well as humans do.

The third observation is that the Motor sub-dimension is a persistent weakness that scaling does not close. As shown in Table 6, Motor scores are weaker overall, and some are very weak, with Gemini 3.1 Pro at 0.119, Gemini 3 Flash at 0.189, and Gemini 3.1 Flash-Lite even falling below zero at -0.058. More importantly, the within-family scaling that lifts Sensory does not lift Motor: Gemma improves only modestly from 0.491 at 4B to 0.614 at 27B (with 12B sitting higher than 27B at 0.693), Llama actually drops from 0.528 at 8B to 0.495 at 70B, and Qwen3 drops from 0.540 at 4B to 0.375 at 8B.

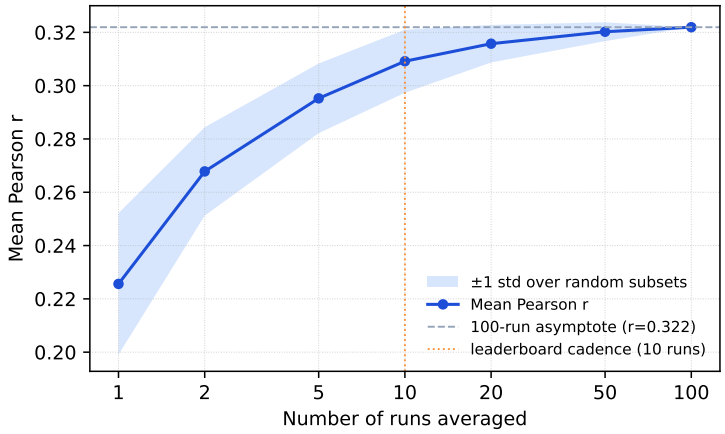


Figure 9: Mean Pearson r vs. number of runs averaged, on Experiment 1, for GEMMA-3-4B-IT coded by Gemini-2.5-Flash-Lite. Solid line: mean over 50 random subsets at each subset size; band: ± 1 std across subsets; dashed: 100-run asymptote; dotted: the 10-run cadence used by the Experiment 1 leaderboard.

E Variance and uncertainty in Experiment 1

The Experiment 1 leaderboard reports each model's Mean Pearson r as the point estimate from 10 independent generation runs, plus a bootstrap standard deviation. Two questions a reader is right to ask: (i) is a 10-run point estimate close to the asymptote that the same model would reach with infinitely many runs, and (ii) what does the bootstrap $\pm$ bar actually quantify? We answer both empirically by running GEMMA-3-4B-IT 100 times, treating that aggregate as the ground truth, and comparing it to what we would have estimated from random 10-run subsets of the same data.

E.1 Convergence: how accurate is the 10-run estimate?

Figure 9 plots Mean r as a function of the number of runs averaged, with shaded ± 1 std bands over 50 random subsets at each subset size. The estimate climbs monotonically toward the 100-run asymptote of $r = +0.322$. At the Experiment 1 leaderboard cadence of 10 runs, the average estimate is $r = +0.309$, lower than the asymptote by $\Delta r = +0.013$.

The curve has two practical implications. *For absolute r values on a single model*, a 10-run point estimate sits about 0.01 below the asymptote it would converge to, so quoted r values are slightly conservative. *For ranking models against each other*, this bias is approximately constant across models, so the relative ordering on the 10-run Experiment 1 leaderboard is stable.

E.2 Bootstrap calibration

The Experiment 1 leaderboard's $\pm$ value next to each model's Mean r is a bootstrap standard deviation computed from $N_{\text{boot}} = 1000$ resamples of the 10 runs with replacement. For each resample we re-aggregate the per-word frequencies, recompute Mean r against the human norms, and take the standard deviation of the resulting distribution.

F Estimation of human correlation ceilings

Each experiment reports a per-category Pearson r human ceiling, defined as the correlation between two independent panels of N raters scored on the same stimuli. This is the largest r a noiseless model could expect to reach against the published norms, which are themselves panel means. We obtain it in one of two ways depending on what each original study released. Kelly et al. [2024] provides participant-level responses, so we compute the ceiling by direct resampling. Harpaintner et al. [2018] and Troche et al. [2017] release only per-word means and SDs, so we estimate the ceiling from those aggregates using the standard ICC+Spearman–Brown chain: ICC_1 recovers single-rater reliability

from the ratio of between-word to total variance, and Spearman–Brown projects single-rater reliability up to the N-vs-N scale via $r_{\text{full}} = 2r_{\text{split}}/(1 + r_{\text{split}})$. We always report r_{full}, because the human norm against which models are correlated is a full-panel mean rather than a single rater.

F.1 Experiment 2: empirical split-half

We draw 1000 random half-splits of the participants per concept (median $N{\approx}30$ raters across 354 concepts), compute per-concept property-category percentages on each half, and average Pearson r across splits. Spearman–Brown then projects the half-panel correlation to $r_{\text{full}} = 0.924$ on the 235-word abstract subset used in the main text and $r_{\text{full}} = 0.920$ on the 119-word concrete subset used in the appendix.

F.2 Validation of the ceiling estimation

Experiment 2 is the only dataset where both methods are available on the same data, since the participant-level release also yields per-word means and SDs. Applying the ICC+SB chain we use for the other experiments to the Experiment 2 category-frequency matrix gives a within-dataset cross-check on the estimate.

Table 7: ICC+Spearman–Brown analytical estimate vs. 1000-split empirical ceiling on Experiment 2, both at r_{full} scale.

Slice	ICC+SB	Empirical	$\lvert\Delta\rvert$
Abstract (235 words)	0.941	0.924	0.017
Concrete (119 words)	0.929	0.920	0.009

The two methods agree to within 0.02 on both slices, with the analytical estimate sitting fractionally above the empirical ceiling. We therefore consider the ICC+SB ceilings reported below as accurate to within $\sim$0.02.

Experiment 1 ceilings. Harpaintner et al. [2018] publish per-word frequencies in four aggregate categories. The resulting r_{full} ceilings are:

Sensorimotor: 0.966, Internal: 0.964, Social: 0.971, Verbal: 0.992 ($\overline{r}_{\text{full}} = 0.974$).

Rating-experiment ceilings. Troche et al. [2017] publish per-word means and SDs on each of the 14 grounding dimensions. We apply ICC+SB dimension-by-dimension with $\sigma^2_{\text{between}}$ from the cross-stimulus variance of per-word means, σ^2_{within} from the mean squared SD, and N from the median raters per word ($\approx$ 24). The 14 r_{full} ceilings range from 0.75 on Space to 0.96 on Visual Form (cross-dimension mean 0.87).

G Prompting setup

This section presents verbatim the prompts used on the three experiments. The literal token `{word}` is substituted by each target concept on each call, and `{properties}` (in the coding prompts) by the comma-separated list of generated properties for that concept.

G.1 Property Generation Prompt (Experiment 1)

The generation prompt for Experiment 1 is shown below in Figure 10. It asks the model for exactly four spontaneous properties per noun, with one in-context example (*sympathy*). It follows the original property listing instructions of Harpaintner et al. [2018].

G.2 Property Generation Prompt (Experiment 2)

The generation prompt for Experiment 2 is shown below in Figure 11. It asks the model for exactly five spontaneous clues per noun, with two in-context examples (*dog* and *dejected*). It follows the human rater instructions in Appendix A of Kelly et al. [2024].

```
You will be performing a word generation task. For the given word '{word}', generate four
unique associations/situations/properties (referred to as 'properties') that come to mind as
spontaneously as possible. Avoid using synonyms. Your response should include only the word
and the four properties, formatted like this example:
word: sympathy
properties: hug, friends, joy, sun
```

Figure 10: Property generation prompt for Experiment 1 Harpaintner et al. [2018], reproduced verbatim. The placeholder `{word}` is replaced by each of the target nouns at inference time.

```
Your job is to define the word "{word}" so that another person, who will only be able to see
your definitions, is able to guess the word you were defining. You can think of this task as
a game in which you need to help your partner understand the meaning of a word that you know
but they don't by giving them clues.

List each clue on a separate numbered line. Clues can be single words or short phrases, but
each clue should contain as few words as possible. You should describe properties of the
idea the word refers to, including its physical features (such as how it looks, feels, or
sounds), where or when you might encounter or experience it, where or when it is used or
experienced, how it is used or experienced, or how it behaves. For example, for the two
words below you might give the following clues:

DOG: Pet, Animal, Has fur, Barks, 4 legs
DEJECTED: Slumped, Shoulders, After you've failed, Feel hopeless, No enthusiasm

A good response will, when all of the clues are combined, define or describe the idea the
word refers to as completely as possible. Keep in mind that your partner will need to guess
the exact word you defined, so try to give clues that will help them separate that word from
other, similar words. Although a word might be used multiple ways, list properties of the
most common usage (the one that you think of first when you read the word).

Remember that you should describe what the word means, and not the word itself — clues such
as "starts with a 't'" or "is a three-letter word" will not be used. Do not use similar
words, or synonyms, as clues. Also refrain from listing whatever comes to mind when you read
the word (e.g., "cat" for the word dog), as these associated things do not describe the
meaning of the word itself.

You must provide exactly 5 clues.

Word: {word}.
```

Figure 11: Property generation prompt used for Experiment 2 Kelly et al. [2024], reproduced verbatim. The placeholder `{word}` is replaced by each target concept at inference time.

G.3 Coding Prompt (Experiment 1)

Each generated property is assigned one of five mutually exclusive categories: **Sensorimotor feature** (a feature experienced by the senses); **Social constellation** (the coexistence or interaction of persons); **Internal state and emotion** (internal cognitive processes such as motivation, emotion, volition); **Association** (thematically or symbolically related but not descriptive); and **Other abstract concept** (an abstract feature that describes the concept). After coding, the two abstract categories: **Association** and **Other abstract concept** are merged to form the final abstract category **Verbal association**. The prompt provides a single in-context example (the *sympathy* properties from the generation prompt, one per category) and asks the coder to emit one labelled line per property. The full coding prompt is shown in Figure 12.

```
You are given an abstract word and four properties that refer to this word. You have to
classify each property into one of the following five categories.

Sensorimotor feature: A feature that can be experienced by our senses. It describes the
meaning of the abstract concept, or the abstract concept can be applied to this feature.
Social constellation: a feature or a situation that describes the coexistence of different
persons or which implies an interaction between at least two different persons.
Internal state and emotion: a feature or a situation that reflects internal, cognitive
processes (e.g., motivation, emotion, volition).
Association: a feature or a situation that does not describe the abstract concept, but which
is thematically or symbolically related to it.
Other abstract concept: an abstract feature that describes the abstract concept or to which
the abstract concept can be applied.

You must output each property and the category, formatted as in the following example
(Abstract word: sympathy, Properties: hug, friends, joy, sun).
hug: Sensorimotor feature
friends: Social constellation
joy: Internal state and emotion
sun: Association

Abstract word: {word}
Properties: {properties}
```

Figure 12: Coding prompt for Experiment 1. Categories are listed verbatim from Harpaintner et al. [2018]; `{word}` and `{properties}` are replaced at inference time by the concept and the comma-separated generated property list.

G.4 Rating Experiment Prompt

The rating experiment Troche et al. [2017] uses 14 cognitive dimensions: *color, taste/smell, tactile, visual form, auditory, emotion, polarity, social, morality, thought, self-motion, space, quantity, time*. Every dimension uses the same instruction skeleton, with only the dimension specific statement substituted in. The full prompt for the *emotion* dimension is shown verbatim in Figure 13; the other thirteen dimensions are produced by replacing the sentence "I relate this word with human emotion." with the analogous sentence for the target dimension, taken verbatim from Troche et al. [2017]. The output is a list of `('word', INT_SCORE)` tuples on a 1–7 Likert scale.

```
In this study we are selecting linguistic stimuli and we need evaluations on a series of
words. For each word, rate your level of agreement with the following statement:

"I relate this word with human emotion."

The scale ranges from 1 to 7, as follows:
Strongly Disagree <- 1      2      3      4      5      6      7 -> Strongly Agree

Keep in mind that the value of "1" corresponds to Strongly Disagree and value of "7"
to Strongly Agree. Please feel free to use the whole range of values provided. When
making your ratings, try to be as accurate as possible, but do not spend too much time
on any one word.

In each line of the output write only the word and its rating in a tuple like this:
('word', INT_SCORE).
```

Figure 13: Rating prompt for the Troche benchmark, shown for the *emotion* dimension. The same instruction skeleton is reused for all 14 dimensions; only the target sentence ("I relate this word with *X*.") changes between dimensions, with the wording for each dimension taken verbatim from Troche et al. [2017].

H Extended mechanistic analysis results

H.1 Feature identification dataset construction

This appendix documents how we built the 425-noun balanced corpus used as the candidate pool for SAE feature identification (Section 5.1). Published norm sets cover individual dimensions of abstract-concept structure but no single resource gives us four balanced experiential categories on the same vocabulary, as well as Troche et al. [2017] does. But we did not want to use our test-set to discover features, so we had to look at other norms. Additionally, we wanted to use only the very highest words per category, and this meant that we needed a larger dataset to subsample only the highest ones. We therefore built our own noun corpus by combining four complementary datasets. Each category is anchored to one primary norm and a single thresholded metric as shown in Table 8 below, so that only the highest-scoring words per category are selected. A word that is not a noun or that satisfies thresholds for more than one grounded category is dropped from both. The final corpus comprises 110 sensorimotor, 101 internal state, 102 social and 112 abstract nouns, for a total of 425 nouns.

Table 8: Norm sources and the metric extracted from each.

Norm	Source	Metric used
Lancaster Sensorimotor	Lynott et al. [2020]	Max across 11 modalities
Diveica Socialness	Diveica et al. [2023]	Socialness 1–7
Glasgow Norms	Scott et al. [2019]	Arousal + |Valence−5|
Brysbaert Concreteness	Brysbaert et al. [2014]	Concreteness 1–5

Table 9: Feature-identification corpus composition. Sensorimotor is further partitioned into 11 Lancaster sub-modalities at 10 words each (Visual, Auditory, Olfactory, Gustatory, Haptic, Interoception, Mouth, Head, Hand_arm, Foot_leg, Torso), enabling per-submodality stratification at the splitting stage.

Category	N	Primary source
Sensorimotor	110	Lancaster Norms (11 submodalities × 10)
Social	102	Diveica Socialness
Internal State	101	Glasgow
Abstract Baseline	112	Brysbaert (low concreteness)

Sentence generation Each of the 425 nouns is paired with 10 context sentences that elicit the noun as the next token. Sentences are produced with Gemini-3.1-pro-preview at temperature 0. For example, for w=*surprise* one of the generated sentences is

> *"If you are feeling sad because you think everyone forgot your birthday, but then they jump out of the dark to make you smile, they have planned a ___".*

The blank is the position where the target noun was originally written.

The prompt instructs Gemini to write 10 numbered sentences ending with the target word, framed as a one-line riddle that makes the missing word easy to recover, and themed by the noun's category:

> *Write* 10 *numbered sentences about everyday life that end with the word '<WORD>'. Your goal in writing these sentences is to make sure that it is obvious that the sentence ends with this word, like a very simple and easy riddle to learn the word. You must also make sure that the sentence is aligned with the theme of '<THEME>'. Start your answer with '*1. *<the first sentence>'.*

The string for <THEME> is filled per subcategory (for Sensorimotor: Visual → "visual perception", Auditory → "auditory perception", etc.) or per category (Internal, Social). For Abstract nouns the theme clause is dropped: there is no grounded modality to anchor on, so we let Gemini choose any framing.

A regex post-processor then strips the target noun from the end of each generated line and verifies it is the same as the intended target. The final $4{,}250$ successful rows generated for the 425 training nouns are released publicly.

Last-word-target probing: per-noun activation signatures Once the dataset is built, for each sentence in $S(w)$, we remove the target noun w and feed the resulting prefix to the model. The SAE activation is read at the last token of the prefix, i.e. the position where the model is about to emit w given everything before it. Per (layer, feature) we take the median across the 10 sentences. Any feature that has a non-zero median for five or more nouns from a single category is considered as a candidate for that category.

Why this construction. The setup ensures (i) The SAE sees a natural mid-sentence residual and therefore stays inside its training distribution, unlike isolated-token probes where many SAE features go silent. (ii) The residual at position $n-1$ is a read-out of the pre-next-token distribution conditioned on the preceding context, which is target-word-specific without requiring the model to actually emit the target. (iii) The aggregation is median over 10 diverse context sentences, which suppresses any single sentence's effect.

Identified features per layer. In Figure 14 we see the features that have at least one non-zero median noun activation in our dataset and the ones that have at least five activations from a single category: the candidates. The features that have only a single activation are of course substantially more: either they fired randomly, or they are very specific to that noun. We observe an interesting phenomenon: until layers 15-17 the model seems to be encoding mostly and understanding the context of the previous and current tokens, so it does not have many features that correspond to the next token content which is our interest. From layer 17 (where it has about 1500 features) and after we see that the number of specific features increases rapidly and at layer 24 it reaches 3500 features and stays constant, indicating that the model starts making decisions about the next word. On the other hand, the candidate features we selected exhibit a different behavior: they increase slightly but noticeably after layer 15 and then they remain constant, until a sudden step increase in the last layer. We hypothesize that this is because these features are mostly responsible for category-level decisions that have to do with the theme and context of the prefix, as intended by their selection process, and not so much for specific decisions dictated by linguistic or reasoning rules.

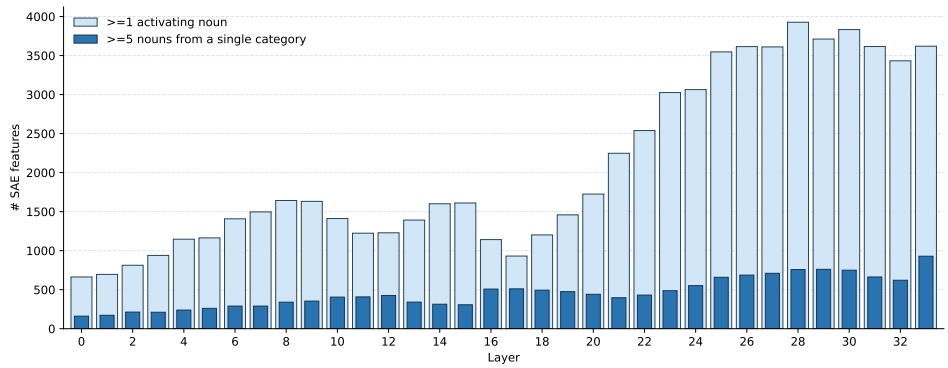


Figure 14: Number of SAE features per layer identified with our detection algorithm in Gemma-3-4B.

H.2 Feature validation

Feature interpretation. Determining the precise functionality of the features highly correlated with grounding dimensions remains challenging, as most lack an obvious semantic meaning. To facilitate interpretation and visualization, we provide links to Neuronpedia for representative features within each category. Additionally, we report automatic interpretations derived from our dataset activations using Claude-Opus-4.6. Note that the highest-correlated features discovered in our analysis are often not hosted on Neuronpedia, as the platform currently only supports a subset of layers (i.e., layers 9, 16, 22, and 29). Consequently, we present the best-performing features currently available.

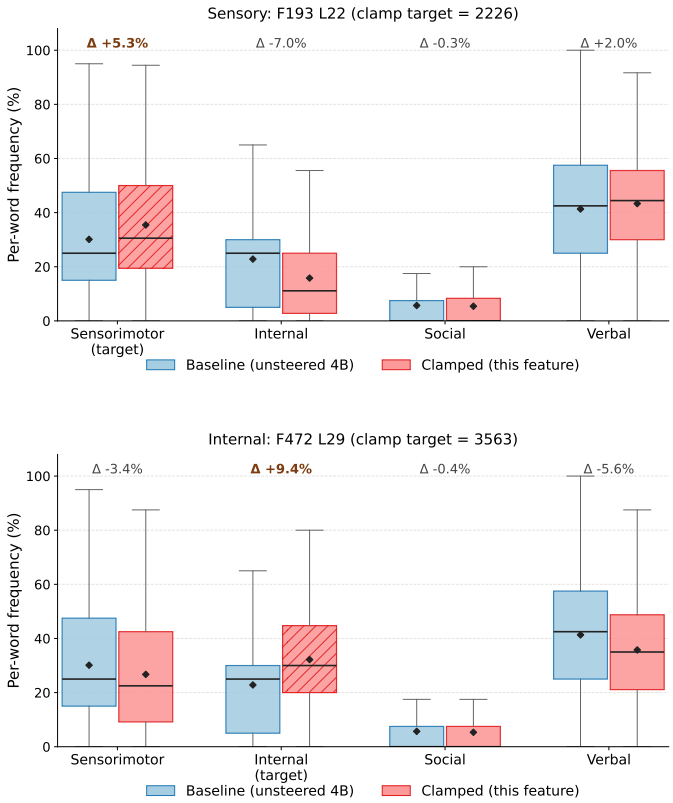


Figure 15: Boxplot for steered model with sensory (top) and internal (bottom) features.

Importantly, while Neuronpedia typically displays the highest positive and negative logits for each feature, our analysis focuses on the highest positive *next-token* logits (i.e., the logits predicted for the token immediately following the active feature). This distinction stems from our dataset's construction and aligns more accurately with our primary task of property-generation.

Representative features from each category available on Neuronpedia:

- **Sensory** ($r = 0.575$)**:** Feature 193, Layer 22.[3]
  Auto-interpretation: Edible foods and gustatory/oral content.
- **Motion-Self** ($r = 0.426$)**:** Feature 132, Layer 29.[4]
  Auto-interpretation: Forceful or momentous events and qualities.
- **Internal** ($r = 0.647$)**:** Feature 472, Layer 29.[5]
  Auto-interpretation: Emotional and motivational states across positive and negative affect.
- **Social** ($r = 0.375$)**:** Feature 166, Layer 22.[6]
  Auto-interpretation: Family and kinship relations.
- **Time** ($r = 0.373$)**:** Feature 733, Layer 29.[7]
  Auto-interpretation: Value-laden outcomes, advantages and setbacks.

Feature steering. For steering, we use clamping [Templeton et al., 2024] and we set the activation to double the median of the highest-activating noun from our identification dataset (Section 5.1). While this heuristic ensures a clear signal, it yields inconsistent intervention strengths. Specifically

[3] https://www.neuronpedia.org/gemma-3-4b-it/22-gemmascope-2-res-16k/193
[4] https://www.neuronpedia.org/gemma-3-4b-it/29-gemmascope-2-res-16k/132
[5] https://www.neuronpedia.org/gemma-3-4b-it/29-gemmascope-2-res-16k/472
[6] https://www.neuronpedia.org/gemma-3-4b-it/22-gemmascope-2-res-16k/166
[7] https://www.neuronpedia.org/gemma-3-4b-it/29-gemmascope-2-res-16k/733

for the internal-feature, we found it had unusually high max-activation and we kept it at 1x, because otherwise the model did not follow instructions properly.

In Figure 15 we present the boxplots for the Sensory and Internal features from above, compared against the baseline model. Both features lead to substantial shifts of the distribution of generated properties towards their target category, while the other categories decline or remain almost constant.

I Controlled comparison of rating and generation

This appendix details the controls summarized in Table 10. All values are means over the 21 models. Ceilings are human split-half reliabilities, projected to the full panel with Spearman-Brown and recomputed on the exact word set and dimensions of each row.

Table 10: Rating vs. generation under matched conditions, averaged over the 21 models. Ceiling: human split-half reliability on the same words and dimensions.

Task	Condition	Mean r	Ceiling	r / ceiling
Rating	all 751 words, 14 dimensions	0.631	0.871	72.5%
Rating	378 most abstract words, 14 dimensions	0.543	0.751	72.4%
Rating	all 751 words, 3 factors	0.636	0.917	69.4%
Rating	93 shared words, 3 shared categories	0.578	0.738	78.4%
Generation	93 shared words, 3 shared categories	0.308	0.966	31.8%

Abstract words only. The rating task of Troche et al. [2017] spans a concreteness continuum, while Experiment 1 uses abstract words only. We therefore restricted the rating task to the 50% most abstract words (A50, 378 words, by the concreteness norms released with the dataset). As a control for the smaller sample, we repeated the analysis on 20 random draws of 376 words.

Split	Mean r (14 dimensions)	Ceiling	r / ceiling
Full (751 words)	0.631	0.871	72.5%
Random 50% (control, 20 draws)	0.630	0.870	72.5%
A50 (378 most abstract)	0.543	0.751	72.4%

The decrease on abstract words comes from the restricted range of the abstract subset, not from the smaller sample: the human ceiling drops in the same proportion, and the fraction of attainable reliability that models reach is unchanged.

Three-factor reduction. Troche et al. [2017] reduce the 14 dimensions to three factors: Endogenous (affective and internal), Exogenous (perceptual and sensory), and Magnitude. Projecting human and model ratings onto these factors and correlating them per factor gives a mean r of 0.636, compared with 0.631 for the 14 per-dimension correlations. On the A50 subset, the three-factor correlation drops only from 0.636 to 0.614, compared with 0.631 to 0.543 in the 14-dimension analysis; the larger drop in the latter is concentrated in individual sensory dimensions.

Matched words and categories. No rating dimension corresponds to the *Verbal Association* category of Experiment 1, and the Magnitude dimensions (Space, Quantity, Time) have no counterpart among its categories. We therefore compared the two tasks on the three categories they share: Sensorimotor (Color, Taste/Smell, Tactile, Visual Form, Auditory, Self-Motion), Internal (Emotion, Polarity, Morality, Thought), and Social (Social), restricted to the 93 words present in both stimulus sets.

	Mean r (3 categories)	Ceiling	r / ceiling
Rating (Troche)	0.578	0.738	78.4%
Generation (Harpaintner)	0.308	0.966	31.8%
Difference	$+0.271$	-0.228	$+46.6$ pp

J Alternative alignment metrics

We tested whether the contrast between the two tasks depends on Pearson correlation by adding Spearman correlation and representational similarity analysis (RSA). For RSA, we build a concept-by-concept dissimilarity matrix from the human category vectors and another from each model's vectors, and correlate the two matrices. RSA therefore asks whether humans and models organize concepts in the same way, rather than whether they assign the same frequency to each category.

Matched setup: 93 shared words, 3 shared categories			
Metric	Rating	Generation	Gap
Pearson	0.578	0.308	+0.271
Spearman	0.506	0.321	+0.185
RSA	0.491	0.104	+0.387
RSA, Claude Opus 4.6	0.538	0.132	+0.406
RSA, GPT-5.4	0.502	0.169	+0.333
RSA, Gemini 3.1 Pro	0.529	0.106	+0.423

The contrast also holds when each task is scored on its own complete word set. Values are means over the 21 models with 95% bootstrap confidence intervals over words (1000 draws).

Task	Metric	Words	Mean	95% CI
Rating	Pearson	751	0.631	[0.620, 0.641]
Rating	Spearman	751	0.591	[0.582, 0.602]
Rating	RSA	751	0.675	[0.651, 0.698]
Generation	Pearson	293	0.281	[0.236, 0.324]
Generation	Spearman	293	0.279	[0.234, 0.321]
Generation	RSA	293	0.082	[0.065, 0.123]

The confidence intervals of the two tasks are disjoint under every metric. The low RSA in generation formalizes the observation of Section 3.2: models can produce roughly human-like aggregate frequencies while assigning properties to different concepts than humans do.

K Property budget

The four-property budget of Experiment 1 might lead models to spend their responses on easily accessible verbal associations, pushing internal-state content past the cutoff. We repeated Experiment 1 with the three frontier models, requesting 4, 8, or 12 properties per concept (10 runs each). For 8 and 12 properties we score both the full list (FULL) and a random subset of four properties (RAND4). SM, IS, SOC, VA denote Sensorimotor, Internal State & Emotion, Social, and Verbal Association; humans produce 32.5% IS and 26.0% VA.

Model	Condition	Mean r	SM%	IS%	SOC%	VA%	SM r	IS r	SOC r	VA r
GPT-5.4	4 (baseline)	0.303	31.1	13.5	10.5	44.7	0.33	0.26	0.41	0.21
	8 FULL	0.315	35.4	14.1	11.6	38.9	0.34	0.32	0.40	0.19
	8 RAND4	0.321	35.4	13.9	11.7	39.0	0.34	0.35	0.38	0.21
	12 FULL	0.320	38.3	13.5	13.2	34.9	0.35	0.34	0.41	0.18
	12 RAND4	0.301	38.6	13.6	13.4	34.4	0.30	0.32	0.39	0.19
Claude Opus 4.6	4 (baseline)	0.266	30.4	13.5	7.8	48.1	0.21	0.20	0.37	0.28
	8 FULL	0.337	30.9	16.0	10.1	43.0	0.31	0.36	0.42	0.25
	8 RAND4	0.281	30.5	16.0	10.0	43.5	0.25	0.30	0.38	0.20
	12 FULL	0.352	33.3	16.5	11.1	39.1	0.34	0.40	0.42	0.24
	12 RAND4	0.331	33.2	16.6	11.1	39.1	0.33	0.37	0.39	0.24
Gemini 3.1 Pro	4 (baseline)	0.259	35.2	7.6	9.1	47.9	0.31	0.16	0.40	0.17
	8 FULL	0.236	41.0	7.5	10.7	40.8	0.34	0.12	0.40	0.09
	8 RAND4	0.224	40.7	7.9	10.8	40.6	0.32	0.12	0.35	0.10
	12 FULL	0.268	43.5	8.0	12.5	36.0	0.34	0.19	0.42	0.12
	12 RAND4	0.208	43.1	8.0	13.0	35.8	0.29	0.14	0.34	0.06

A larger budget slightly lowers the share of verbal associations and raises sensorimotor and social content, but internal-state content stays at about half the human rate or less, and Mean r stays far below the human ceiling.

L Prompt framing

We compared the original neutral instruction with two variants, each inserting one clause at the same position of the original prompt (Figure 10), so that the task description, output format, and worked example are identical across conditions.

- *First-person experiential:* "As you do this, recall specific moments when you personally experienced or encountered '{word}', and let the properties come from those memories."
- *Free-associate as a person:* "Do not aim to be complete or accurate; simply write the first things an ordinary person would associate with the word."

Framing	GPT-5.4	Claude Opus 4.6	Gemini 3.1 Pro
Neutral (original)	0.303	0.266	0.259
First-person experiential	0.242	0.212	0.159
Free-associate as a person	0.306	0.343	0.302

The experiential framing lowers alignment for all three models. The free-association framing yields modest gains for Claude Opus 4.6 and Gemini 3.1 Pro and none for GPT-5.4. The best value under either framing, 0.343, remains far below the human ceiling, so the gap is not explained by the neutral wording of the original prompt or by models treating the task as a request for comprehensive or correct properties.

M Base versus instruction-tuned models

Base models cannot follow the chat-style instruction of Experiment 1; with the original one-shot prompt in completion format, more than half of their outputs could not be parsed as four properties. We therefore used a three-shot completion prompt: the original instruction verbatim, followed by

```
word: sympathy
properties: hug, friends, joy, sun

word: selection
properties: wine, choice, variety, quality

word: stability
properties: balance, rock, foundation, calm

word: {word}
properties:
```

The first example is the original one; the other two words were drawn at random from the stimulus set. The same prompt was given to the base and the instruction-tuned checkpoints. Gemma 3 12B and 27B base checkpoints did not produce reliably parsable output even under this prompt, so we report Gemma 3 4B and Qwen3 8B. Pretrained-only checkpoints are also unavailable for the frontier models we study.

Model	Format	Examples	Mean r	Internal state (%)	Verbal (%)
Humans (norms)	n/a	n/a	0.974[†]	32.5	26.0
Gemma-3-4B-it	chat	1	0.270	22.7	41.4
Gemma-3-4B-it	chat	3	0.309	20.9	43.1
Gemma-3-4B-it	completion	3	0.369	17.1	41.9
Gemma-3-4B-pt (base)	completion	3	0.411	23.9	42.1
Qwen3-8B-it	chat	1	0.373	15.9	39.6
Qwen3-8B-it	chat	3	0.286	17.7	45.8
Qwen3-8B-it	completion	3	0.359	17.6	44.0
Qwen3-8B-Base	completion	3	0.453	22.5	48.9

[†]Human ceiling.

Under the matched completion format, both base checkpoints are more aligned than their instruction-tuned counterparts and produce more internal-state content, moving toward the human rate of 32.5%. Verbal associations, however, remain strongly over-produced, and are highest for the Qwen base model. This comparison is not a clean causal ablation: when the two additional examples were chosen to emphasize one grounding category, base-model generations shifted toward that category by about 6 to 8 percentage points, compared with about 3 to 4 points for the instruction-tuned models, so the demonstrations needed to make base models follow the task can themselves prime the category distribution. Instruction tuning may therefore influence the grounding profile, but it does not account for the gap.

N Statistical units and pooled-model profiles

In Harpaintner et al. [2018], each word received four properties from each of 10 participants; our model profiles likewise aggregate 10 independently sampled response sets of four properties per word. The number of response sets per word is therefore matched; what differs is that human response sets come from distinct people while model response sets are samples from one model. The diagonal of Figure 3a shows that repeated samples from one model form a profile about as stable as the human aggregate.

To test a more heterogeneous notion of the generating "individual", we built pooled profiles in which each word's response sets are drawn from several models rather than one. Pooling all 21 models at 240 properties per word gives Mean $r = 0.391$ (40.2% of the ceiling). Even pooled profiles remain far below the human ceiling. This analysis asks a different question from our main comparison, whether heterogeneity across models produces a more human-like profile, and does not replace the model-specific results.

O Multimodal models

A clean test of multimodality is not possible with current frontier models, which are natively multimodal without text-only counterparts. The closest pairs in our panel are Qwen3 vs. Qwen3-VL at 4B and 8B. These pairs may also differ in training data, objectives, and post-training, so the comparison is suggestive rather than causal.

Pair (VL − text)	Δ Mean r	Δ Sensorimotor r
Qwen3-4B vs. Qwen3-VL-4B	-0.019	$+0.009$
Qwen3-8B vs. Qwen3-VL-8B	-0.100	-0.038

In the rating experiment, the text-only model is better in both matched pairs (Qwen3-8B 0.703 vs. VL 0.517; Qwen3-4B 0.507 vs. VL 0.457). Multimodal capability by itself therefore does not narrow the grounding gap in our data.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction accurately summarize the paper's claims, experiments, and scope.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses key limitations under section 6, including replication experiments, LLM-coder bias, and SAE coverage.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: The paper does not present formal theoretical results or proofs.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper provides datasets, prompts, model details, coding procedures, metrics, and uncertainty methods needed to reproduce the main results.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Code, prompts, model outputs, and analysis scripts are publicly available at `https://github.com/odychlapanis/grounding-gap/`, with per-experiment instructions for reproducing the main results.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: The experimental setup specifies stimuli, prompts, models, runs, coding, temperatures, and evaluation metrics.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: The paper reports bootstrap uncertainty, human ceilings, coder-error margins, and convergence analyses.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper identifies the compute setup, including API use for closed models and 8 x Quadro RTX 6000 runs for open models.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work uses existing datasets and model outputs, preserves anonymity, and presents no apparent ethics-policy violation.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Although the paper does not have a significant societal impact, it discusses why grounding differences matter as LLMs are used in more human-facing and agentic roles.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release a high-risk model, scraped dataset, or artifact requiring misuse safeguards.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Existing datasets, models, SAE resources, and tools are credited through citations.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.

- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not introduce a standalone new dataset, model, or benchmark as its main contribution.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not conduct new crowdsourcing or human-subject experiments.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not conduct new crowdsourcing or human-subject experiments.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: LLMs are a core component of the method: they code generated properties into grounding categories (Section 3.1 and Appendix A), generate the cloze sentences for SAE feature identification, and produce feature auto-interpretations (Appendix H). Each use is described and validated in the paper. LLMs were also used for text editing.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.