

Do Large Language Models appeal to the Senses? Narrative Embodiment in Stories told by Llama and Qwen

Stefan Arnold and Rene Gröbner

Friedrich-Alexander-Universität Erlangen-Nürnberg
Lange Gasse 20, 90403 Nürnberg, Germany
(stefan.st.arnold, rene.edgar.gröbner)@fau.de

Abstract

Language Models (LMs) produce highly fluent narratives, yet cognitive probing suggests their semantic representations lack grounding in sensorimotor experiences. By projecting model-generated texts into a multidimensional sensorimotor vector space, we quantify narrative embodiment in storytelling. Specifically, we evaluate the impact of parameter scaling and conditional multimodality on perceptual grounding. We find that (1) model capacity monotonically widens the divergence from the human level of embodiment, demonstrating that scale alone cannot resolve narrative disembodiment, and (2) audio-language models demonstrate a distinct superiority over vision-language models, an effect we contextualize through the auditory scaffolding hypothesis.

Ultimately, we demonstrate that authentic sensorimotor grounding is not an emergent property of autoregressive decoding or conditional encoding, highlighting a critical limitation in current generative architectures.

1 Introduction

Storytelling functions as an essential cognitive architecture through which humans organize experiences, construct meaning, and transmit knowledge (Boyd, 2009). Since language models are increasingly deployed to generate narratives, it becomes critical to evaluate the depth, authenticity, and structure of the worlds they construct. Tian et al. (2024) recently analyzed discourse-level progression by mapping macro-level story arcs, meso-level turning points, and micro-level affective dimensions. Such evaluation revealed significant discrepancies between human and model storytelling, demonstrating that model-generated narratives are frequently homogeneously positive, lack essential tension, and fall short of human narrative diversity.

While both plot architecture and affective tracking capture the temporal progression of a narrative,

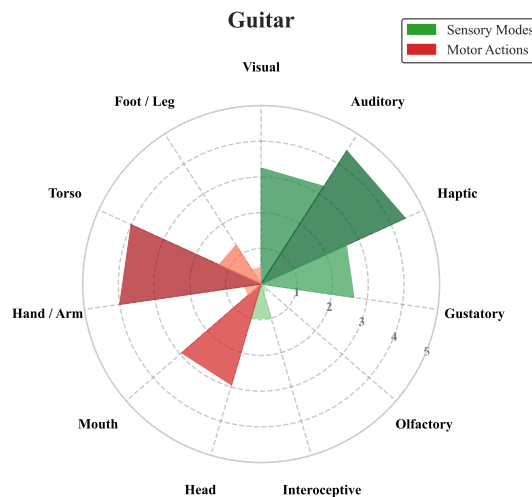


Figure 1: Multidimensional sensorimotor footprint of the lexical item 'guitar'. Ratings derived from Lancaster sensorimotor norms (Lynott et al., 2020) quantify the sensory and motor grounding of individual words.

they do not measure how deeply a story is anchored in human sensory experience. Because authentic meaning is deeply rooted in sensorimotor realities, a truly immersive narrative must also reflect this embodied dimension, and evaluating a model's ability to capture it remains critical to quantifying the gap between human and model storytelling.

Figure 1 illustrates that a single lexical item such as *guitar* evokes a multidimensional physical experience, simultaneously activating auditory representations of sound alongside the haptic dimensions of the hands and arms. Because language carries this inherent perceptual grounding, a truly immersive story cannot merely execute a sequence of plot turning points; rather, it must actively anchor the reader in a physical reality.

Xu et al. (2025) and more recently Hagihara and Miyazawa (2026) have both highlighted a severe *embodiment gap*, demonstrating that while ungrounded language models can recover abstract cognitive features like imageability, they systemati-

cally fail to capture sensorimotor ratings. Because these cognitive probes evaluated words in isolation rather than continuous corpora, it remains unknown how this disembodiment manifests in storytelling, where a model must dynamically select vocabulary to construct narrative depth.

Contribution. We investigate the sensorimotor grounding of model-generated storytelling by shifting the analytical focus of prior stylometric analysis (Reinhart et al., 2025) from lexicogrammatical markers to psycholinguistic norms. By projecting generated stories into a vector space defined by sensory modes and action effectors, we can directly quantify the narrative’s perceptual footprint.

Our sensorimotor analysis reveals that text-only language model suffer from a severe embodiment gap that cannot be bypassed by parameter scaling, demonstrating that scale alone cannot resolve narrative disembodiment. We further notice that conditional modality can help embodied writing. Counterintuitively, vision-language models show virtually no improvement over text-only models, whereas audio-language models emerge as significantly more physically grounded. This indicates that the sequential nature of acoustic waveforms aligns more naturally with the descriptive nuances of grounded text than static visual representations.

2 Related Work

From the perspective of language production, several recent studies reveal that human-written and model-generated text differs substantially despite similar grammatical fluency. Through a multidimensional corpus analyses, Reinhart et al. (2025) showed distinct grammatical and rhetorical styles, while diversity benchmarks report recurrent syntactic templates (Shaib et al., 2024) and reduced or differently structured linguistic variation, particularly across extended passages (Tripto et al., 2025) or in story generation (Guo et al., 2025).

From the perspective of narrative storytelling, stories written by humans and those generated by models differ in terms of narrative progression and affective dynamics (Tian et al., 2024), while psychological-depth measures captured divergence in their experiential richness (Harel-Canada et al., 2024). Furthermore, narratological analyses reveal differences in character variety (Brei et al., 2026) and expose distinct patterns in the social organization of stories (Nonaka and Perry, 2025).

We extend this line of research to grounded sto-

Prompt: The Mage, the Warrior, and the Priest

Story: A light breeze swept the ground, and carried with it still the distant scents of dust and time-worn stone. The Warrior led the way, heaving her mass of armour and muscle over the uneven terrain. She soon crested the last of the low embankments, which still bore the unmistakable fingerprints of haste and fear. She lifted herself up onto the top the rise, and looked out at the scene before her. [...]

Table 1: Example prompt followed by a human-written story from the WRITING PROMPTS dataset (Fan et al., 2018). Note that we only leverage the story segment.

rytelling. Xu et al. (2025) tested the alignment of psycholinguistic and sensorimotor ratings between human cognition and model estimates. The comparison revealed an uneven correspondence that notably declined when shifting from psycholinguistic features to specific sensory and motor dimensions, a finding further corroborated by Hagihara and Miyazawa (2026). We test whether the same sensorimotor pattern appears in long-form storytelling and whether scaling or multimodality shifts generated narratives toward human profiles.

3 Sensorimotor Analysis

To systematically assess the embodiment gap in generative storytelling, we designed an experimental setup that maps free-form narrative text into a structured sensorimotor vector space.

3.1 Corpus Collection

We conducted our evaluation using the WRITING PROMPTS corpus (Fan et al., 2018), a large-scale dataset consisting of human-written stories, averaging approximately 734 words in length. Table 1 exemplifies the creative narrative style of writing prompts. To ensure sufficient context for our prompt-continuation layout, we strictly filtered the corpus to include only samples with length ≥ 1000 , yielding a subset of 1,700. For each instance, models were conditioned on a 500-words context of the human-written text and tasked with autonomously generating the 500-word continuation.

3.2 Feature Extraction

Grounded cognition links conceptual meaning to perceptual experience, bodily states, and action effectors (Barsalou, 2008). To systematically quantify the physical groundedness of the narratives, we projected the textual stories into a vector space using the *Lancaster sensorimotor norms* (Lynott et al.,

2020). This set of norms operationalizes human cognition by assigning continuous scalar ratings to 39,707 lemmatized English concepts across six sensory modalities (*visual, auditory, haptic, gustatory, olfactory, and interoceptive*) and five motor effectors (*head, mouth/throat, hand/arm, torso, and foot/leg*). Crucially, because these norms capture the graded intensity of a sensorimotor experience, relying on raw frequency counts or simple usage ratios would flatten vital psychological nuance by treating highly evocative physical descriptions identically to weakly associated terms.

To capture the true experiential depth of a narrative, we designed a metric that evaluates sensorimotor intensity normalized by document length. We define the sensorimotor intensity of a narrative as its average load per word. Beyond preserving the continuous nature of the sensorimotor ratings, normalizing by length is essential for evaluation as accumulative sums would artificially inflate the embodiment scores of models due to their well-documented verbosity bias (Saito et al., 2023).

We lemmatized all words in a story and queried the vocabulary against the sensorimotor lexicon. For a given dimension d , the sensorimotor load L_d is formulated as:

$$L_d = \frac{\sum_{i=1}^{|T|} \mathbf{v}_d(w_i)}{|T|}$$

where $\mathbf{v}_d(w_i)$ represents the sensorimotor scalar for a mapped word w_i , and $|T|$ denotes the cardinality of the set of all non-punctuation words. By summing intensity scalars instead of binary occurrences, this equation accurately quantifies the average perceptual weight of the generated story while naturally controlling for sequence length.

Since generative models can occasionally collapse into pathological repetition loops during extended sampling (Holtzman et al., 2020), we strictly controlled for text degeneration. We applied an automated regular expression filter to truncate the text at the exact onset of any continuous repetition. This guarantees that our intensity metric reflects intentional lexical choices rather than algorithmic stuttering (and ensures the validity of the denominator $|T|$, which controls for sequence length).

To establish a baseline for the human intensity of sensorimotor embodiment, we applied this feature extraction pipeline using spaCy to the human-written stories of our corpus subset. Establishing the natural variance of embodied writing is critical as it defines the human centroid and variance

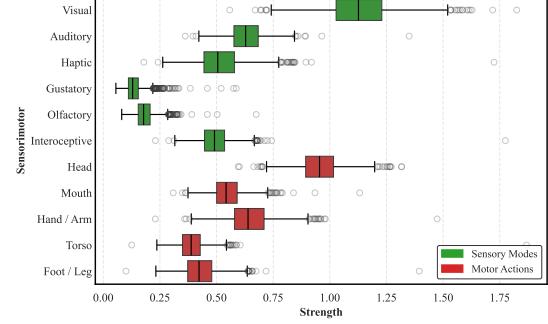


Figure 2: Distribution of the average sensorimotor load per story across six perceptual and five motor dimensions. We note that human-written stories are predominantly grounded in visual and auditory modalities.

against which model divergence is subsequently measured. Figure 2 shows that human narratives exhibit a non-uniform perceptual profile. Audiovisual perception and head-related actions dominate the embodiment space, whereas olfactory and gustatory dimensions are sparsely utilized, reflecting its highly descriptive and action-oriented nature.

3.3 Model Selection

To investigate the drivers of sensorimotor divergence, we evaluated two distinct architectural dimensions: *parameter scaling* and *conditional multimodality*. While parameter scaling is often presumed to resolve semantic deficits, we sought to isolate the effect of model capacity on embodied writing. We utilized the Llama-3 (Grattafiori et al., 2024) family across 1B, 3B, and 8B configurations to determine whether increased model capacity natively produces more embodied narratives. To ascertain whether encoders inject physical situatedness into generated stories, we compared the text-only Qwen-2 (Yang et al., 2025) against its vision-language (Bai et al., 2023; Wang et al., 2024) and audio-language (Chu et al., 2023) variants.

4 Embodiment Gap

4.1 Parameter Scaling

To characterize the embodiment shift induced by generative language models, we quantify the divergence between the human baseline and the model-generated stories using a custom distance metric inspired by *Burrows’s Delta* (Burrows, 2002). Instead of measuring shifts in discrete word frequencies, our metric operates directly on the continuous sensorimotor loads. We z -standardize our sensorimotor loads of the generated stories using the em-

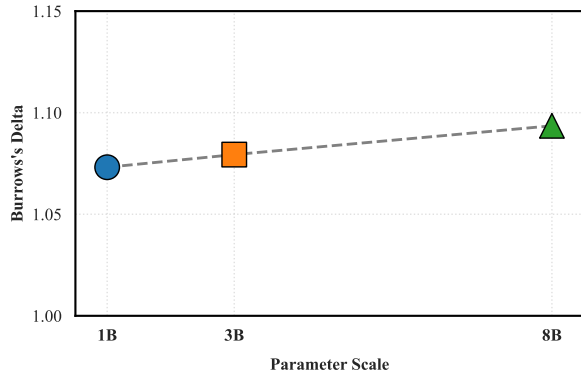


Figure 3: Embodiment divergence for Llama-3. Scale increases the embodiment gap, pulling text further from the human embodiment intensity.

pirical mean and standard deviation of the human baseline. The divergence is then computed as the mean absolute difference of these z -scores. This yields a single scalar that serves as a proxy for the physical fingerprint of a text. A lower score indicates close adherence to human sensorimotor properties, whereas a higher score reflects increased deviation and a widening of the embodiment gap.

Figure 3 presents the scaling trajectories for the Llama-3 family across parameter configurations of {1B, 3B, 8B}. While parameter scaling is widely recognized for improving syntactic fluency and abstract reasoning, our analysis reveals a stark barrier: larger capacities do not yield closer alignment with the human embodiment but instead scaling widens the divergence score monotonically.

To understand this embodiment gap, we must consider the nature of the language modeling objective. Because the raw representational capacity of the model expands, one might initially assume the generated narratives would naturally incorporate a richer sensory vocabulary. However, our findings indicate that text-only models are inherently biased towards disembodied language patterns. Scaling up provides exponentially more parameters to capture the human distribution, but rather than grounding the text perceptually, it reinforces the model’s reliance on non-sensorimotor words. This demonstrates that raw computational scale alone cannot bridge the gap between abstract symbol manipulation and grounded human experience.

4.2 Conditional Multimodality

Since text-only scaling fails to bridge the embodiment gap, we continue our investigation by examining whether multimodality can serve as a functional

proxy for physical experience. To assess whether cross-modal conditioning inject real-world physical grounding, we evaluated the relative embodiment shift of the text-language, vision-language, and audio-language models of the Qwen-2 family against the human embodiment.

Figure 4 visualizes a universal underrepresentation of embodied grounding. Although multimodal architectures employ encoders to align non-textual sensory with the language model’s representation space, each model produces sensorimotor vocabulary strictly below the human intensity. This highlights a persistent deficit in translating externally encoded modalities into grounded narrative prose.

Vision-Language. To isolate the influence of image sensory to perceptual writing, we compared the vision-language model to the text-only model. We find that the vision-language model performs virtually identically to its text-only counterpart in its core dimensions including *visual* and *haptic*.

Audio-Language. While visual conditioning provides negligible gains, audio conditioning demonstrates a distinct superiority in embodied grounding. We observe that the audio-language model consistently shifts notably closer to human embodiment intensities than the vision-language model across all sensorimotor dimensions. This advantage is naturally most pronounced in the *auditory* dimension, but it strikingly persists even within the *visual* and *haptic* dimensions. We argue that this unique cross-modal dominance occurs because the temporal features extracted by audio encoders align more naturally with the sequential generation required for cohesive storytelling than the spatial features extracted by vision encoders.

5 Conclusion

To assess the properties of model-generated stories beyond narrative structures (Tian et al., 2024), we investigated whether language models natively ground their narratives in human-validated physical experiences. By projecting long-form narrative text into a multidimensional sensorimotor vector space, we quantified the embodiment of distinct architectures across scales and conditioning against an average human style of writing.

We identify a severe *embodiment gap* in stories generated by language models that persists regardless of model capacity or cross-modal grounding. Evidenced by the systematic failure of models with

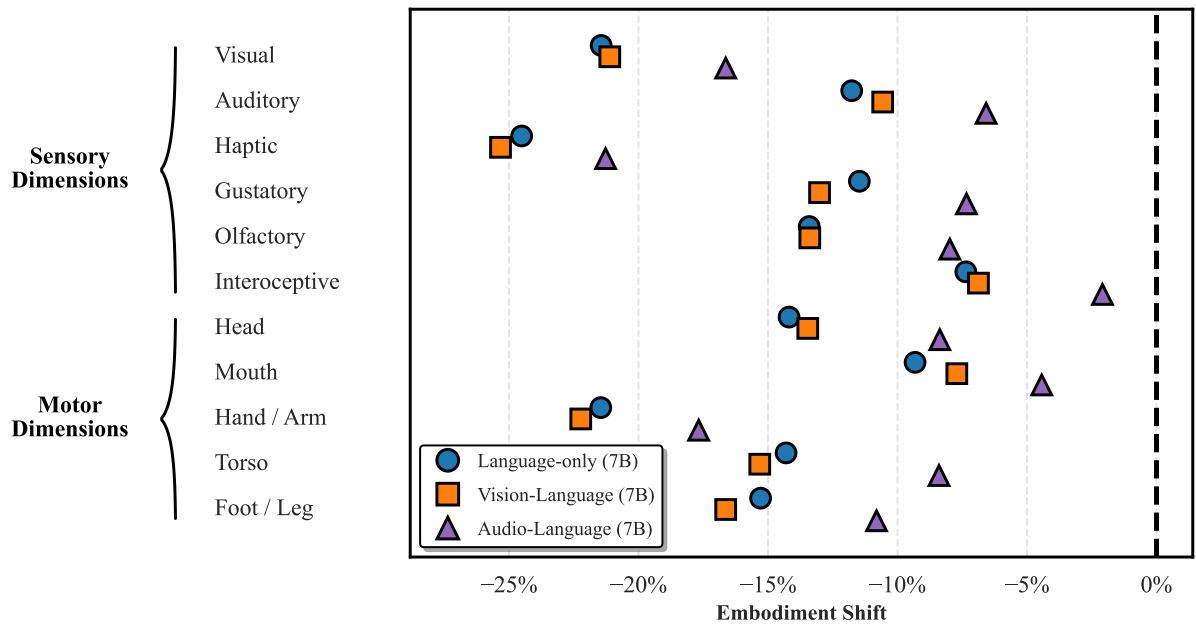


Figure 4: Relative embodiment gap of the Qwen text-only and cross-modal architectures compared to the human intensity of embodied writing. While all models systematically underrepresent sensory modes and motor actions, the audio-language model compared to the vision-language model steers vocabulary better to produce more perceptually-grounded stories across the entire sensorimotor subspace.

larger capacities to induce perceptually grounded vocabulary usage, parameter scaling cannot remedy this deficit of embodied writing. Although cross-modal sensory conditioning also fails to resolve the embodiment gap, audio-language conditioning distinctly and consistently outperforms vision-language conditioning.

We ground this cross-modal asymmetry in the *auditory scaffolding hypothesis* (Conway et al., 2009), which posits that sound serves as a primary cognitive scaffold for processing temporal information. Given that narrative construction is inherently a sequential process, the temporal features extracted by audio-frequency encoders provide a natively compatible structural foundation for physical grounding. In contrast, spatial features obtained by image-patch encoders fail to transfer physical situatedness into continuous prose. Ultimately, these findings demonstrate that sensorimotor grounding is not an emergent property and will require targeted training (Wu et al., 2026).

Limitations. We acknowledge that our investigation is restricted to conditional architectures, which utilize modality-specific encoders to project visual and audio features into the latent space of the language model. Unlike *unified* models where vision and audio sensory flow directly into the model back-

bone, these intermediary components create a bottleneck between the non-textual sensory and the language model’s internal representation. We plan to extend this line of research to unified architectures such as some of the Gemma-4 models. These unified models will allow us to determine whether early-stage sensory integration can natively resolve the narrative embodiment gap.

References

- Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*.
- Lawrence W. Barsalou. 2008. *Grounded cognition*. *Annual Review of Psychology*, 59:617–645.
- Brian Boyd. 2009. *On the origin of stories: Evolution, cognition, and fiction*. Harvard University Press.
- Anneliese Brei, Abhishek Sharma, Nicholas Sanaie, Lu Wang, and Snigdha Chaturvedi. 2026. *CASPER in the machine: Insights into character variety in LLM-generated stories*. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14815–14840, San Diego, California, United States. Association for Computational Linguistics.

- John Burrows. 2002. ‘delta’: a measure of stylistic difference and a guide to likely authorship. *Literary and linguistic computing*, 17(3):267–287.
- Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. 2023. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv preprint arXiv:2311.07919*.
- Christopher M Conway, David B Pisoni, and William G Kronenberger. 2009. The importance of sound for cognitive sequencing abilities: The auditory scaffolding hypothesis. *Current directions in psychological science*, 18(5):275–279.
- Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Yanzhu Guo, Guokan Shang, and Chlo   Clavel. 2025. [Benchmarking linguistic diversity of large language models](#). *Transactions of the Association for Computational Linguistics*, 13:1507–1526.
- Hiromichi Hagihara and Kazuki Miyazawa. 2026. How well do large language models mirror human cognition of word concepts?: A comparison of psychological ratings for early-acquired english words. *Behavior Research Methods*, 58(2):58.
- Fabrice Y. Harel-Canada, Hanyu Zhou, Sreya Muppalla, Zeynep Senahan Yildiz, Miryung Kim, Amit Sahai, and Nanyun Peng. 2024. [Measuring psychological depth in language models](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17162–17196, Miami, Florida, USA. Association for Computational Linguistics.
- Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.
- Dermot Lynott, Louise Connell, Marc Brysbaert, James Brand, and James Carney. 2020. The lancaster sensorimotor norms: multidimensional measures of perceptual and action strength for 40,000 english words. *Behavior research methods*, 52(3):1271–1291.
- Hiroshi Nonaka and K. E. Perry. 2025. [Evaluating LLM story generation through large-scale network analysis of social structures](#). *Preprint*, arXiv:2510.18932. Presented at the NeurIPS 2025 Workshop on Evaluating the Evolving LLM Lifecycle: Benchmarks, Emergent Abilities, and Scaling.
- Alex Reinhardt, Ben Markey, Michael Laudénbach, Kachad Pantusen, Ronald Yurko, Gordon Weinberg, and David West Brown. 2025. Do llms write like humans? variation in grammatical and rhetorical styles. *Proceedings of the National Academy of Sciences*, 122(8):e2422455122.
- Keita Saito, Akifumi Wachi, Koki Wataoka, and Youhei Akimoto. 2023. Verbosity bias in preference labeling by large language models. *arXiv preprint arXiv:2310.10076*.
- Chantal Shaib, Yanai Elazar, Junyi Jessy Li, and Byron C. Wallace. 2024. [Detection and measurement of syntactic templates in generated text](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6416–6431, Miami, Florida, USA. Association for Computational Linguistics.
- Yufei Tian, Tenghao Huang, Miri Liu, Derek Jiang, Alexander Spangher, Muhao Chen, Jonathan May, and Nanyun Peng. 2024. [Are large language models capable of generating human-level narratives?](#) In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17659–17681, Miami, Florida, USA. Association for Computational Linguistics.
- Nafis Irtiza Tripto, Saranya Venkatraman, Mahjabin Nahar, and Dongwon Lee. 2025. [Beyond checkmate: Exploring the creative choke points for AI generated texts](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 11953–11970, Suzhou, China. Association for Computational Linguistics.
- Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. 2024. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.
- Minghua Wu, Javier Conde, Pedro Reviriego, and Marc Brysbaert. 2026. How does fine-tuning improve sensorimotor representations in large language models? *arXiv preprint arXiv:2603.03313*.
- Qihui Xu, Yingying Peng, Samuel A Nastase, Martin Chodorow, Minghua Wu, and Ping Li. 2025. Large language models without grounding recover non-sensorimotor but not sensorimotor features of human concepts. *Nature human behaviour*, 9(9):1871–1886.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.