The Mechanistic Emergence of Symbol Grounding in Language Models

Ziqiao Ma[*1] **Shuyu Wu**[*1] **Xiaoxi Luo**[*2] **Yidong Huang**[1,4] **Josue Torres-Fonseca**[1]
Freda Shi[‡2,3] **Joyce Chai**[‡1]
[1]University of Michigan [2]University of Waterloo [3]Vector Institute [4]UNC at Chapel Hill

1 Introduction

Symbol grounding [29] refers to the problem of how abstract and discrete symbols, such as words, acquire meaning by connecting to perceptual or sensorimotor experiences. Extending to the context of multimodal machine learning, grounding has been leveraged as an explicit pre-training objective for vision-language models (VLMs), by explicitly connecting linguistic units to the world that gives language meanings [34, 39]. Through supervised fine-tuning with grounding signals, such as entity-phrase mappings, modern VLMs have achieved fine-grained understanding at both region [74, 48, 66] and pixel [79, 53, 78] levels.

With the rising of powerful autoregressive language models [LMs; 46, 1, 16, *inter alia*] and their VLM extensions, there is growing interest in identifying and interpreting their emergent capabilities. Recent work has shown preliminary correlational evidence that grounding may emerge in LMs [57, 61, 70] and VLMs [9, 8, 59] trained at scale, even when solely optimized with the simple next-token prediction objective. However, the potential underlying mechanisms that lead to such an emergence are not well understood. To address this limitation, our work seeks to understand the emergence of symbol grounding in LMs, causally and mechanistically tracing how symbol grounding arises within the internal computations.

We begin by constructing a minimal testbed, motivated by the annotations provided in the CHILDES corpora [41], where child–caregiver interactions provide cognitively plausible contexts for studying symbol grounding alongside verbal utterances. In our framework, each word is represented in two distinct forms: one token that appears in non-verbal scene descriptions (e.g., a *box* in the environment) and another that appears in spoken utterances (e.g., *box* in dialogue). We refer to these as environmental tokens (`<ENV>`) and linguistic tokens (`<LAN>`), respectively. A deliberately simple word-level tokenizer assigns separate vocabulary entries to each form, ensuring that they are treated as entirely different tokens by the language model. This framework enforces a structural separation between scenes and symbols, preventing correspondences from being reduced to trivial token identity. Under this setup, we can evaluate whether a model trained from scratch is able to predict the linguistic form from its environmental counterpart.

We quantify the level of grounding using surprisal: specifically, we compare how easily the model predicts a linguistic token (`<LAN>`) when its matching environmental token (`<ENV>`) is present versus when unrelated cues are given instead. A lower surprisal in the former condition indicates that the model has learned to align environmental grounds with linguistic forms. We find that LMs do learn to ground: the presence of environmental tokens consistently reduces surprisal for their linguistic counterparts, in a way that simple co-occurrence statistics cannot fully explain. To study the underlying mechanisms, we apply saliency analysis [65] and the tuned lens [3], which converge on the result that grounding relations are concentrated in the middle layers of the network. Further analysis of attention heads reveals patterns consistent with the aggregate mechanism [4], where attention heads support the prediction of linguistic forms by retrieving their environmental grounds in the context.

[*] Authors contributed equally to this work.
[‡] Advisors contributed equally to this work.

1st Workshop on Interpreting Cognition in Deep Learning Models (CogInterp @ NeurIPS 2025)

Table 1: Training and test examples across datasets with target word *book*. The training examples combine environmental tokens (<ENV>; shaded) with linguistic tokens (<LAN>). Test examples are constructed with either matched (*book*) or mismatched (*toy*) environmental contexts, paired with corresponding linguistic prompts. Note that in child-directed speech and caption-grounded dialogue, $book_{<ENV>}$ and $book_{<LAN>}$ are two distinct tokens received by LMs.

Dataset	Training Example		Test Example		
	<ENV>	**<LAN>**	**<ENV> Match**	**<ENV> Mismatch**	**<LAN>**
Child-Directed Speech	⟨*CHI*⟩ *takes book from mother*	⟨*CHI*⟩ *what's that* ⟨*MOT*⟩ *a book in it ...*	⟨*CHI*⟩ *asked for a new book*	⟨*CHI*⟩ *asked for a new toy*	⟨*CHI*⟩ *I love this* _____
Caption-Grounded Dialogue	*a dog appears to be reading a book with a full bookshelf behind*	⟨*Q*⟩ *can you tell what book it's reading* ⟨*A*⟩ *the marriage of true minds by stephen evans*	*this is a book*	*this is a toy*	⟨*Q*⟩ *can you name this object* ⟨*A*⟩ _____
Image-Grounded Dialogue		⟨*Q*⟩ *can you tell what book it's reading* ⟨*A*⟩ *the marriage of true minds by stephen evans*			*what do we have here?* _____

Finally, we demonstrate that these findings generalize beyond the minimal CHILDES data and Transformer models. They appear in a multimodal setting with the Visual Dialog dataset [19], and in state-space models (SSMs) such as Mamba-2 [17]. In contrast, we do not observe grounding in unidirectional LSTMs, consistently with their sequential state compression and lack of content-addressable retrieval. Taken together, our results show that symbol grounding can mechanistically emerge in autoregressive LMs, while also delineating the architectural conditions under which it can arise.

Figure 3 is an illustration of the symbol grounding mechanism through information aggregation.

2 Method

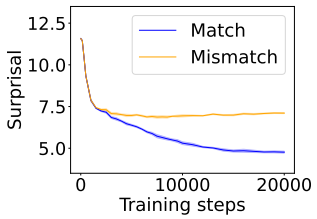
2.1 Dataset and Tokenization

To capture the emergent grounding from multimodal interactions, we design a minimal testbed with a custom word-level tokenizer, in which every lexical item is represented in two corresponding forms: one token that appears in non-verbal descriptions (e.g., a *book* in the scene description) and another that appears in utterances (e.g., *book* in speech). We refer to these by environmental (<ENV>) and linguistic tokens (<LAN>), respectively. For instance, $book_{<ENV>}$ and $book_{<LAN>}$ are treated as distinct tokens with separate integer indices; that is, the tokenization provides no explicit signal that these tokens are related, so any correspondence between them must be learned during training rather than inherited from their surface form. We instantiate this framework in three datasets, ranging from child-directed speech transcripts to image-based dialogue. A detailed explanation of these three different datasets is shown in the Appendix Section A.1

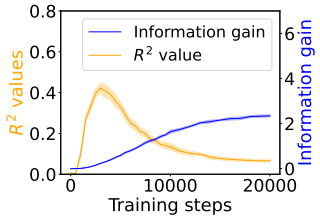
2.2 Evaluation Protocol

We assess symbol grounding with a contrastive test that asks whether a model assigns a higher probability to the correct linguistic token when the matching environmental token is in context, following the idea of priming in psychology. This evaluation applies uniformly across datasets (Table 1): in CHILDES and caption-grounded dialogue, environmental priming comes from descriptive contexts; in image-grounded dialogue, from ViT-derived visual tokens. We compare the following conditions:

- **Match (experimental condition)**: The context contains the corresponding <ENV> token for the target word, and the model is expected to predict its <LAN> counterpart.
- **Mismatch (control condition)**: The context is replaced with a different <ENV> token. The model remains tasked with predicting the same <LAN> token; however, in the absence of corresponding environmental cues, its performance is expected to be no better than chance.



(a) Surprisal



(b) Correlation

Figure 1: Results on 12-layer Transformer. Left: Average surprisal of the experimental and control conditions over training steps. Right: Grounding information gain and its correlation to the co-occurrence of linguistic and environment tokens over training steps.

For example (first row in Table 1), when evaluating the word $book_{\texttt{<LAN>}}$, the input context is

$$\langle CHI\rangle\ asked_{\texttt{<ENV>}}\ for_{\texttt{<ENV>}}\ a_{\texttt{<ENV>}}\ new_{\texttt{<ENV>}}\ book_{\texttt{<ENV>}}\ \langle CHI\rangle\ I_{\texttt{<LAN>}}\ love_{\texttt{<LAN>}}\ this_{\texttt{<LAN>}}\ \_\_\_\_\_, \tag{1}$$

where the model is expected to predict $book_{\texttt{<LAN>}}$ for the blank, and the role token $\langle CHI\rangle$ indicates the involved speaker or actor's role being a child. In the control (mismatch) condition, the environmental token $box_{\texttt{<ENV>}}$ is replaced by another valid noun such as $toy_{\texttt{<ENV>}}$.

Context templates. For a target word v with linguistic token $v_{\texttt{<LAN>}}$ and environmental token $v_{\texttt{<ENV>}}$, we denote $\overline{C}_v$ as a set of context templates of v. For example, when $v = book$, a $\overline{c} \in \overline{C}_v$ can be

$$\langle CHI\rangle\ asked_{\texttt{<ENV>}}\ for_{\texttt{<ENV>}}\ a_{\texttt{<ENV>}}\ new_{\texttt{<ENV>}}\ \texttt{[FILLER]}\ \langle CHI\rangle\ I_{\texttt{<LAN>}}\ love_{\texttt{<LAN>}}\_\_\_\_\_, \tag{2}$$

where `[FILLER]` is to be replaced with an environmental token, and the blank indicates the expected prediction as in Eq. (1). In the match condition, the context $\overline{c}(v)$ is constructed by replacing `[FILLER]` with $v_{\texttt{<ENV>}}$ in $\overline{c}$. In the mismatch condition, the context $\overline{c}(u)$ uses $u_{\texttt{<ENV>}} (u \neq v)$ as the filler, while the prediction target remains $v_{\texttt{<LAN>}}$.

For the choices of v and u, we construct the vocabulary V with 100 nouns from the MacArthur–Bates Communicative Development Inventories [24] that occur frequently in our corpus. Each word serves once as the target, with the remaining $M = 99$ used to construct mismatched conditions. For each word, we create $N = 10$ context templates, which contain both <ENV> and <LAN> tokens. Details of the vocabulary and context template construction can be found in the Appendix A.

Grounding information gain. Following prior work, we evaluate how well an LM learns a word using the mean surprisal over instances. The surprisal of a word w given a context c is defined as $s_{\boldsymbol{\theta}}(w \mid c) = -\log P_{\boldsymbol{\theta}}(w \mid c)$, where $P_{\boldsymbol{\theta}}(w \mid c)$ denotes the probability, under an LM parameterized by $\boldsymbol{\theta}$, that the next word is w conditioned on the context c.

The *grounding information gain* $G_{\boldsymbol{\theta}}(v)$ for v is defined as

$$G_{\boldsymbol{\theta}}(v) = \frac{1}{N}\sum_{n=1}^{N}\left(\frac{1}{M}\sum_{u\neq v}^{M}\left[s_{\boldsymbol{\theta}}\left(v_{\texttt{<LAN>}} \mid \overline{c}_n\left(u_{\texttt{<ENV>}}\right)\right) - s_{\boldsymbol{\theta}}\left(v_{\texttt{<LAN>}} \mid \overline{c}_n\left(v_{\texttt{<ENV>}}\right)\right)\right]\right).$$

We report $G_{\boldsymbol{\theta}} = \frac{1}{|V|}\sum_{v\in V} G_{\boldsymbol{\theta}}(v)$, and track $G_{\boldsymbol{\theta}^{(t)}}$ across training steps t to analyze how grounding emerges over time. An analysis of the formula is shown in the Appendix Section C.1.

We train LMs from random initialization. The details about the configurations of different models are shown in the Appendix Section B.4.

3 Behavioral Evidence

3.1 Behavioral Evidence of Emergent Grounding

In this section, we ask: **Does symbol grounding emerge behaviorally in autoregressive LMs?** We first test whether models show systematic surprisal reduction when predicting a linguistic token when its environmental counterpart is in context (Figure 1, where the gap between the lines represent the grounding information gain). For Transformers (Figures 1a), surprisal in the match condition decreases steadily while that in the mismatch condition enters a high-surprisal plateau early, indicating that the

models leverage environmental context to predict the linguistic form. In contrast, the unidirectional LSTM (Figure 8f) shows little separation between the conditions, reflecting the absence of grounding. Overall, these results provide behavioral evidence of emergent grounding: in sufficiently expressive architectures (Transformers and Mamba-2), the correct environmental context reliably lowers surprisal for its linguistic counterpart, whereas LSTMs fail to exhibit this effect, marking an architectural boundary on where grounding can emerge.

3.2 Behavioral Effects Beyond Co-occurrence

A natural concern is that the surprisal reductions might be fully explainable by shallow statistics: **the models might have simply memorized frequent co-occurrences of <ENV> and <LAN> tokens, without learning a deeper and more general mapping.** We test this hypothesis by comparing the tokens' co-occurrence with the grounding information gain in the child-directed speech data.

We define co-occurrence between the corresponding <ENV> and <LAN> tokens at the granularity of a 512-token training chunk. For each target word v, we count the number of chunks in which both its <ENV> and <LAN> tokens appear. Following standard corpus-analysis practice, these raw counts are log-transformed. For each model checkpoint, we run linear regression between the log co-occurrence and the grounding information gain of words, obtaining an R^2 statistic as a function of training time.

Figure 1b shows the R^2 values (orange) alongside the grounding information gain (blue) for transformer models. More structures are shown in Figure 9. In both the Transformer and Mamba-2, R^2 rises sharply at the early steps but then goes down, even if the grounding information gain continues increasing. These results suggest that grounding in Transformers and Mamba-2 cannot be fully accounted for by co-occurrence statistics: while models initially exploit surface co-occurrence regularities, later improvements in grounding diverge from these statistics, indicating reliance on richer and more complicated features acquired during training. In contrast, LSTM shows persistently increasing R^2 but little increase in grounding information gain over training steps, suggesting that it encodes co-occurrence but lacks the architectural mechanism to transform it into predictive grounding.

We also test whether the grounding effects observed in CHILDES generalize to multimodal dialogue, using the Visual Dialog dataset. These results are shown in the Appendix Section C.2.

4 Mechanistic Explanation

In this section, we provide a mechanistic and interpretable account of the previous observation. We focus on a 12-layer Transformer trained on CHILDES with 5 random seeds, and defer broader generalization to the discussion (Appendix E).

4.1 The Emergence of Symbol Grounding

To provide a mechanistic account of symbol grounding, i.e., when it emerges during training and how it is represented in the network, we apply two interpretability analyses.

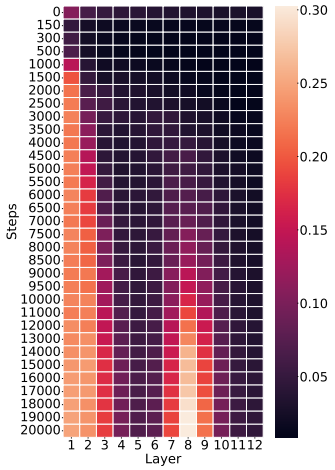
Saliency flow. For each layer ℓ, we compute a saliency matrix following Wang et al. [65]: $I_\ell = \left|\sum_h A_{h,\ell} \odot \frac{\partial \mathcal{L}}{\partial A_{h,\ell}}\right|$, where $A_{h,\ell}$ denotes the attention matrix of head h in layer ℓ. Each entry of I_ℓ quantifies the contribution of the corresponding attention weight to the cross-entropy loss $\mathcal{L}$, averaged across heads. Our analysis focuses on ground-to-symbol connections, i.e., flows from environmental ground (<ENV>) tokens to the token immediately preceding (and predicting) their linguistic forms (<LAN>).

Probing with the Tuned Lens. We probe layer-wise representations using the Tuned Lens [3], which trains affine projectors to map intermediate activations to the final prediction space while keeping the LM output head frozen.

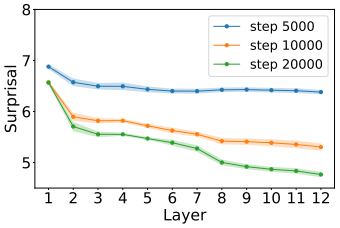
Results. Ground-to-symbol saliency is weak in the early stages of training but rises sharply later, peaking in layers 7–9 (Figure 2a), suggesting that mid-layer attention plays a central role in establishing symbol–ground correspondences. In addition, Figure 2b shows that early layers remain poor predictors even at late training stages (e.g., after 20,000 steps), whereas surprisal begins to drop markedly from layer 7 at intermediate stages (step 10,000), suggesting a potential representational shift in the middle layers.

4.2 Hypothesis: Gather-and-Aggregate Heads Implement Symbol Grounding

Building on these results, we hypothesize that specific Transformer heads in the middle layers enable symbol grounding. To test this, we examine attention saliencies for selected heads (Figure 7). We find that several heads exhibit patterns consistent with the gather and aggregate mechanisms described by Bick et al. [4]: gather heads (e.g., Figures 7a and 7b) compress relevant information into a subset of positions, while aggregate heads (e.g., Figures 7c and 7d) redistribute this information to downstream tokens. In our setups, saliency often concentrates on environmental tokens such as *train*$_{\texttt{<ENV>}}$, where gather heads pool contextual information into compact, retrievable states. In turn, aggregate heads broadcast this information from environmental ground (*train*<ENV>) to the token immediately preceding the linguistic form, thereby supporting the prediction of *train*$_{\texttt{<LAN>}}$. Taking these observations together, we hypothesize that the gather-and-aggregate heads implement the symbol grounding mechanism.



(a) Saliency of layer-wise attention from environmental to linguistic tokens across training steps.



(b) Layer-wise tuned lens to predict the <LAN> token in match condition.

Figure 2: Overtime mechanistic analysis on GPT-CHILDES.

4.3 Causal Interventions of Attention Heads

We then conduct causal interventions of attention heads to validate our previous hypothesis.

Operational definition. We identify attention heads as gather or aggregate following these standards:

- **Gather head.** An attention head is classified as a gather head if at least 30% of its total saliency is directed toward the environmental ground token from the previous ones.
- **Aggregate head**: An attention head is classified as an aggregate head if at least 30% of its total saliency flows from the environmental ground token to the token immediately preceding the corresponding linguistic token.

Causal intervention methods. In each context, we apply causal interventions to the identified head types and their corresponding controls. Following Bick et al. [4], interventions are implemented by zeroing out the outputs of heads. For the control, we mask an equal number of randomly selected heads in each layer, ensuring they do not overlap with the identified gather or aggregate heads.

Results and discussions. As training progresses, the number of both gather and aggregate heads increases (Table 2), suggesting that these mechanisms emerge over the course of learning. Causal interventions reveal a clear dissociation: zeroing out aggregate heads consistently produces significantly higher surprisal compared to controls, whereas the gather head interventions have no such effect. This asymmetry suggests that gather heads serve in a role less critical in our settings, where the input template is semantically light and the environmental evidence alone suffices to shape the linguistic form. Layer-wise patterns further support this division of labor: gather heads cluster in shallow layers (3-4), while aggregate heads concentrate in mid layers (7-8). This resonates with our earlier probing results, where surprisal reductions became prominent only from layers 7-9. Together, these findings highlight aggregate heads in the middle layers as the primary account of grounding in the model.

Table 2: Causal intervention results on identified gather and aggregate heads across training checkpoints (ckpt.). **Avg. Count** denotes the average number of heads of each type over inference times, and **Avg. Layer** denotes the average layer index where they appear. **Interv. Sps.** reports surprisal after zeroing out the identified heads, while **Ctrl. Sps.** reports surprisal after zeroing out an equal number of randomly selected heads. **Original** refers to the baseline surprisal without any intervention. *** indicates a significant result ($p < 0.001$) where the intervention surprisal is higher than that in the corresponding control experiment.

Ckpt.	Gather Head				Aggregate Head				Original
	Avg. Count	Avg. Layer	Interv. Sps.	Ctrl. Sps.	Avg. Count	Avg. Layer	Interv. Sps.	Ctrl. Sps.	
5000	0.35	3.32	6.37	6.38	2.28	7.38	**6.51** (***)	6.39	6.38
10000	3.26	3.67	5.25	5.32	5.09	7.28	**5.86** (***)	5.29	5.30
20000	5.76	3.59	4.69	4.79	6.71	7.52	**5.62** (***)	4.76	4.77

The result can be generalized to visual dialog with images, shown in the Appendix Section C.3.

References

[1] Anthropic. The claude 3 model family: Opus, sonnet, haiku, March 2024. URL https://www.anthropic.com/news/claude-3-family.

[2] Aryaman Arora, Neil Rathi, Nikil Roashan Selvam, Róbert Csórdas, Dan Jurafsky, and Christopher Potts. Mechanistic evaluation of transformers and state space models. *arXiv preprint arXiv:2505.15105*, 2025.

[3] Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Igor Ostrovsky, Lev McKinney, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*, 2023.

[4] Aviv Bick, Eric P. Xing, and Albert Gu. Understanding the skill gap in recurrent models: The role of the gather-and-aggregate mechanism. In *Forty-second International Conference on Machine Learning*, 2025.

[5] Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pages 2397–2430. PMLR, 2023.

[6] Alberto Bietti, Vivien Cabannes, Diane Bouchacourt, Herve Jegou, and Leon Bottou. Birth of a transformer: A memory viewpoint. *Advances in Neural Information Processing Systems*, 2023.

[7] Terra Blevins, Hila Gonen, and Luke Zettlemoyer. Analyzing the mono-and cross-lingual pretraining dynamics of multilingual language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3575–3590, 2022.

[8] Walid Bousselham, Felix Petersen, Vittorio Ferrari, and Hilde Kuehne. Grounding everything: Emerging localization properties in vision-language transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3828–3837, 2024.

[9] Shengcao Cao, Liang-Yan Gui, and Yu-Xiong Wang. Emerging pixel grounding in large multimodal models without grounding supervision. In *ICCV Findings*, 2025.

[10] Tyler A Chang and Benjamin K Bergen. Word acquisition in neural language models. *Transactions of the Association for Computational Linguistics*, 10:1–16, 2022.

[11] Tyler A Chang, Zhuowen Tu, and Benjamin K Bergen. Characterizing learning curves during language model pre-training: Learning, forgetting, and stability. *Transactions of the Association for Computational Linguistics*, 12:1346–1362, 2024.

[12] Jierun Chen, Fangyun Wei, Jinjing Zhao, Sizhe Song, Bohuai Wu, Zhuoxuan Peng, S-H Gary Chan, and Hongyang Zhang. Revisiting referring expression comprehension evaluation in the era of large multimodal models. *arXiv preprint arXiv:2406.16866*, 2024.

[13] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023.

[14] Xuweiyi Chen, Ziqiao Ma, Xuejun Zhang, Sihan Xu, Shengyi Qian, Jianing Yang, David Fouhey, and Joyce Chai. Multi-object hallucination in vision language models. *Advances in Neural Information Processing Systems*, 37:44393–44418, 2024.

[15] Eve V Clark. *The lexicon in acquisition*. Number 65. Cambridge University Press, 1995.

[16] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

[17] Tri Dao and Albert Gu. Transformers are ssms: Generalized models and efficient algorithms through structured state space duality. In *International Conference on Machine Learning*, pages 10041–10071. PMLR, 2024.

[18] Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *The Twelfth International Conference on Learning Representations*, 2024.

[19] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. Visual dialog. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 326–335, 2017.

[20] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020.

[21] Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. https://transformer-circuits.pub/2021/framework/index.html.

[22] Linnea Evanson, Yair Lakretz, and Jean-Rémi King. Language acquisition: do children and language models follow similar learning stages? In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12205–12218, 2023.

[23] Afsaneh Fazly, Afra Alishahi, and Suzanne Stevenson. A probabilistic computational model of cross-situational word learning. *Cognitive Science*, 34(6):1017–1063, 2010.

[24] Larry Fenson, Virginia A Marchman, Donna J Thal, Phillip S Dale, J Steven Reznick, and Elizabeth Bates. Macarthur-bates communicative development inventories. *PsycTESTS Dataset*, 2006.

[25] Lila R Gleitman and Barbara Landau. *The acquisition of the lexicon*. MIT Press, 1994.

[26] Noah Goodman, Joshua Tenenbaum, and Michael Black. A bayesian framework for cross-situational word-learning. *Advances in neural information processing systems*, 20, 2007.

[27] Reto Gubelmann. Pragmatic norms are all you need–why the symbol grounding problem does not apply to llms. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11663–11678, 2024.

[28] Thilo Hagendorff. Machine psychology: Investigating emergent capabilities and behavior in large language models using psychological methods. *arXiv preprint arXiv:2303.13988*, 2023.

[29] Stevan Harnad. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1-3): 335–346, 1990.

[30] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8): 1735–1780, 1997.

[31] Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13418–13427, 2024.

[32] Zhangqi Jiang, Junkai Chen, Beier Zhu, Tingjin Luo, Yankun Shen, and Xu Yang. Devils in middle layers of large vision-language models: Interpreting, detecting and mitigating object hallucinations via attention lens. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 25004–25014, 2025.

[33] Sara Kangaslahti, Elan Rosenfeld, and Naomi Saphra. Hidden breakthroughs in language model training. *arXiv preprint arXiv:2506.15872*, 2025.

[34] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975, 2022.

[35] Tom Lieberum, Matthew Rahtz, János Kramár, Neel Nanda, Geoffrey Irving, Rohin Shah, and Vladimir Mikulik. Does circuit analysis interpretability scale? Evidence from multiple choice capabilities in chinchilla. *arXiv preprint arXiv:2307.09458*, 2023.

[36] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[37] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in neural information processing systems*, volume 36, pages 34892–34916, 2023.

[38] Sheng Lu, Irina Bigoulaeva, Rachneet Sachdeva, Harish Tayyar Madabushi, and Iryna Gurevych. Are emergent abilities in large language models just in-context learning? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5098–5139, 2024.

[39] Ziqiao Ma, Jiayi Pan, and Joyce Chai. World-to-words: Grounded open vocabulary acquisition through fast mapping in vision-language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 524–544, 2023.

[40] Ziqiao Ma, Zekun Wang, and Joyce Chai. Babysit a language model from scratch: Interactive language learning by trials and demonstrations. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 991–1010, 2025.

[41] Brian MacWhinney. The childes project: Tools for analyzing talk: Volume i: Transcription format and programs, volume ii: The database, 2000.

[42] Jiayuan Mao, Chuang Gan, Pushmeet Kohli, Joshua B. Tenenbaum, and Jiajun Wu. The neuro-symbolic concept learner: Interpreting scenes, words, sentences from natural supervision. *International Conference on Learning Representations (ICLR)*, 2019.

[43] Jiayuan Mao, Freda H. Shi, Jiajun Wu, Roger P. Levy, and Joshua B. Tenenbaum. Grammar-based grounded lexicon learning. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021.

[44] Kevin Meng, David Bau, Alex J Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, 2022.

[45] Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. In-context learning and induction heads. *Transformer Circuits Thread*, 2022. https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html.

[46] OpenAI. Hello gpt-4o, May 2024. URL https://openai.com/index/hello-gpt-4o/.

[47] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research Journal*, pages 1–31, 2024.

[48] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, Qixiang Ye, and Furu Wei. Grounding multimodal large language models to the world. In *The Twelfth International Conference on Learning Representations*, 2024.

[49] Sarah Pratt, Mark Yatskar, Luca Weihs, Ali Farhadi, and Aniruddha Kembhavi. Grounded situation recognition. In *European Conference on Computer Vision*, pages 314–332. Springer, 2020.

[50] Shaolin Qu and Joyce Yue Chai. Context-based word acquisition for situated dialogue in a virtual world. *Journal of Artificial Intelligence Research*, 37:247–277, 2010.

[51] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[52] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[53] Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M Anwer, Erix Xing, Ming-Hsuan Yang, and Fahad S Khan. Glamm: Pixel grounding large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[54] Terry Regier. The emergence of words: Attentional learning in form and meaning. *Cognitive science*, 29(6):819–865, 2005.

[55] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[56] Deb K Roy and Alex P Pentland. Learning words from sights and sounds: A computational model. *Cognitive science*, 26(1):113–146, 2002.

[57] Masoud Jalili Sabet, Philipp Dufter, François Yvon, and Hinrich Schütze. Simalign: High quality word alignments without parallel training data using static and contextualized embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020.

[58] Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. Are emergent abilities of large language models a mirage? *Advances in Neural Information Processing Systems*, 36, 2023.

[59] Dominik Schnaus, Nikita Araslanov, and Daniel Cremers. It's a (blind) match! Towards vision-language correspondence without parallel data. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24983–24992, 2025.

[60] Thibault Sellam, Steve Yadlowsky, Ian Tenney, Jason Wei, Naomi Saphra, Alexander D'Amour, Tal Linzen, Jasmijn Bastings, Iulia Raluca Turc, Jacob Eisenstein, et al. The multiberts: Bert reproductions for robustness analysis. In *International Conference on Learning Representations*, 2021.

[61] Haoyue Shi, Luke Zettlemoyer, and Sida I. Wang. Bilingual lexicon induction via unsupervised bitext construction and word alignment. In *ACL*, 2021.

[62] Jeffrey Mark Siskind. A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition*, 61(1-2):39–91, 1996.

[63] Oskar van der Wal, Pietro Lesci, Max Müller-Eberstein, Naomi Saphra, Hailey Schoelkopf, Willem Zuidema, and Stella Biderman. Polypythias: Stability and outliers across fifty language model pre-training runs. In *Proceedings of the Thirteenth International Conference on Learning Representations (ICLR 2025)*, pages 1–25, 2025.

[64] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[65] Lean Wang, Lei Li, Damai Dai, Deli Chen, Hao Zhou, Fandong Meng, Jie Zhou, and Xu Sun. Label words are anchors: An information flow perspective for understanding in-context learning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9840–9855, 2023.

[66] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Song XiXuan, et al. Cogvlm: Visual expert for pretrained language models. *Advances in Neural Information Processing Systems*, 37:121475–121499, 2024.

[67] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *Transactions on Machine Learning Research*, 2022.

[68] Sarah Wiegreffe, Oyvind Tafjord, Yonatan Belinkov, Hannaneh Hajishirzi, and Ashish Sabharwal. Answer, assemble, ace: Understanding how LMs answer multiple choice questions. In *The Thirteenth International Conference on Learning Representations*, 2025.

[69] Wenhao Wu, Yizhong Wang, Guangxuan Xiao, Hao Peng, and Yao Fu. Retrieval head mechanistically explains long-context factuality. In *The Thirteenth International Conference on Learning Representations*, 2025.

[70] Zhaofeng Wu, Dani Yogatama, Jiasen Lu, and Yoon Kim. The semantic hub hypothesis: Language models share semantic representations across languages and modalities. In *ICML*, 2025.

[71] Mengzhou Xia, Mikel Artetxe, Chunting Zhou, Xi Victoria Lin, Ramakanth Pasunuru, Danqi Chen, Luke Zettlemoyer, and Ves Stoyanov. Training trajectories of language models across scales. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13711–13738, 2023.

[72] Zhuofan Xia, Dongchen Han, Yizeng Han, Xuran Pan, Shiji Song, and Gao Huang. Gsva: Generalized segmentation via multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[73] Fei Xu and Joshua B Tenenbaum. Word learning as bayesian inference. *Psychological review*, 114(2):245, 2007.

[74] Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. Ferret: Refer and ground anything anywhere at any granularity. In *The Twelfth International Conference on Learning Representations*, 2024.

[75] Chen Yu. The emergence of links between lexical acquisition and object categorization: A computational study. *Connection science*, 17(3-4):381–397, 2005.

[76] Chen Yu and Dana H Ballard. A unified model of early word learning: Integrating statistical and social cues. *Neurocomputing*, 70(13-15):2149–2165, 2007.

[77] Haonan Yu and Jeffrey Mark Siskind. Grounded language learning from video described with sentences. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 53–63, 2013.

[78] Tao Zhang, Xiangtai Li, Hao Fei, Haobo Yuan, Shengqiong Wu, Shunping Ji, Chen Change Loy, and Shuicheng Yan. Omg-llava: Bridging image-level, object-level, pixel-level reasoning and understanding. *Advances in neural information processing systems*, 37:71737–71767, 2024.

[79] Yichi Zhang, Ziqiao Ma, Xiaofeng Gao, Suhaila Shakiah, Qiaozi Gao, and Joyce Chai. Groundhog: Grounding large language models to holistic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[80] Rosie Zhao, Naomi Saphra, and Sham M. Kakade. Distributional scaling laws for emergent capabilities. In *NeurIPS 2024 Workshop on Scientific Methods for Understanding Deep Learning*, 2024.

A Dataset Details

A.1 Three Different Datasets

Child-directed speech. The Child Language Data Exchange System [CHILDES; 41] provides transcripts of speech enriched with environmental annotations.[‡] We use the spoken utterances as the linguistic tokens (`<LAN>`) and the environmental descriptions as the environment tokens (`<ENV>`). The environmental context is drawn from three annotation types:

- **Local events**: simple events, pauses, long events, or remarks interleaved with the transcripts.
- **Action tiers**: actions performed by the speaker or listener (e.g., `%act: runs to toy box`). These also include cases where an action replaces speech (e.g., `0 [% kicks the ball]`).
- **Situational tiers**: situational information tied to utterances or to larger contexts (e.g., `%sit: dog is barking`).

Caption-grounded dialogue. The Visual Dialog dataset [19] pairs MSCOCO images [36] with sequential question-answering based multi-turn dialogues that exchange information about each image. Our setup uses MSCOCO captions as the environmental tokens (`<ENV>`) and the dialogue turns form the linguistic tokens (`<LAN>`). In this pseudo cross-modal setting, textual descriptions of visual scenes ground natural conversational interaction. Compared to CHILDES, this setup introduces richer semantics and longer utterances, while still using text-based inputs for both token types, thereby offering a stepping stone toward grounding in fully visual contexts.

Image-grounded dialogue. To move beyond textual proxies, we consider an image-grounded dialogue setup, using the same dataset as the caption-grounded dialogue setting. Here, a frozen vision transformer [ViT; 20] directly tokenizes each RGB image into patch embeddings, with each embedding treated as an `<ENV>` token, analogously to the visual tokens in modern VLMs. We use DINOv2 [47] as our ViT tokenizer, as it is trained purely on vision data without auxiliary text supervision [in contrast to models like CLIP; 52], thereby ensuring that environmental tokens capture only visual information. The linguistic tokens (`<LAN>`) remain unchanged from the caption-grounded dialogue setting, resulting in a realistic multimodal interaction where conversational utterances are grounded directly in visual input.

A.2 Context Templates

We select the target tokens following the given procedure:

1. Get a list of words, with their ENV and LAN frequency both greater than or equal to 100 in the CHILDES dataset;
2. Get another list of nouns from CDI;
3. Take intersection and select top 100 words (by frequency of their ENV token) as target token list.

In CHILDES, all contexts are created with `gpt-4o-mini` followed by human verification if the genrated contexts are semantically light. We adopt the following prompt:

In visual dialogue (caption version and VLM version), we pre-define 10 sets of templates for each version:

A.3 Word Lists

CHILDES and Visual Dialog (Text Only). [box, book, ball, hand, paper, table, toy, head, car, chair, room, picture, doll, cup, towel, door, mouth, camera, duck, face, truck, bottle, puzzle, bird, tape, finger, bucket, block, stick, elephant, hat, bed, arm, dog, kitchen, spoon, hair, blanket, horse, tray, train, cow, foot, couch, necklace, cookie, plate, telephone, window, brush, ear, pig, purse, hammer, cat, shoulder, garage, button, monkey, pencil, shoe, drawer, leg, bear, milk, egg, bowl, juice, ladder, basket, coffee, bus, food, apple, bench, sheep, airplane, comb, bread, eye, animal, knee, shirt, cracker, glass, light, game, cheese, sofa, giraffe, turtle, stove, clock, star, refrigerator, banana, napkin, bunny, farm, money]

Visual Dialog (VLM). [box, book, table, toy, car, chair, doll, door, camera, duck, truck, bottle, bird, elephant, hat, bed, dog, spoon, horse, train, couch, necklace, cookie, plate, telephone, window, pig,

[‡] See the manual for data usage: https://talkbank.org/0info/manuals/CHAT.pdf

Prompt Templates for CHILDES

```
Given the word “{word}”, create 3 pairs of sentences
that follow this
requirement:
1. The first sentence has a subject “The child”, describing an
event or
situation, and has the word “{word}”. Make sure to add a
newline to the end of
this first sentence
2. The second sentence is said by the child (only include the
speech itself,
don’t include “the child say”, etc.), and the word “{word}”
also appears in
the sentence said by the child. Do not add quote marks either
3. Print each sentence on one line. Do not include anything
else.
4. Each sentence should be short, less than 10 words.
5. The word “{word}” in both sentence have the same meaning
and have a clear
indication or an implication relationship.
6. “{word}” should not appear at the first/second word of each
sentence.
Generate 3 pairs of such sentences, so there should be 6 lines
in total.
You should not add a number.
For each line, just print out the sentence.
```

Prompt Templates for Visual Dialogue (Caption Version)

```
this:<ENV> is:<ENV> [FILLER]:<ENV> <Q> what:<LAN> is:<LAN> it:
<LAN> <A>
(predict [FILLER]:<LAN>)

this:<ENV> is:<ENV> [FILLER]:<ENV> <Q> what:<LAN> do:<LAN> you:
<LAN>
call:<LAN> this:<LAN> <A> (predict [FILLER]:<LAN>)

this:<ENV> is:<ENV> [FILLER]:<ENV> <Q> can:<LAN> you:<LAN>
name:<LAN> this:<LAN> object:<LAN> <A>
(predict [FILLER]:<LAN>)

this:<ENV> is:<ENV> [FILLER]:<ENV> <Q> what’s:<LAN>
this:<LAN> called:<LAN> <A>
(predict [FILLER]:<LAN>)

this:<ENV> is:<ENV> [FILLER]:<ENV> <Q> what:<LAN>
this:<LAN> thing:<LAN> is:<LAN> <A>
(predict [FILLER]:<LAN>)
```

cat, monkey, drawer, bear, milk, egg, bowl, juice, ladder, bus, food, apple, sheep, bread, animal, shirt, cheese, giraffe, clock, refrigerator, accordion, aircraft, alpaca, ambulance, ant, antelope, backpack, bagel, balloon, barrel, bathtub, beard, bee, beer, beetle, bicycle, bidet, billboard, boat, bookcase, boot, boy, broccoli, building, bull, burrito, bust, butterfly, cabbage, cabinetry, cake, camel, canary, candle, candy, cannon, canoe, carrot, cart, castle, caterpillar, cattle, cello, cheetah, chicken, chopsticks, closet, clothing, coat, cocktail, coffeemaker, coin, cosmetics]

Prompt Templates for Visual Dialogue (Caption Version) (continued)

```
this:<ENV> is:<ENV> [FILLER]:<ENV> <Q> what:<LAN>
would:<LAN> you:<LAN> name:<LAN> this:<LAN> <A>
(predict [FILLER]:<LAN>)

this:<ENV> is:<ENV> [FILLER]:<ENV> <Q>
what's:<LAN> the:<LAN> name:<LAN> of:<LAN> this:<LAN>
item:<LAN> <A> (predict [FILLER]:<LAN>)

this:<ENV> is:<ENV> [FILLER]:<ENV> <Q> how:<LAN>
do:<LAN> you:<LAN> identify:<LAN> this:<LAN> <A>
(predict [FILLER]:<LAN>)

this:<ENV> is:<ENV> [FILLER]:<ENV> <Q> what:<LAN>
do:<LAN> we:<LAN> have:<LAN> here:<LAN> <A>
(predict [FILLER]:<LAN>)

this:<ENV> is:<ENV> [FILLER]:<ENV> <Q> how:<LAN>
do:<LAN> you:<LAN> call:<LAN> this:<LAN>
object:<LAN> <A> (predict [FILLER]:<LAN>)
```

Prompt Templates for Visual Dialogue (VLM Version)

```
"<image> \nwhat is it ?",
"<image> \nwhat do you call this ?",
"<image> \ncan you name this object ?",
"<image> \nwhat is this called ?",
"<image> \nwhat this thing is ?",
"<image> \nwhat would you name this ?",
"<image> \nwhat is the name of this item ?",
"<image> \nhow do you identify this ?",
"<image> \nwhat do we have here ?",
"<image> \nhow do you call this object ?"
```

B Implementation Details

We outline the key implementation details in this section and provide links to the GitHub repositories:

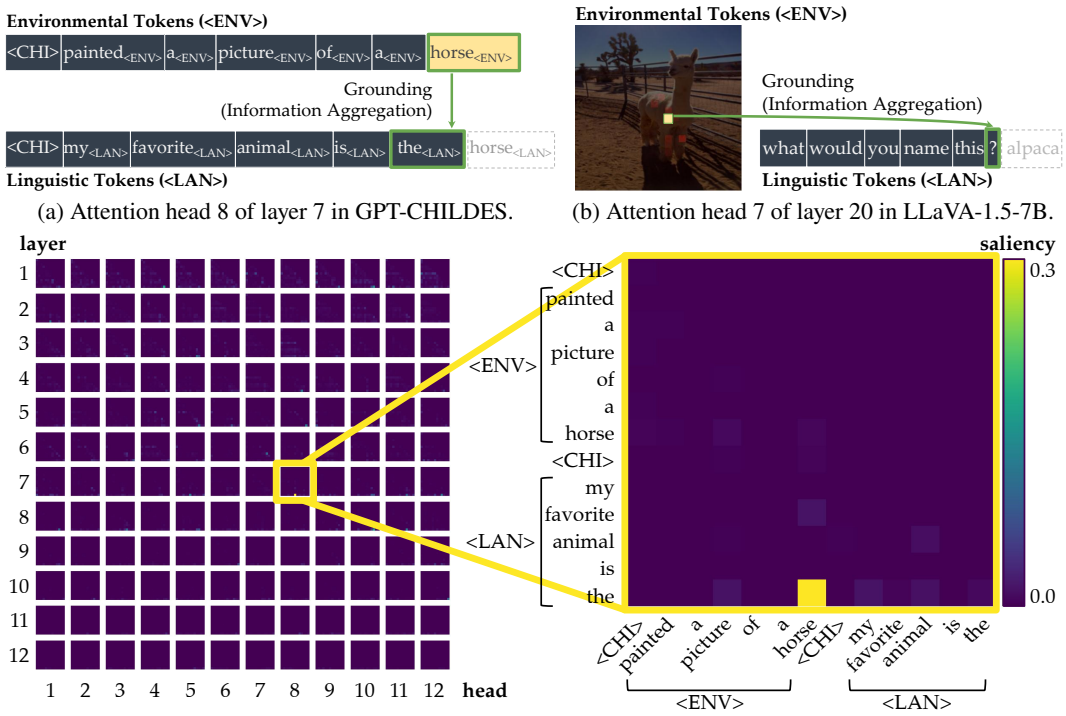
- Model Training: https://github.com/Mars-tin/TraBank
- CHILDES Processing: https://github.com/Mars-tin/PyChildes

B.1 Checkpointing

We save 33 checkpoints in total for text-only experiments and 16 checkpoints for the VLM setting.

CHILDES and Visual Dialog (Text Only). We save the intermediate steps: [0, 150, 300, 500, 1000, 1500, 2000, 2500, 3000, 3500, 4000, 4500, 5000, 5500, 6000, 6500, 7000, 7500, 8000, 8500, 9000, 9500, 10000, 11000, 12000, 13000, 14000, 15000, 16000, 17000, 18000, 19000, 20000]

Visual Dialog (VLM). We save the intermediate steps: [10000, 20000, 40000, 60000, 80000, 100000, 120000, 140000, 160000, 180000, 200000, 220000, 240000, 260000, 280000, 300000]



(a) Attention head 8 of layer 7 in GPT-CHILDES.

(b) Attention head 7 of layer 20 in LLaVA-1.5-7B.

(c) Left: saliency over tokens of each head in each layer for the prompt $\langle CHI \rangle$ *painted*$_{\langle ENV \rangle}$ *a*$_{\langle ENV \rangle}$ *picture*$_{\langle ENV \rangle}$ *of*$_{\langle ENV \rangle}$ *a*$_{\langle ENV \rangle}$ *horse*$_{\langle ENV \rangle}$ $\langle CHI \rangle$ *my*$_{\langle LAN \rangle}$ *favorite*$_{\langle LAN \rangle}$ *animal*$_{\langle LAN \rangle}$ *is*$_{\langle LAN \rangle}$ *the*$_{\langle LAN \rangle}$. Right: among all, only one of them (head 8 of layer 7) is identified as an aggregate head, where information flows from *horse*$_{\langle ENV \rangle}$ to the current position, encouraging the model to predict *horse*$_{\langle LAN \rangle}$ as the next token.

Figure 3: Illustration of the symbol grounding mechanism through information aggregation. Lighter colors denote more salient attention, quantified by saliency scores, i.e., gradient $\times$ attention contributions to the loss [65]. When predicting the next token, aggregate heads [4] emerge to exclusively link environmental tokens (visual or situational context; <ENV>) to linguistic tokens (words in text; <LAN>). These heads provide a mechanistic pathway for symbol grounding by mapping external environmental evidence into its linguistic form.

B.2 Training details.

For the text-only Transformer, Mamba2, and LSTM models, we randomly initialize them from scratch. The training process is conducted five times, each with a different random seed (using seeds 42, 142, 242, 342, and 442, respectively). The batch size is 16.

For VLM models, we randomly initialize the language model backbone from scratch and keep the DINOv2 vision encoder frozen. The training process is conducted five times for 300k steps, each with a different random seed (using seed 42, 142, 242, 342, and 442, respectively).

All the models use a word-level tokenizer. A list of hyperparameters is shown below:

Transformer and LSTM Model.

- model_max_length: 512
- learning rate: 5e-5
- learning rate schedule: linear
- warmup_steps: 1000
- hidden_size: 768
- beta1: 0.9
- beta2: 0.95
- weight_decay: 0
- batch_size: 16
- grad_clip_norm: 1.0

Mamba2 Model.

- model_max_length: 512
- learning rate: 4e-4
- learning rate schedule: linear
- warmup_steps: 2000
- hidden_size: 768
- beta1: 0.9
- beta2: 0.95
- weight_decay: 0.4
- batch_size: 16
- grad_clip_norm: 1.0

VLM Model.

- model_max_length: 1024
- learning rate: 2e-5
- learning rate schedule: cosine
- warmup_steps: 9000
- hidden_size: 768
- beta1: 0.9
- beta2: 0.95
- weight_decay: 0
- batch_size: 16
- grad_clip_norm: 1.0

B.3 Computational resources.

Each Transformer, Mamba2, and LSTM model is trained on a single A40 GPU within 5 hours. For VLM models, training is conducted on 2 A40 GPUs over 15 hours, using a batch size of 8 per device.

B.4 Model specifications

We train LMs from random initialization, ensuring that no prior linguistic knowledge influences the results. Our training uses the standard causal language modeling objective, as in most generative LMs. To account for variability, we repeat all experiments with 5 random seeds, randomizing both model initialization and corpus shuffle order. Our primary architecture is Transformer [64] in the style of GPT-2 [51] with 18, 12, and 4 layers, with all of them having residual connections. We extend the experiments to 4-layer unidirectional LSTMs [30] with no residual connections, as well as 12- and 4-layer state-space models [specifically, Mamba-2; 17]. For fair comparison with LSTMs, the 4-layer Mamba-2 models do not involve residual connections, whereas the 12-layer ones do. For multimodal settings, while standard LLaVA [37] uses a two-layer perceptron to project ViT embeddings into the language model, we bypass this projection in our case and directly feed the DINOv2 representations into the LM. We obtain the developmental trajectory of the model by saving checkpoints at various training steps, sampling more heavily from earlier steps, following Chang and Bergen [10].

C Addendum to Results

C.1 Mathematics of Grounding Information Gain

The *grounding information gain* $G_{\boldsymbol{\theta}}(v)$ for v is defined as

$$G_{\boldsymbol{\theta}}(v) = \frac{1}{N}\sum_{n=1}^{N}\left(\frac{1}{M}\sum_{u\neq v}^{M}\left[s_{\boldsymbol{\theta}}\left(v_{\texttt{<LAN>}} \mid \bar{c}_n\left(u_{\texttt{<ENV>}}\right)\right) - s_{\boldsymbol{\theta}}\left(v_{\texttt{<LAN>}} \mid \bar{c}_n\left(v_{\texttt{<ENV>}}\right)\right)\right]\right).$$

This is a sample-based estimation of the expected log-likelihood ratio between the match and mismatch conditions

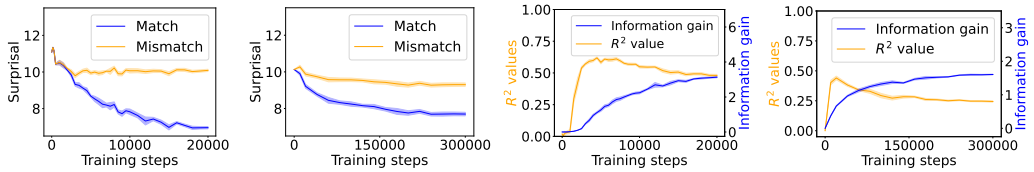
$$G_{\boldsymbol{\theta}}(v) = \mathbb{E}_{c,u}\left[\log\frac{P_{\boldsymbol{\theta}}(v_{\texttt{<LAN>}} \mid c, v_{\texttt{<ENV>}})}{P_{\boldsymbol{\theta}}(v_{\texttt{<LAN>}} \mid c, u_{\texttt{<ENV>}})}\right],$$

which quantifies how much more information the matched ground provides for predicting the linguistic form, compared to a mismatched one. A positive $G_{\boldsymbol{\theta}}(v)$ indicates that the matched environmental token increases the predictability of its linguistic form.

C.2 Visual Dialogue with Captions and Images

We also test whether the grounding effects observed in CHILDES generalize to multimodal dialogue, using the Visual Dialog dataset. In this setting, the environmental ground is supplied either by captions or by image features (Table 1). For caption-grounded dialogue, the mismatch context is constructed in the same way as for CHILDES (Equation 2). For image-grounded dialogue, mismatch contexts are generated via Stable Diffusion 2 [55]–based image inpainting, which re-generates the region defined by the ground-truth mask corresponding to the target word's referent.

We train 12-layer Transformers with 5 random seeds. Similarly as Figures 1a and Figures 1b, when captions serve as the environmental ground, Transformers show a clear surprisal gap between match and mismatch conditions (Figure 4a), with the grounding information gain increasing steadily while R^2 peaks early and declines (Figure 4c). Directly using image as grounds yields the same qualitative



(a) Surprisal curves (w/ caption). (b) Surprisal curves (w/ image). (c) R^2 and information gain (w/ caption). (d) R^2 and information gain (w/ image).

Figure 4: Average surprisal of the experimental and control conditions in caption- and image-grounded dialogue settings, as well as the grounding information gain and its correlation to the co-occurrence of linguistic and environment tokens over training steps. All results are from a 12-layer Transformer model on grounded dialogue data.

pattern (Figures 4b and 4d), although the observed effect is smaller. Both settings confirm that emergent grounding cannot be fully explained by co-occurrence statistics.

Overall, our findings demonstrate that Transformers are able to exploit environmental grounds in various modalities to facilitate linguistic prediction. The smaller but consistent gains in the image-grounded case suggest that while grounding from visual tokens is harder, the same architectural dynamics identified in textual testbeds still apply.

C.3 Causal Intervention Analysis to Visual Dialog with Images

We also conduct causal interventions of attention heads on the VLM model to further validate our previous hypothesis.

Operational definition. We identify attention heads as aggregate following this standard (We do not define gather head): An attention head is classified as an aggregate head if at least a certain threshold (70% or 90% in our experiment settings) of its total image patch to end saliency flows from the patches inside bounding box to the token immediately preceding the corresponding linguistic token.

Causal intervention methods. In each context, we apply causal interventions to the identified head types and their corresponding controls in the language backbone of the model. Similar to section 5.3, interventions are implemented by zeroing out a head's

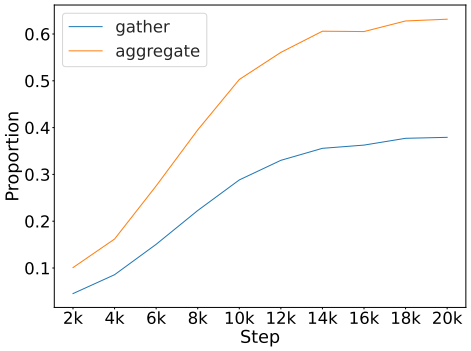


Figure 5: Gather-and-aggregate overtime.

Thres.	Ckpt.	Aggregate Head				Original
		Avg. Count	Avg. Layer	Interv. Sps.	Ctrl. Sps.	
70%	20k	32.30	7.78	9.96	9.95	9.21
	100k	35.63	7.71	**9.42** (***)	8.84	8.24
	200k	34.99	7.80	**8.95** (***)	8.15	7.76
	300k	34.15	7.76	**8.96** (***)	8.11	7.69
90%	20k	10.66	8.33	**9.51** (***)	9.43	9.21
	100k	13.90	8.26	**8.95** (***)	8.50	8.24
	200k	13.47	8.46	**8.41** (***)	7.88	7.76
	300k	12.73	8.42	**8.40** (***)	7.87	7.69

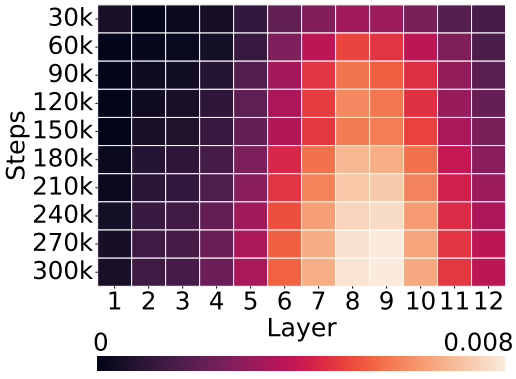


Figure 6: Mechanistic analysis in the image-grounded visual dialogue setting. Left: Causal intervention results on identified aggregate heads across training checkpoints, where intervention on aggregate heads consistently yields significantly higher surprisal ($p < 0.001$, ***) compared to the control group ones. Right: Saliency of layer-wise attention from environmental tokens (i.e., image tokens corresponding to patches within the bounding boxes of the target object) to linguistic tokens across training steps.

outputs. For the control, we mask an equal number
of randomly selected heads in each layer, ensuring they do not overlap with the identified aggregate heads.

Results and discussions. As training progresses, the number of aggregate heads increases first and then becomes steady (Figure 6), suggesting that these mechanisms emerge over the course of learning. Causal interventions reveal that zeroing out aggregate heads consistently produces significantly higher surprisal rises compared to controls. The average layer also align with the saliency heatmap, also shown in Figure 6.

C.4 Behavioral Analysis

We show the complete behavioral evidence for all models in Figure 8, and co-occurrence analysis in Figure 9.

C.5 Mechanistic Analysis

After identifying the set of gather and aggregate heads for each context, we conduct an overtime analysis to determine the proportion of saliency to the total saliency, as illustrated in Figure 5.

D Related Work

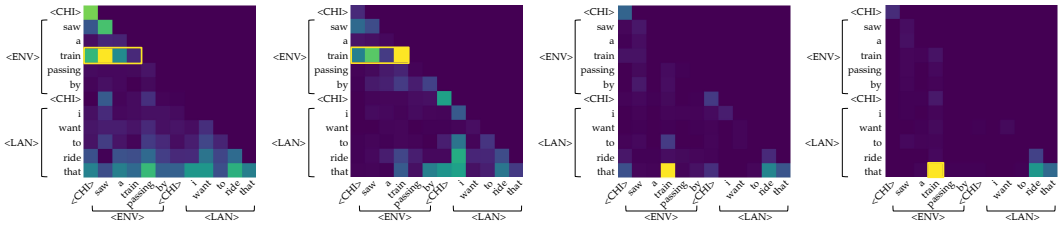
D.1 Language Grounding

Referential grounding has long been framed as the lexicon acquisition problem: how words map to referents in the world [29, 25, 15]. Early work focused on word-to-symbol mappings, designing learning mechanisms that simulate children's lexical acquisition and explain psycholinguistic phenomena [62, 54, 26, 23]. Subsequent studies incorporated visual grounding, first by aligning words with object categories [56, 75, 73, 76, 77], and later by mapping words to richer visual features [50, 42, 43, 49]. More recently, large-scale VLMs trained with paired text–image supervision have advanced grounding to finer levels of granularity, achieving region-level [34, 39, 13, 74, 66] and pixel-level [72, 53, 79] grounding, with strong performance on referring expression comprehension [12].

Recent work suggests that grounding emerges as a property of VLMs trained without explicit supervision, with evidence drawn from attention-based spatial localization [9, 8] and cross-modal geometric correspondences [59]. However, all prior work focused exclusively on static final-stage models, overlooking the training trajectory, a crucial aspect for understanding when and how grounding emerges. In addition, existing work has framed grounding through correlations between visual and textual signals, diverging from the definition by Harnad [29], which emphasizes causal links from symbols to meanings. To address these issues, we systematically examine learning dynamics throughout the training process, applying causal interventions to probe model internals and introducing control groups to enable rigorous comparison.

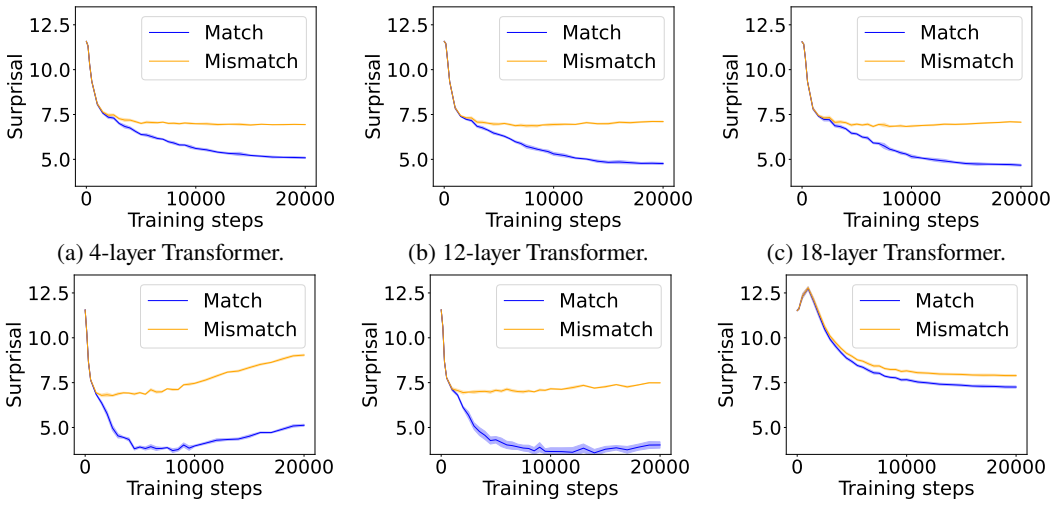
D.2 Emergent Capabilities and Learning Dynamics of LMs

A central debate concerns whether larger language models exhibit genuinely new behaviors: Wei et al. [67] highlight abrupt improvements in tasks, whereas later studies argue such effects are artifacts of thresholds or in-context learning dynamics [58, 38]. Beyond end performance, developmental analyses show that models acquire linguistic abilities in systematic though heterogeneous orders with variability across runs and checkpoints [60, 7, 5, 71, 63]. Psychology-inspired perspectives further emphasize controlled experimentation to assess these behaviors [28], and comparative studies



(a) Gather: L4 H7. (b) Gather: L4 H8. (c) Aggregate: L7 H5. (d) Aggregate: L8 H5.

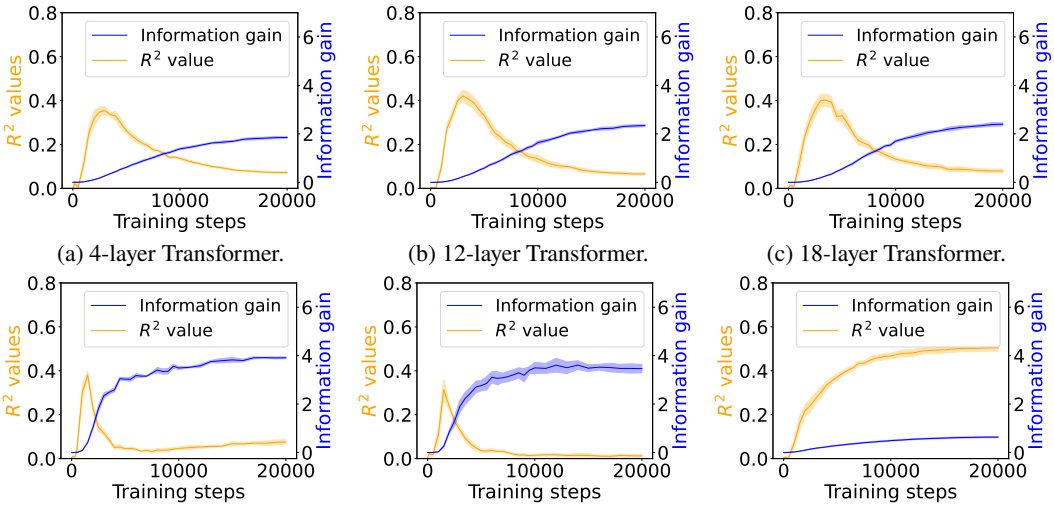
Figure 7: Examples of gather and aggregate heads identified in GPT-CHILDES. L: layer; H: head.



(a) 4-layer Transformer. (b) 12-layer Transformer. (c) 18-layer Transformer.

(d) 12-layer Mamba 2. (e) 4-layer Mamba 2. (f) 4-layer LSTM.

Figure 8: Average surprisal of the experimental and control conditions over training steps.



(a) 4-layer Transformer. (b) 12-layer Transformer. (c) 18-layer Transformer.

(d) 12-layer Mamba 2. (e) 4-layer Mamba 2. (f) 4-layer LSTM.

Figure 9: Grounding information gain and its correlation to the co-occurrence of linguistic and environment tokens over training steps.

reveal both parallels and divergences between machine and human language learning [10, 22, 11, 40]. At a finer granularity, hidden-loss analyses identify phase-like transitions [33], while distributional studies attribute emergence to stochastic differences across training seeds [80]. Together, emergent abilities are not sharp discontinuities but probabilistic outcomes of developmental learning dynamics. Following this line of work, we present a probability- and model internals–based analysis of how symbol grounding emerges during language model training.

D.3 Mechanistic Interpretability of LMs

Mechanistic interpretability has largely focused on attention heads in Transformers [21, 45, 44, 6, 35, 69]. A central line of work established that *induction heads* emerge to support in-context learning [ICL; 21, 45], with follow-up studies tracing their training dynamics [6] and mapping factual recall circuits [44]. At larger scales, Lieberum et al. [35] identified specialized *content-gatherer* and *correct-letter* heads, and Wu et al. [69] showed that a sparse set of *retrieval heads* is critical for reasoning and long-context performance. Relatedly, Wang et al. [65] demonstrated that label words in demonstrations act as *anchors*: early layers gather semantic information into these tokens, which later guide prediction. Based on these insights, Bick et al. [4] proposed that retrieval is implemented through a coordinated *gather-and-aggregate (G&A)* mechanism: some heads collect content from relevant tokens, while others aggregate it at the prediction position. Other studies extended this line of

work by analyzing failure modes and training dynamics [68] and contrasting retrieval mechanisms in Transformers and SSMs [2]. Whereas prior analyses typically investigate ICL with repeated syntactic or symbolic formats, our setup requires referential alignment between linguistic forms and their environmental contexts, providing a complementary testbed for naturalistic language grounding.

E Discussions

Generalization to full-scale VLMs. As an additional case study, we extend our grounding-as-aggregation hypothesis to a full-scale VLM, LLaVA-1.5-7B [37]. Even in this heavily engineered architecture, we identify many attention heads exhibiting aggregation behavior consistent with our earlier findings (Figure 3b), reinforcing the view that symbol grounding arises from specialized heads. At the same time, full-scale VLMs present additional complications. Models like LLaVA use multiple sets of visual tokens, including CLIP-derived embeddings that already encode language priors, and global information may be stored in redundant artifact tokens rather than object-centric regions [18]. Moreover, the large number of visual tokens (environmental tokens, in our setup) substantially increases both computational cost and the difficulty of isolating genuine aggregation heads. These factors make systematic identification and intervention at scale a nontrivial challenge. For these reasons, while our case study highlights promising evidence of grounding heads in modern VLMs, systematic detection and causal evaluation of such heads at scale remains an open challenge. Future work will need to develop computationally viable methods for (i) automatically detecting aggregation heads across diverse VLMs, and (ii) applying causal interventions to validate their role in grounding. Addressing these challenges will be crucial for moving from anecdotal case studies to a more principled understanding of grounding in modern VLMs.

The philosophical roots of grounding, revisited. Our findings highlight the need to sharpen the meaning of grounding in multimodal models. Prior work has often equated grounding with statistical correlations between visual and textual signals, such as attention overlaps or geometric alignments [8, 9, 59]. While informative, such correlations diverge from the classic formulation by Harnad [29], which requires symbols to be causally anchored to their referents in the environment. On the other extreme, Gubelmann [27] argued that the symbol grounding problem does not apply to LLMs as they "are connectionist, statistical devices that have no intrinsic symbolic structure." In contrast, we discover emergent symbolic structure as an intrinsic mechanistic property: one that can be traced along training, observed in the specialization of attention heads, and validated through causal interventions. This provides not only a practical diagnostic protocol that reveals when and how models genuinely tie symbols to meaning beyond surface-level correlations, but also challenges the view that grounding is philosophically irrelevant to systems without explicit symbolic structure.

Practical implications to LM hallucinations. Our findings have practical implications for improving the reliability of LM outputs: by identifying aggregation heads that mediate grounding between environmental and linguistic tokens, we provide a promising mechanism to detect model reliability before generation. Our findings echo a pathway to mitigate hallucinations by focusing on attention control: many hallucination errors stem from misallocated attention in intermediate layers [32, 14]. Such attention-level signals can serve as early indicators of overtrust or false grounding, motivating practical solutions like decoding-time strategies to mitigate and eventually prevent hallucination [31].