PNAS RESEARCH ARTICLE | PSYCHOLOGICAL AND COGNITIVE SCIENCES

Sensory context improves language prediction in humans and LLMs

Thomas L. Botch[a,1] and Emily S. Finn[a,1]

Edited by Susan Goldin-Meadow, The University of Chicago, Chicago, IL; received January 7, 2026; accepted July 29, 2026

Language is a fundamental human capacity. Large language models (LLMs) have presented the first viable model of language outside of humans, yet how these models learn and use language differs significantly from humans. Here, we compare LLMs and humans predicting language with varying levels of sensory information—from disembodied written text to audiovisual videos of speakers—to demonstrate that, in both humans and LLMs, sensory context is critical for optimal performance. We asked human participants to predict upcoming words within narratives presented as either audiovisual, audio-only, or written language (N = 1,500 total; 500 per modality), focusing on content words that convey essential meaning. We compared these predictions across modalities as well as to predictions generated by LLMs. Human predictions were overall more accurate than LLMs, regardless of the sensory modality of presentation. Compared to written language, both audiovisual and audio-only language increased the accuracy and consensus of human predictions and decreased alignment with LLMs. We identified that prosody, a signal that conveys important linguistic information via the auditory channel, partially drove both the observed advantage in accuracy within humans and divergence from LLMs. Integrating multimodal information—i.e., prosody, auditory, or audiovisual information—with the representation of language learned during LLM training improved models' next-word prediction performance and increased the efficiency of language learning. These findings demonstrate that sensory contexts are foundational to human-like language behavior, and that these contexts can enrich and accelerate language acquisition within LLMs similar to what is observed within human development.

cognitive science | human behavior | language | large language models

Significance

Sensory context—the auditory and visual channels through which language is transmitted—is thought to be critical for how humans evolved and develop language. Large language models (LLMs) provide the first viable model of language outside humans, yet learn language strictly through text. We compared humans predicting language while reading, listening, or watching audiovisual videos with text-based LLM predictions. While humans outperformed LLMs regardless of modality, increased sensory context further improved human performance and decreased alignment with LLMs. Integrating multimodal information during LLM training improved next-word prediction and language learning efficiency. These findings demonstrate that sensory contexts are foundational for human-like language behavior, highlighting the importance of embodied, multimodal experience for language processing.

Humans evolved language on a species level and develop language on an individual level through embodied interactions in diverse sensory and social contexts (1, 2). For humans, these contexts play a critical role in facilitating the language development (3, 4): observations of children and simulations of language learning in adults show that extralinguistic context—nonverbal elements, the physical environment, and broader world knowledge—unequivocally disambiguates linguistic signals (5, 6) and speeds vocabulary acquisition (7, 8).

Large language models (LLMs) show impressive abilities in both language understanding and generation (9–12), as well as in higher-order cognitive behaviors such as theory of mind (13, 14) and analogical reasoning (15, 16). Yet, LLMs significantly differ from humans in how they learn and use language (17–19): while humans rapidly acquire language in the first years of life despite a dearth of input (20, 21), LLMs are exposed to orders of magnitude more data (22), but this data is purely text-based and therefore removed from the sensory and social pressures that shape human language development.

The correspondence between humans and LLMs in their language representations and behaviors is of current debate (23–27). To what extent does the additional sensory context that humans, but not LLMs, have access to explain observed differences in their behavior? In other words, how essential are these contexts for cultivating "human-like" language behavior? Because LLMs are the first viable candidate of language outside of humans (28), understanding when and how language diverges between humans and LLMs provides a promising opportunity for identifying components that undergird the development of language writ large.

In this study, we develop a comparative framework for evaluating when and how sensory context—specifically, multimodal, audiovisual information—supports moment-to-moment language processing. We operationalized language processing as next-word prediction, the core training task of LLMs which also serves as a partial index for human linguistic abilities; this provides a straightforward point of comparison between humans and models. Our main hypothesis was that human language prediction leverages sensory

Author affiliations: [a]Department of Psychological and Brain Sciences, Dartmouth College, Hanover, NH 03755

Author contributions: T.L.B. and E.S.F. designed research; T.L.B. performed research; T.L.B. analyzed data; and T.L.B. and E.S.F. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: thomas.l.botch@dartmouth.edu or emily.s.finn@dartmouth.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2600317123/-/DCSupplemental.

Published August 26, 2026.

information that is not currently captured within text-based LLMs. To this end, we collected a large dataset (N = 1,500 participants) of human predictions of language during real-world, first-person narratives presented as video (audiovisual), audio-only, or writing (text). We compared the accuracy and distributions of human and LLM predictions of upcoming words. Then, to directly test the role of sensory context in language development, we integrated multimodal information during LLM training (as a proxy for human language development) and evaluated its effect on prediction accuracy and rate of language acquisition. Results show that access to increasingly rich sensory context enhances language prediction in both humans and LLMs, underscoring the fundamental role of multimodal information for acquiring and using language in human-like ways.

Results

Access to Sensory Information Boosts Accuracy of Humans' Next-Word Predictions. We asked human participants (N = 1,500 total; 500 per modality) to either watch (audiovisual condition; *AV*), listen to (audio-only condition; *A*), or read (written condition; *W*) a real-world narrative (one of three distinct personal stories; see *Materials and Methods*). At intermittent points during the story, we paused the video, audio, or transcript and asked participants to generate a prediction of the upcoming word. In contrast to prior studies comparing human and LLM predictions of language (29–31), which have assessed performance at any and all words (including, e.g., stop-words such as "the," "and," "is") and even nonwords (e.g., punctuation, emojis), we limited our focus to content words—i.e., nouns, verbs, adjectives, and adverbs that convey information and concepts essential to the narrative's meaning. We further subsampled target words based on LLMs' performance (*Materials and Methods*) to include words (tokens) for which LLMs were more vs. less confident in their predictions, allowing us to compare human and LLM performance at these particular moments (Fig. 1).

We first investigated whether processing language within sensory contexts boosts the accuracy of human language predictions. Indeed, human participants were more likely to predict the correct next word within richer sensory contexts (Fig. 2*A*; logistic mixed-effects model, $\beta_{AV>A>W} = -0.91$, SE (SE) = 0.03, odds ratio (OR) = 0.40, $z = -26.70$, $P < 0.001$). This boost in accuracy from increased sensory information extended to analyses of the human population. For each target word, we aggregated responses across participants to create a prediction distribution and selected the word most frequently predicted in each distribution as the representative prediction. Across the human population, predictions were more accurate with greater availability of sensory information (see *SI Appendix*, Fig. S1; logistic mixed-effects model, $\beta_{AV>A>W} = -1.49$, $SE = 0.16$, $OR = 0.23$, $z = -9.56$, $P < 0.001$).

However, binary accuracy imposes a stringent criterion that may not fairly capture the general correctness of predictions. For example, predicting the word "kid" when the correct word was "child" would be labeled an incorrect response, despite these two words being very similar in semantic content. Thus, to get a continuous measure of accuracy that takes into account semantic similarity, we embedded words within a word-level semantic space [*fasttext*, 300 dimensions; (32)] and calculated accuracy as the cosine similarity between the predicted and ground-truth words. Using this continuous metric, human predictions still exhibited higher accuracy (greater semantic similarity to the ground-truth word) within richer sensory contexts (Fig. 2*B*; $\beta_{AV>A>W} = -0.09$, $SE = 0.01$, $t(2011) = -11.08$, $P < 0.001$).

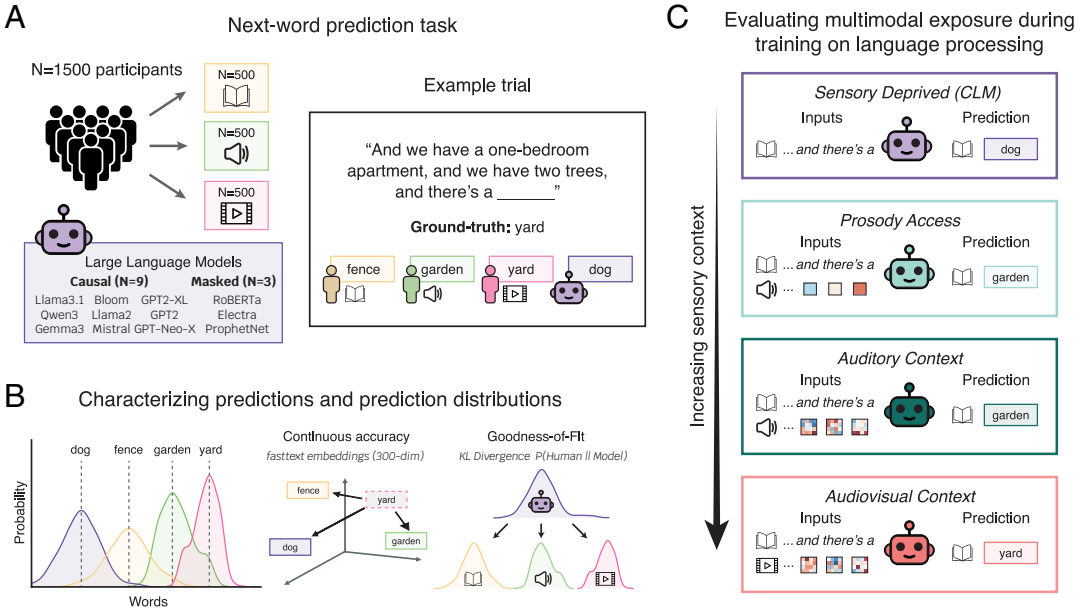

Fig. 1. Overview of methods. (*A*) Human participants (N = 1,500 total) predicted upcoming words within real-world narratives presented as either audiovisual, audio-only, or written language (N = 500 per modality). Large language model (LLM) predictions were sampled from both causal (N = 9) and masked (N = 3) language models for the same words. (*B*) Schematic of predictions and distributions. Words predicted by humans were aggregated into separate prediction distributions for each modality (audiovisual, audio-only, and written language) and compared to prediction distributions from LLMs. Continuous accuracy (i.e., semantic similarity) of predicted words to ground-truth words was compared across human and LLMs. Alignment of LLM predictions and prediction distributions was evaluated in relation to human behavior. (*C*) We assessed how access to increasing levels of sensory context (multimodal information) interacts with language prediction and acquisition within LLMs. We trained four different LLMs to predict the upcoming word using 1) written language alone (*sensory deprived*), or using written language integrated with 2) prosodic cues (a subset of the auditory signal; *prosody access*), 3) the full auditory signal (*auditory context*), or 4) full audiovisual features representing the holistic sensory context (*audiovisual context*).

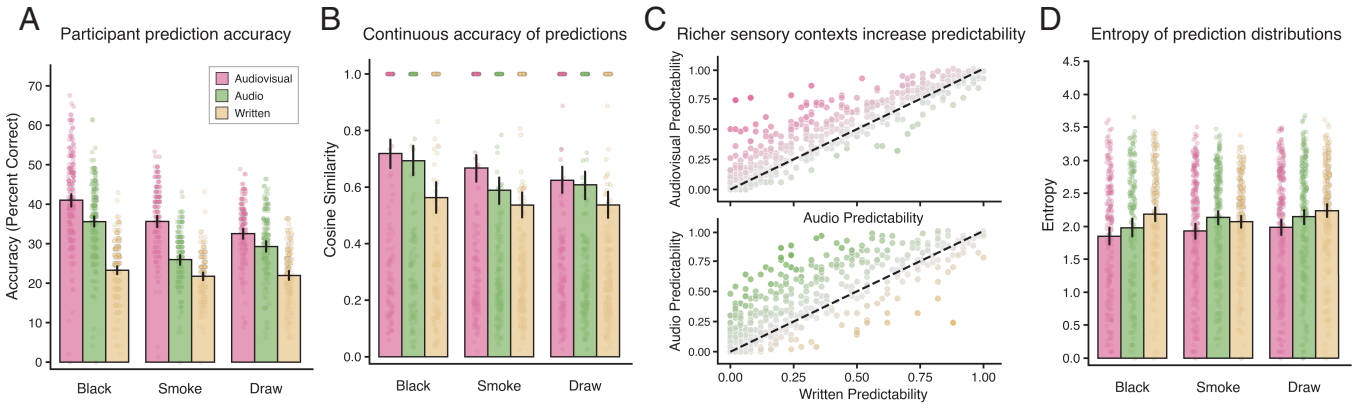

Fig. 2. Human next-word prediction performance. (*A*) In all three stories (x-axis), participant prediction accuracy increased with the amount of sensory context (binary accuracy averaged across all predicted words, y-axis); points represent individual participants. (*B*) Population-level top-1 predictions of audiovisual language were more semantically similar to the ground-truth word than predictions of both audio-only and written language; points represent predicted words. Richer sensory contexts (*C*) increased the proportion of participants predicting the correct word (predictability) and (*D*) decreased the entropy of prediction distributions; points represent predicted words. In (*C*), dashed lines indicate the identity line between two modalities; in (*D*), bars indicate mean values, error bars indicate 95% CI of the mean.

While these results suggest that sensory contexts improve humans' ability to predict language, it is possible that participants with access to the auditory signal (audiovisual and audio-only conditions) were unintentionally advantaged due to "leakage" of auditory information (i.e., hearing the first phoneme of the upcoming, target word). We therefore performed a post hoc analysis that identified moments of potential leakage (*Materials and Methods*) and reran the above analyses after removing these potentially confounded moments. Even when stringently controlling for potential leakage (removing 11% of target words), human predictions of both audiovisual and audio-only language continued to be more accurate than predictions of written language across all metrics of accuracy (*participant average accuracy*: $\beta_{AV>A>W} = -0.10$, $t(1495) = -25.76$; *binary accuracy*: $\beta_{AV>A>W} = -1.42$, $OR = 0.24$, $z = -8.74$; *continuous accuracy*: $\beta_{AV>A>W} = -0.08$, $t(1798) = -10.26$; all $P < 0.001$). These findings suggest that access to the sensory context that naturally surrounds language helps humans predict upcoming words.

Sensory Context Increases the Consensus of Predictions Across Human Participants. Language is often used in social contexts, when a speaker needs to convey a message to a listener or multiple listeners. Successful communication is generally accompanied by a cohesion of understanding across an audience (33–36). Does language informed by sensory context facilitate this group-level alignment better than written language? One way to conceptualize group-level alignment is a reduction in uncertainty, or an increase in "consensus," across the audience.

In all three stories, we found that language informed by sensory context exhibited greater predictability, operationalized as the proportion of participants predicting the correct word, than written text (Fig. 2*C*; $\beta_{AV>A>W} = -0.26$, $SE = 0.01$, $t(2011) = -25.61$, $P < 0.001$). However, higher predictability does not necessarily entail a reduction in uncertainty: other words within the distribution may still be frequently predicted across the population. To account for this possibility, we evaluated the uncertainty of the human prediction distributions for audiovisual, audio-only, and written language, operationalized as Shannon's entropy. Prediction distributions became less entropic within richer sensory contexts (Fig. 2*D*; $\beta_{AV>A>W} = 0.17$, $SE = 0.01$, $t(2011) = 11.23$, $P < 0.001$). Importantly, our results remained consistent when using bias-corrected entropy estimation methods for small sample data (Chao-Shen entropy: $\beta_{AV>A>W} = 0.17$, $SE = 0.02$, $t(2011) = 9.83$, $P < 0.001$; NSB entropy: $\beta_{AV>A>W} = 0.15$, $SE = 0.01$, $t(2011) = 10.82$, $P < 0.001$). These findings suggest that having access to the sensory context of language increases the consensus and certainty of predictions across the population of human participants.

Humans Outperform Large Language Models at Predicting Language. Humans acquire and use language within diverse sensory contexts. These contexts are often embedded within social scenarios, which are in turn entwined with physical interactions and communicative goals. In contrast to these rich, embodied environments, large language models (LLMs) are commonly trained solely on corpora of disembodied text (e.g., books, internet forums). Here, we asked whether this impoverished view of language learned by LLMs impacts their ability to predict language compared to human participants.

We first compared the binary accuracy (one-shot) between humans and language models. Human predictions in all conditions (audiovisual, audio-only, and written) were significantly more accurate than predictions from both causal (CLM) and masked (MLM) language models (*SI Appendix*, Fig. S2; $\beta_{AV>A>W>CLM>MLM} = -5.19$, $SE = 0.46$, $OR = 0.01$, $z = -11.20$, $P < 0.001$). This finding was stable even when accounting for word-level semantics with our continuous accuracy metric: human predictions were more semantically similar to the ground-truth word than predictions from either form of language model (Fig. 3*A*; $\beta_{AV>A>W>CLM>MLM} = -0.29$, $SE = 0.02$, $t(10074) = -12.77$, $P < 0.001$). Critically, this effect emerged consistently in each of the three stories for both binary (*SI Appendix*, Fig. S3; "*black*:" $\beta = -5.13$, $z = -10.97$; "*smoke*:" $\beta = -5.45$, $z = -10.96$; "*draw*:" $\beta = -4.98$, $z = -8.24$; all $P < 0.001$) and continuous accuracy (*SI Appendix*, Fig. S4; "*black*:" $\beta = -0.32$, $t(2904) = -13.67$; "*smoke*:" $\beta = -0.29$, $t(3549) = -13.76$; "*draw*:" $\beta = -0.26$, $t(3609) = -9.25$; all $P < 0.001$), underscoring that the human advantage for predicting language is robust across distinct stimuli. The effect also persisted when varying the context window (number of words) provided to LLMs (*SI Appendix*, Fig. S5; $\beta_{AV>A>W>CLM>MLM} = -0.37$, $SE = 0.03$, $t(167573) = -13.29$, $P < 0.001$; note that while performance of some of the largest models began to match or

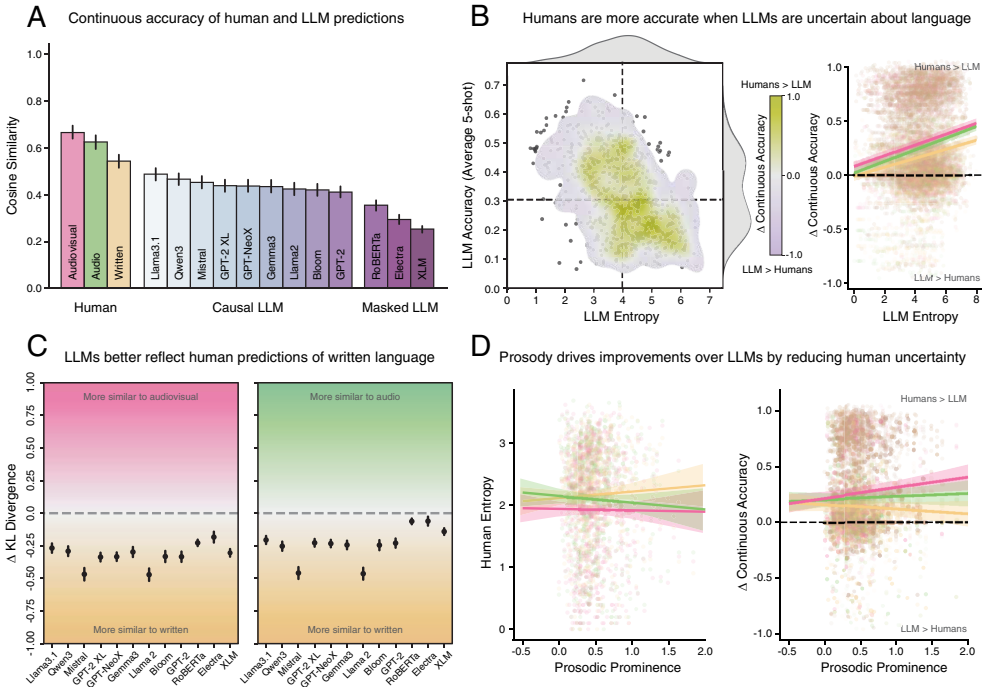

Fig. 3. Comparing humans' and LLMs' next-word predictions. (*A*) Continuous accuracy (semantic similarity) of predictions to the ground-truth word. Human predictions were more accurate than LLMs regardless of stimulus modality and LLM architecture. (*B*) *Left*: Difference of human and LLM continuous accuracy as a function of LLM accuracy (average 5-shot) and entropy (uncertainty). While accuracy generally decreased as LLM entropy increased, regardless of modality, humans' accuracy advantage over LLMs was stronger when LLMs were more uncertain (*Lower Right* quadrant). Dashed lines indicate the 50th percentile of LLM accuracy and entropy. *Right*: Improvements in continuous accuracy over LLMs increased with greater access to sensory information and as LLMs became more uncertain. (*C*) KL divergence between LLM and human prediction distributions. LLM prediction distributions exhibited better fits to human written language distributions. (*D*) *Left*: Prosodic features present in the window leading up to a target word reduced uncertainty of predictions of that word specifically for humans with access to sensory information. *Right*: Prosodic features increased the accuracy of human predictions of audiovisual and audio-only language over LLMs. In panels (*B* and *D*), points represent predicted words. In all panels, error bars or confidence bands indicate 95% CI of the mean.

exceed that of humans in the written condition at the largest context window sizes, no models ever achieved performance equivalent to humans in either sensory condition at any window size).

Given that humans were overall more accurate than LLMs, we next aimed to understand whether humans exhibited a particular advantage over models at certain moments (words) in the stories. We reasoned that if humans predicting language benefit from sensory context and embodied experience, either in the moment (audiovisual and audio-only conditions) or over a lifetime of experience (written condition), this information might confer a particular advantage at moments of otherwise high uncertainty (quantified as the entropy of the LLM prediction distributions). Of note, we sampled words a priori for the next-word prediction task to test this hypothesis about high vs. low uncertainty moments (*Materials and Methods*). To this end, we evaluated if and how the entropy of LLM prediction distributions for the upcoming, target word affected the observed difference in continuous prediction accuracy (human vs. LLM). Regardless of sensory context, the accuracy advantage for humans over LLMs increased as LLM predictions became more uncertain (Fig. 3 *B*, *Left*; $\beta = 0.05$, $SE = 1.9e^{-3}$, $t(24099) = 26.11$, $P < 0.001$). Importantly, humans in all conditions outperformed LLMs even when LLMs provided correct predictions for target words in both the high-accuracy/low-entropy quadrant (*SI Appendix*, Fig. S6; $\beta_{AV>CLM} = 0.18$, $z = 6.38$; $\beta_{A>CLM} = 0.19$, $z = 13.35$; $\beta_{W>CLM} = 0.13$, $z = 15.58$; all $P < 0.001$) and the high-accuracy/high-entropy quadrant (*SI Appendix*, Fig. S6; $\beta_{AV>CLM} = 0.21$, $z = 7.66$; $\beta_{A>CLM} = 0.18$, $z = 13.58$; $\beta_{W>CLM} = 0.11$, $z = 14.58$; all $P < 0.001$). Critically, there was also an interaction with sensory context such that this entropy-specific human accuracy advantage became increasingly pronounced as humans were provided with greater amounts of sensory context (Fig. 3 *B*, *Right*; quadratic term, $\beta_{AV>A>W} = -0.01$, $SE = 1.8e^{-3}$, $t(24099) = -3.63$, $P < 0.001$). Together, these results suggest that sensory information may serve to stabilize and aid human predictions of language particularly at moments of high uncertainty.

Large Language Models Are Better Aligned with Human Predictions of Written Text. Regardless of modality (audiovisual, audio-only, or written), human predictions of our target words were more accurate than those of LLMs. Yet, if our hypothesis is correct—that access to sensory context drives differences between humans across conditions and between humans and models—it would predict not only that models are less accurate, but also that models' predictions are more similar to humans in the written condition, since both are similarly deprived of in-the-moment sensory information. We therefore tested whether the modality through which humans process language affects the alignment of LLMs with human behavior.

We first evaluated the alignment of per-word "predictability" between humans and LLMs. In line with past work, we defined predictability as the probability assigned to the correct, ground-truth word (i.e., for humans, the proportion of participants predicting the correct word). This measure is therefore defined with reference to the correct response rather than the most probable response. Prior studies have suggested a strong relationship between human and LLM predictability (29, 30), yet these studies assessed human predictions under less naturalistic conditions (i.e., having participants predict every word in a transcript, including uninformative [stop] words). Here, we found a weak

relationship between human and LLM predictability, driven by an interaction with stimulus modality (audiovisual/audio-only/written). Specifically, LLM predictability exhibited a weak, positive relationship with human predictability of written language which decreased as humans were provided with additional sensory context (*SI Appendix*, Fig. S7; $\beta_{AV>A>W} = 4.5e^{-3}$, $SE = 5.3e^{-4}$, $z = 8.34$, $P < 0.001$). In other words, LLM predictability was, at best, weakly aligned with human predictability of written language. This relationship stands in contrast to prior studies—potentially due to their focus on written language and oversampling common words—and hints that the alignment of humans and LLMs may depend on the modality of the stimulus.

On its own, predictability provides incomplete insight into the structure of the prediction distribution. We therefore quantified the extent of the match between the LLM and human prediction distributions using the Kullback–Leibler (KL) divergence between the two. Importantly, this measure is directional and reflects how well the LLM distribution can be used in place of the human distribution. Across all tested models, LLM distributions were most similar to human prediction distributions of written language (Fig. 3*C*), a difference that increased with greater sensory context (*SI Appendix*, Fig. S8; $\beta_{AV>A>W} = -0.23$, $SE = 0.01$, $t(24162) = -23.29$, $P < 0.001$). We consistently observed that LLMs better aligned with human written distributions using a second distance metric (Earth mover's distance: $\beta_{AV>A>W} = -2.2e^{-4}$, $SE = 1.6e{-}5$, $t(24162) = -13.26$, $P < 0.001$), demonstrating that these findings were not driven by the truncation of LLM distributions necessary to calculate KL divergence. Thus, further supporting our hypothesis, LLMs trained only on text data are better aligned with the behavior of humans who were deprived of in-the-moment sensory cues when predicting language.

Prosodic Information Improves Humans' Predictions of Language. Our findings thus far establish a clear rank order in performance: humans with access to sensory information outperform humans in the written condition, who in turn outperform (but are better matched to) LLMs. What information is available to participants with access to sensory context, but not for those deprived of this context? One particularly important feature of language is prosody, an auditory property that captures the intonation and stress given to words (or phonemes) and is closely tied to linguistic structure (e.g., syntax, pragmatics). Importantly, while prosody is interconnected with and inseparable from linguistic structure, the prosodic information itself is carried within the auditory signal. In our next set of analyses, we directly tested the hypothesis that linguistic information cued by sensory context—namely, prosody—informs human predictions of language and contributes to the divergence between humans and LLMs.

We tested whether humans are indeed leveraging prosodic cues to improve their predictions of upcoming words while processing audiovisual and audio-only language. To determine whether prosodic cues accounted for the observed differences between humans with and without access to sensory information (audiovisual/audio-only and written conditions, respectively), as well as differences between humans and LLMs, we quantified the prosodic prominence of each word using an established composite measure of auditory spectral information (*Materials and Methods*). While this measurement allows prosodic prominence and prosodic boundaries to be extracted from the same signal, we focus our analyses on prosodic prominence as it reflects the importance a speaker gives to words within the same utterance (37). We then averaged the prosodic prominence values of the five words before a target word to measure the prosody of the segment preceding a prediction. In other words, we investigated whether, in a coarse-grained sense, having more prosody available in the segment immediately preceding a target word boosted prediction performance.

Within human participants, the majority of the observed differences between the audiovisual/audio-only conditions and the written condition were accounted for by an interaction with prosody: the more prosody in the speech segment leading up to a target word, the larger the boost for humans with sensory context in terms of both accuracy (*binary accuracy*: $\beta_{AV>A>W} = -2.02$, $OR = 0.13$, $z = -3.59$; *continuous accuracy*: $\beta_{AV>A>W} = -0.10$, $t(2008) = -3.70$, $P < 0.001$) and consensus (Fig. 3 *D*, *Left*; *entropy*: quadratic term, $\beta_{AV>A>W} = 0.12$, $t(2008) = 2.37$, $P = 0.018$; *predictability*: quadratic term, $\beta_{AV>A>W} = -0.07$, $t(2008) = -2.18$, $P = 0.029$). In turn, the presence of prosodic information was also related to the observed divergences between humans processing sensory-informed language and LLMs. Specifically, differences in continuous accuracy between humans and LLMs increased with greater prosody for both audiovisual and audio-only language, but decreased for written language (Fig. 3 *D*, *Right*; $\beta = -0.03$, $t(24093) = -4.94$, $P < 0.001$). While it is unclear how much of the advantage conferred by prosody is driven by the sensory signal of prosody itself, or various types of linguistic information cued by that signal (e.g., syntactic structure, pragmatics), results nevertheless demonstrate the importance of prosody as a fundamental linguistic signal conveyed through the auditory modality. Together, these results suggest that prosodic information—a subset of information provided via the auditory signal—supports language processing to partially drive the observed boost in performance for humans with access to sensory context.

Sensory Context Improves Language Acquisition and Prediction in LLMs. A final prediction that arises from our claim—that sensory context is fundamentally important for human-like language behavior—is that exposing LLMs to varying amounts of sensory context during language acquisition (model training) will at least partially "rescue" their performance relative to humans in terms of both prediction accuracy and uncertainty. To test this, we trained four LLMs under different language acquisition conditions. The first model (*sensory deprived*; SD) was trained on only written text and represents a "control" case similar in spirit to the larger LLMs (e.g., *GPT-2*, *Llama2*) tested in our previous analyses. The second model (*prosody access*; PA) simulates a learner that benefits specifically from prosodic cues during language acquisition. The third model (*auditory context*; AC) reflects a learner that can leverage the full auditory signal (e.g., articulations, speaker, and environmental characteristics) to inform predictions of language. The final model (*audiovisual context*; AVC) represents a learner that integrates auditory and visual information (e.g., facial expressions, gestures) to learn a holistic representation of language. We note that these "baby" LLMs are trained on much less data than the larger, publicly available LLMs evaluated in the prior sections, and therefore the translation of results to industry-grade LLMs, which differ in architecture, scale, and emergent properties, is not guaranteed. Still, the relative performance between these smaller models provides a proof-of-principle for the impact of sensory context on language acquisition. We used these four models in a comparative

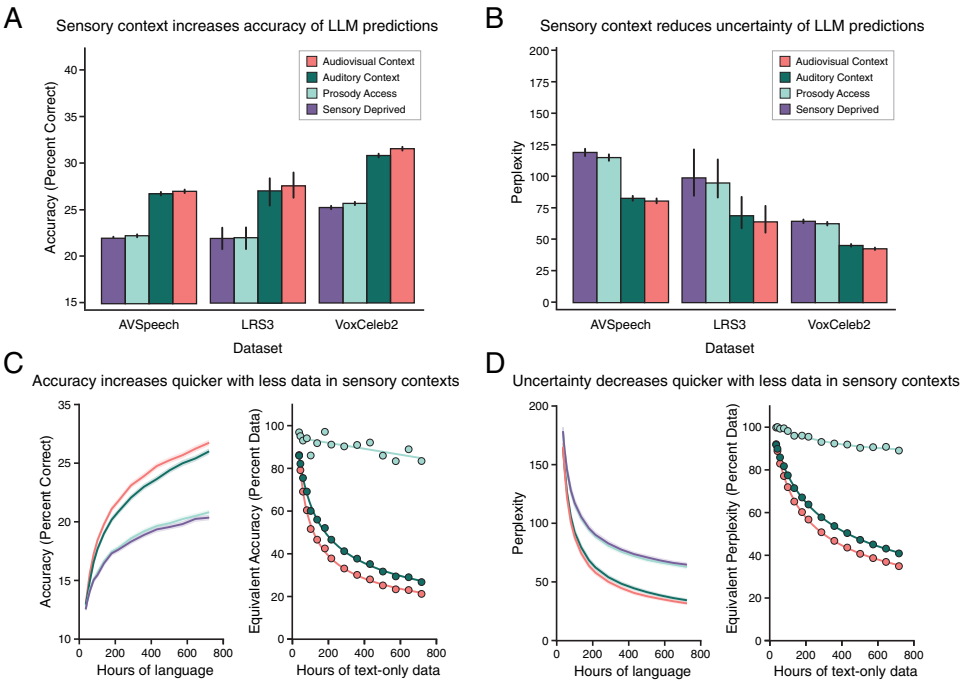

Fig. 4. Access to sensory information during training improves LLM performance. We trained and evaluated four different LLMs to understand how increasing amounts of sensory information (*sensory deprived*: text-only; *prosody access*: prosody & language; *auditory context*: audio & language; *audiovisual context*: audiovisual features & language) facilitates language processing and acquisition. Across three datasets, integrating sensory information within model representations of language (*A*) increased the accuracy and (*B*) decreased the uncertainty of LLM predictions. (*C* and *D*) Sensory information enhanced the rate of learning within LLMs (*Left* panels), such that models with access to increasingly rich sensory information required less data for equivalent performance to the text-only model (*sensory deprived*, *Right* panels). In all panels, error bars or confidence bands indicate 95% CI of the mean.

case study to determine whether access to sensory information can enhance language processing within LLMs similar to what is observed in humans.

We trained and evaluated these models within three open-access audiovisual language datasets that contained between 400 and 1,000 h of paired multimodal and written language. Across all three datasets, access to increasingly multimodal information—from just prosodic to the full audiovisual signal—improved the accuracy (Fig. 4*A*; $\beta_{AVC>AC>PA>SD} = -0.05$, $SE = 3e^{-4}$, $t(6554) = -163.58$, $P < 0.001$) and reduced the uncertainty (Fig. 4*B*; $\beta_{AVC>AC>PA>SD} = 26.41$, $SE = 0.2$, $t(6554) = 129.51$, $P < 0.001$) of model predictions. We also observed a nonlinear effect of sensory context on model performance (*accuracy* cubic term: $\beta_{AVC>AC>PA>SD} = 0.02$, $SE = 3e^{-4}$, $t(6554) = 66.27$; *perplexity* cubic term: $\beta_{AVC>AC>PA>SD} = -10.11$, $SE = 0.2$, $t(6554) = -49.58$; all $P < 0.001$). Across all models, the benefit of added sensory information scaled with the amount of sensory information provided to the model: the *audiovisual context* model outperformed the *auditory context* model (*accuracy*: $\beta_{AVC>AC} = 4.8e^{-3}$, $z = 11.46$; *perplexity*: $\beta_{AVC>AC} = -2.43$, $z = -8.42$; all $P < 0.001$), which in turn outperformed the *prosody access* model (*accuracy*: $\beta_{AC>PA} = 0.05$, $z = 114.59$; *perplexity*: $\beta_{AC>PA} = -25.37$, $z = -87.99$; all $P < 0.001$), which outperformed the *sensory deprived* model (*accuracy*: $\beta_{PA>SD} = 3.5e^{-3}$, $z = 8.17$; *perplexity*: $\beta_{PA>SD} = -3.11$, $z = -10.78$; all $P < 0.001$). Taken together with the nonlinear improvements, these results suggest that the auditory signal carries a majority of the informative variance for improving language predictions.

Thus far, we have evaluated the benefits of sensory information using two common performance metrics (accuracy and perplexity). We next aimed to validate that these performance gains were related to the language-congruent information contained within the sensory features, rather than simple increases in model dimensionality. Therefore, for each dataset, we trained "sensory control" models for each form of sensory context (i.e., prosody, auditory, audiovisual). Specifically, we randomized the dimensions of the sensory embeddings for each sample and token, effectively retaining the input dimensionality while destroying the representational structure of the sensory information. All sensory control models performed worse than the *sensory deprived* model in terms of both accuracy (*SI Appendix*, Fig. S9; $\beta_{SD>AVC/C} = 2.6e^{-3}$, $z = 8.15$; $\beta_{SD>AC/C} = 1.1e^{-3}$, $z = 3.26$; $\beta_{SD>PA/C} = 7.7e^{-3}$, $z = 23.96$; all $P < 0.001$) and perplexity (*SI Appendix*, Fig. S10; $\beta_{SD>AVC/C} = -1.98$, $z = -14.08$, $P < 0.001$; $\beta_{SD>AC/C} = -0.34$, $z = -2.41$, $P = 0.02$; $\beta_{SD>PA/C} = -7.08$, $z = -50.24$, $P < 0.001$), suggesting that the observed performance improvements were driven by the representational structure of the sensory signals.

We also aimed to estimate the added information contributed by each sensory modality. Given that directly quantifying the intrinsic, language-related information within these sensory signals is nontrivial, we focused on estimating the additional information provided by each sensory modality for next-word prediction. We operationalized information gain as conditional mutual information: the reduction in uncertainty the sensory modality provides for an upcoming word, above and beyond the word's linguistic context. We found that all sensory modalities contributed unique information for prediction not present within linguistic context (*SI Appendix*, Fig. S11; $\beta_{PA} = 0.03$, $t(4915) = 18.32$; $\beta_{AC} = 0.36$, $t(4915) = 205.76$; $\beta_{AVC} = 0.41$, $t(4915) = 229.91$; all $P < 0.001$). In contrast, the sensory control models exhibited no such increase in information when compared to the *sensory deprived* model—if anything, the sensory control models showed a slight decrease in information ($\beta_{PA/C} = -0.07$, $t(4915) = -62.04$, $P < 0.001$; $\beta_{AC/C} = -2.7e^{-3}$, $t(4915) = -2.37$, $P = 0.02$; $\beta_{AVC/C} = -0.02$, $t(4915) = -17.10$, $P < 0.001$). Critically, this added information scaled with the amount of sensory context provided to the model (quadratic term: $\beta_{AVC>AC>PA} = 0.12$, $SE = 1.3e^{-3}$, $t(4915) = 92.53$, $P < 0.001$), mirroring the nonlinear effects we previously observed in accuracy and perplexity. These findings suggest that the representational structure of sensory signals provide unique

information above and beyond linguistic context for predicting language.

In human development, sensory contexts are believed to improve the efficiency of language acquisition by reducing the total exposure needed to achieve a certain level of competence (19, 38–40). If sensory context plays a similar facilitating role for language acquisition within LLMs, we would expect that sensory-informed LLMs would show not only absolute gains in performance, but also relative gains such that less data is required to reach equivalent performance to the *sensory deprived* LLM. We therefore compared model performance after training our models on increasing amounts of data (15 to nearly 750 h). As expected, all models became more accurate (Fig. 4 *C*, *Left*; $\beta = 0.15$, $SE = 1.9e^{-4}$, $t(53126) = 792.74$, $P < 0.001$) and less uncertain (Fig. 4 *D*, *Left*; $\beta = -0.43$, $SE = 2.7e^{-4}$, $t(53126) = -1597.71$, $P < 0.001$) with increasing amounts of data. However, all models provided with sensory information (*audiovisual context*, *auditory context*, *prosody access*) exhibited faster performance gains than the *sensory deprived* model that scaled with the amount of training data provided to the model (*accuracy*: $\beta_{AVC>AC>PA>SD} = 0.05$, $SE = 3.9e^{-4}$, $t(53126) = 131.93$; *perplexity*: $\beta_{AVC>AC>PA>SD} = -0.10$, $SE = 5e^{-4}$, $t(51175) = -190.45$; all $P < 0.001$). Similar to before, these improvements in model performance were, in part, nonlinear, with the largest gains observed with access to the complete auditory signal (*accuracy* cubic term: $\beta_{AVC>AC>PA>SD} = -0.02$, $SE = 3.9e^{-4}$, $t(51175) = -51.77$; *perplexity* cubic term: $\beta_{AVC>AC>PA>SD} = 0.03$, $SE = 4.9e^{-4}$, $t(51167) = 60.28$; all $P < 0.001$).

Finally, we quantified how the *efficiency* of language acquisition changes with sensory information. Specifically, for a given number of hours of text-only data, we identified the amount of time required for the *prosody access*, *auditory context* and *audiovisual context* LLMs to exhibit equivalent performance (i.e., accuracy, uncertainty) to the *sensory deprived* model (*Materials and Methods*). We then fit a curve that described the relative efficiency of language acquisition for each sensory-informed model compared to the *sensory deprived* model. All multimodal models were well described by an exponential decay function for accuracy and a power law function for perplexity (all $R^2 > 0.99$), demonstrating that text-only data provides diminishing returns for language learning compared with multimodal data. Critically, we found that performance benefits scaled with the richness of sensory information (Fig. 4 *C* and *D*, *Right*; *accuracy*: $\beta_{AVC>AC>PA} = -0.30$, $SE = 7e^{-4}$, $t(39844) = -430.57$; *perplexity*: $\beta_{AVC>AC>PA} = -0.17$, $SE = 2.8e^{-4}$, $t(39844) = -622.41$; all $P < 0.001$). Furthermore, these benefits were nonlinear, again emphasizing the particular importance of the auditory signal for language prediction (*accuracy* quadratic term: $\beta_{AVC>AC>PA} = 0.05$, $SE = 7.2e^{-4}$, $t(39844) = 74.26$; *perplexity* quadratic term: $\beta_{AVC>AC>PA} = 0.05$, $SE = 2.9e^{-4}$, $t(39846) = 170.35$; all $P < 0.001$). Thus, multisensory context can play a similar role for LLMs as it does for humans: namely, it can improve language processing by reducing uncertainty and accelerating the rate of language acquisition.

Discussion

Language researchers have long believed that sensory context—the auditory and visual channels through which language is experienced—is critical for language processing and acquisition. Despite broad theoretical agreement, empirically quantifying this importance has been difficult, as it would be not only unethical but also impossible to teach humans language in a fully disembodied manner. Large language models (LLMs), which achieve seemingly human-level language proficiency via statistical learning from written text alone, capture a disembodied experience of language divergent from that of humans. Thus, these models present an opportunity to directly test the assumption that sensory context is critical for human-like language processing.

Here, we showed that sensory context critically supports language prediction for humans and structures language acquisition in large language models (LLMs). While depriving humans of sensory information by having them read (instead of watch or listen) reduced performance, even readers provided more accurate predictions than LLMs. Human predictions informed by sensory contexts were also less aligned with LLMs than those based on written language, driven in part by prosodic features conveyed via the auditory signal. Finally, we demonstrated proof-of-principle that—at least within smaller LLMs—access to multimodal information during training improves language predictions and accelerates language acquisition. Together, our findings underscore the central role of sensory context for human language prediction and highlight a systematic divergence between humans and LLMs.

Human language, like communication in many other species, is commonly expressed via vocalizations and gestures and thus inextricably tied to the auditory and visual domains (1, 2, 41). Exposure to auditory cues during development, both prenatally and in infancy, boosts the efficiency of language acquisition (42–46). When combined with visual information (e.g., facial expressions, mouth movements, gestures), these cues help disambiguate speech (6), direct attention (47), and scaffold human vocabulary acquisition (7, 8). In contrast, reading-related behaviors and brain areas (e.g., the visual word form area) emerge much later in development (48, 49). We suggest that our observed pattern of prediction accuracy—whereby humans with more sensory context outperformed humans reading, who in turn outperformed LLMs—arises from differences in both momentary (perceptual) access and lifetime (latent) exposure to sensory-grounded language. Specifically, humans watching or listening benefit from in-the-moment sensory context, humans reading perform at an intermediate level because lifetime sensory-grounded experience scaffolds linguistic representations even without real-time cues, and LLMs perform worst due to wholesale lack of any sensory context for language.

To support this idea, we partially "rescued" LLM performance by integrating sensory cues (i.e., prosody, auditory, and audiovisual features) with language during training, thereby improving predictions and accelerating acquisition. While our "baby LLMs" differ from larger, industry-grade models that often exhibit emergent properties as a result of scale [e.g., dataset size, architecture; (50, 51)], this demonstration nevertheless underscores the integral role of sensory context in language learning. While it is difficult to disentangle whether improved performance stems from sensory information per se or simply higher-dimensional input, we trained control models with shuffled sensory dimensions to show that structured, representational content drives most of these improvements. Furthermore, quantifying conditional mutual information confirmed that sensory modalities contribute unique predictive information beyond linguistic context alone. Our findings dovetail with prior work that suggests substantial shared information between linguistic content and prosody (52, 53), while demonstrating that the full sensory representation

contains even more useful and unique information for language prediction. Importantly, in real-world settings, these sensory signals are inseparable from linguistic content and provide this additional information "for free." Future work could explore these effects more mechanistically, for example by varying the type, timing, or integration of sensory cues. Our results align with the trend toward multimodal LLM training (54, 55), suggesting that integrating sensory signals meaningfully enhances prediction in artificial systems, as in human learners.

Historically, language has been studied as an information-transmission system in which less predictable utterances carry more information (56–58). Studies in this tradition often quantify how predictability—here, a property inversely related to the information carried by a given unit of interest (e.g., characters, words)—changes as a result of context (e.g., surrounding letters, preceding words). For humans, less predictable words evoke stronger neural responses (59, 60) and slower reading times (61, 62), thought to reflect linguistic processing effort (63) and individual world models (5). Past work has shown that LLM surprisal (log-probabilities of predictions) predicts human reading times (64) and predictions of written text (29, 30), suggesting a tight link between LLMs' representations of language and predictability in the information-theoretic tradition. However, most past studies rely on single-sentence "Cloze" tasks [i.e., fill-in-the-blank sentences, (65)] presented as written text and compare human and LLM predictions under less naturalistic conditions—i.e., by oversampling common words (e.g., stop words) or having participants predict every item (29), including nonwords like emojis and symbols (31). In contrast, we used real-world narratives and sampled only content words, thereby focusing on words that carry much of the narrative's meaning while minimizing interruptions. The picture that emerged from our approach was more complicated: while, on a word-by-word basis, LLM predictability was related—albeit weakly—to human predictability while reading, this relationship disappeared or even perhaps flipped direction when humans were listening or watching. This divergence from past work further demonstrates that humans can and do integrate extralinguistic (i.e., sensory) cues to inform predictions about upcoming semantic content, and further underscores the need to study human language processing under more naturalistic conditions.

While our findings establish sensory context as indispensable for human-like language, sensory signals are but one of many information sources that shape how humans use and extract meaning from language—and likely one of many factors driving divergences between humans and LLMs (12, 66). More broadly, human language differs in form and functions from LLMs: humans develop and use language in an embodied manner to interact with other agents and act on the world (67, 68). In our study, participants passively watched rehearsed stories told by single narrators, meaning that the visual cues present in these videos (e.g., gestures, facial expressions) are only a fraction of the visual information available in naturalistic settings, where the semantic structure of the environment (69, 70) and joint attention between speakers and listeners (71, 72) further shape how humans understand and predict the world. Accordingly, if anything, our findings likely underestimate the role of the visual modality in informing and contextualizing language predictions. Developing a more nuanced understanding of how real-world, embodied contexts inform language will be critical to understanding behavioral and representational divergences of humans and LLMs.

Importantly, these divergences do not invalidate LLMs as models of language. Rather, identifying differences between humans and LLMs provides an exciting opportunity to improve models and reveal the ingredients central to human language. Here, we evaluated behavioral alignment on the primary task used for LLM training: next-word prediction. While next-word prediction captures only a subset of human language behaviors (12, 73), it provides a strong self-supervised language objective that has been shown to relate to the alignment between LLMs and human neural representations (29, 74). Our findings hint at potential divergences between human and LLM representations rooted in the modality of language processing and behavior under investigation [i.e., comprehension vs. production vs. prediction; (17, 23, 75, 76)]. It will be important to evaluate how sensory context impacts alignment with other language behaviors and informs other language contexts beyond oral monologues, such as natural conversations where sensory signals are likely to be more informative for processing and prediction (77).

Language is inseparable from the world in which it is learned and used: sensory, social, and environmental contexts are not mere niceties but rather essential features of language itself. While human language has traditionally been studied with these contexts in mind, LLMs, as the first viable model of language outside humans, provide an opportunity to directly quantify their impact on language behavior and representation. We showed that access to sensory context produces systematic divergences in language processing between humans and text-only LLMs, and that exposing LLMs to multimodal information during training facilitates rapid language acquisition akin to that observed in human children. More generally, our findings demonstrate the potential of comparative approaches of LLM and human behavior to evaluate theories of how we, as humans, acquire and use language.

Materials and Methods

Participants. We recruited human adults ($N = 1,500$; mean age = 38.15 ± 11.97; 743 female, 727 male, 10 nonbinary, 6 other, 14 do not wish to report) from an online crowd-sourced research platform (Prolific). All participants provided informed consent, confirmed no prior experience listening to *The Moth* podcast, and indicated English as their first language. Participants were compensated above the living wage ($12 per hour), and all procedures within the study were approved by the Committee for the Protection of Human Subjects at Dartmouth College.

Behavioral Experiment.

Stimuli. We used three stories recorded for *The Moth* podcast ("*I Knew You Were Black*:" 13:20 min; "*Where There's Smoke*:" 10:02 min; "*How to Draw a Nekkid Man*:" 12:09 min). All stories are first-person personal narratives about a past event within each speaker's life. Specifically, "*Black*" focuses on race, activism, and identity, "*Smoke*" on confronting addiction and friendships, and "*Draw*" on divorce and transitioning between jobs. These narratives occupy a middle ground between free-form conversation and books, as each narrative is drafted in advance, edited, and practiced yet intended to be performed aloud in front of a live audience.

For all stories, forced-aligned transcripts and audio recordings were initially sourced from two previously published fMRI datasets (78, 79). We also obtained the original video recordings of these narratives from *The Moth* podcast's YouTube channel. We manually validated each video's audio by checking the overlap against the original audio used for the transcript forced-alignment. We next performed manual adjustments to the transcripts in two ways to mitigate differences in presentation between the three conditions (audiovisual, audio-only, written). First, we added punctuation to the transcripts to ensure that the written text read similarly to how the original spoken version sounded. Second,

and crucially, we manually adjusted word-level boundaries using Praat (80) to ensure that participants with access to auditory information (audiovisual and audio-only conditions) did not receive an advantage due to "leakage" of auditory information (e.g., hearing the first phoneme of the next, to-be-predicted word).

Controlling for auditory leakage. Even with careful manual adjustment of word timings, it is sometimes difficult to pause the auditory stream in a way that gives complete access to word *N-1* without including "leakage" from the target word *N*. It is therefore possible that participants with access to auditory information (audiovisual and audio-only conditions) may have been unintentionally advantaged by hearing information from the to-be-predicted word (i.e., the first phoneme). To rule out the possibility that this confound contributed significantly to our observed pattern of results, we performed a post hoc analysis to remove moments of potential "leakage" that may have driven the observed differences in human behavior across modalities.

To identify moments of leakage, we first limited our analyses to errors (incorrect predictions) and tested whether participants in the audiovisual and audio-only language conditions were more likely to predict a word with the same first phoneme as the ground-truth word. Specifically, for each target word predicted in our experiment, we compared the first phoneme of each wrong answer to the first phoneme of the ground-truth word, separately counting the number of correct and incorrect predictions for the audiovisual, audio-only, and written language conditions. We then performed Barnard's exact test on each 2×2 contingency table (*audiovisual/audio-only vs. written*, *correct vs. incorrect*) to quantify moments where participants in the audiovisual or audio-only language conditions were more likely to predict a word that was incorrect but nonetheless started with the correct phoneme than participants in the written language condition.

Each significant contingency table represents a moment (i.e., target word) where participants with access to sensory context may have been unintentionally advantaged. We therefore took the intersection of significant values between the audiovisual and audio-only conditions to determine these potentially advantaged moments. Using the outcome of these contingency tables, we excluded any trial that exhibited evidence of possible leakage and reran comparisons between humans in the audiovisual, audio-only, and written conditions to ensure that all reported effects still hold after controlling for this possible confound. Critically, this exclusion criterion provides a conservative test, as other factors beyond true "leakage" such as phoneme transitions and prosody likely also help inform the first phoneme of the upcoming word in audio conditions. We also note that this analysis is not intended to remove the effects of longer-range, coarticulatory cues; rather than confounds, we consider these to be features of interest that may help boost predictive performance for humans with access to the auditory stream (*Prosody Annotations*).

Selecting candidate words for prediction. One of our goals was to assess human predictions of language at moments where LLMs either succeeded or failed at performing the same task (next-word prediction). To this end, we imposed a few constraints on the words presented to participants.

First, we focused our experiment on content words–removing stop-words (e.g., "the," "a," "an") and named entities (e.g., people, locations)–as these words are most often subject to the contextual dependencies within stories. Second, we selectively sampled these content words based on 1) the average continuous accuracy of five-shot model predictions (*Metrics*) and 2) the entropy of these predictions. Specifically, we identified an LLM representative of the population of tested LLMs (*GPT-2 XL*) based on the similarity of continuous accuracy and entropy (evaluated at content words) across models. We then divided the candidate content words into four quadrants based on moments where *GPT-2 XL* exhibited high/low continuous accuracy and high/low entropy. We sampled words proportionally from each quadrant to match the native distribution of words across quadrants. This resulted in the following proportion of words sampled from the accuracy-entropy quadrants: 17% high–high, 33% high–low, 34% low–high, 17% low–low. Last, we ensured that the selected words were spaced apart by a minimum of 10 words. This final constraint ensured that our participants were able to experience the story as naturally as possible with minimal interruptions (30).

For each sampled word, we estimated the human prediction distribution by collecting 50 predictions of that word, one from each of 50 distinct participants. To do this, we divided the "candidate words" from the procedure above into separate presentation orders that satisfied the outlined constraints. This resulted in three different presentation orders for "*Where There's Smoke*" and "*How to Draw a Nekkid Man*" ($N = 150$ participants per condition), and four presentations orders for "*I Knew You Were Black*" ($N = 200$ participants per condition). This procedure resulted in sampling approximately 12.5% of words from each story (total of 672 words) and limited interruptions for each participant to less than 4% of the story.

Experimental design. We divided participants into three groups ($N = 500$ each) that either 1) watched the speaker perform the story (audiovisual language), 2) listened to the story without seeing the speaker or the transcript (audio-only language), or 3) read the story word-by-word without hearing the accompanying audio (written language). Participants in all conditions provided responses to the same words, and in the written condition, the sampled words were visually presented at the rate at which they were spoken in the original version to mitigate timing differences between conditions.

Upon enrollment, participants provided informed consent, confirmed no prior experience listening to *The Moth* podcast, and read instructions that explained the task. Specifically, we instructed participants to watch, listen to, or read the story and, whenever the story was paused, provide the word that they believed was most likely to come next. Participants in both conditions gave predictions by typing them into a text box. Following these instructions, participants completed a short practice trial that mimicked the structure of the experiment before starting the main experiment. After the experiment, we asked participants to complete a standard demographics survey and provide any comments about their experience.

Large Language Models. We compared human responses to 12 openly available models (81) sourced from the *transformers* library (82). These LLMs included both causal language (number of parameters: *GPT-2*, 137*M*; *GPT-2 XL*, 1.61*B*; *GPT-Neo-X*, 1.37*B*; *Bloom*, 559*M*; *Mistral*, 7.24*B*; *Llama2*, 6.74*B*; *Gemma3*, 1*B*; *Qwen3*, 32*B*; *Llama3.1*, 70.6*B*) and masked language (number of parameters: *RoBERTa*, 125*M*; *Electra*, 110*M*; *XLM-ProphetNet*, 391*M*) training objectives. For each LLM, we extracted the one-shot and five-shot predictions (i.e., top-1 prediction and top-5 predictions respectively) for each word in the stimulus using the prior 25 words as context. Given the importance of evaluating humans and LLMs under comparable conditions (83), we selected this context window to provide a coarsely matched equivalent to human temporal windows of language processing (84, 85). Similar to previous studies (29, 86), we calculated the entropy of the prediction distribution for each word as a measure of prediction uncertainty.

Data Preprocessing. We performed a series of preprocessing steps to standardize responses for human participants. First, we manually checked the spelling of each response and corrected any typos. This resulted in spell-correcting 4.5% of audiovisual responses, 3.5% of audio responses, and 3.4% of written responses. On the occasion that participants responded with multiple words, we selected the first word within the response that was not a repetition of the previous words, a stop word, or a named entity. In addition, we standardized across slang words (e.g., "yah" $\rightarrow$ "yes," "fag" $\rightarrow$ "cigarette") to remove unintentional variability in the responses. This resulted in shortening 3.4% of audiovisual responses, 3.4% of audio responses, and 0.2% of written responses. Last, we lemmatized both human and LLM responses, as well as the ground-truth word, to account for differences in responses based on inflected form. All analyses are conducted on the lemmatized forms of responses and ground-truth words.

Metrics. Below, we define the metrics used to evaluate the performance of human and LLMs on next-word prediction, as well as the correspondence between human and LLM prediction distributions.

Accuracy. We calculated the *binary accuracy* (exact match; 0 or 1) of human and LLM one-shot predictions. Given that binary measures provide a limited picture of LLM task performance (87), we also calculated a continuous measure of accuracy. Specifically, we defined *continuous accuracy* as the cosine similarity between the one-shot prediction and the ground-truth word. To avoid biasing continuous accuracy for LLMs based on different transformer architectures, we

used a word-embedding model to estimate the semantic similarity of predictions for both humans and LLMs. We chose *fasttext* (32) as the word model given its ability to provide vectors for out-of-vocabulary words.

Predictability. In line with prior work (29, 30), we quantified human *predictability* as the percentage of human participants identifying the correct word. We quantified LLM predictability as the probability assigned to the ground-truth word. Also in line with prior work (88), we evaluated the correspondence between human and LLM predictability scores within a logarithmic space to account for the distribution of the data.

Uncertainty. We quantified the uncertainty of predictions as the Shannon's entropy of prediction distributions. Specifically, for human participants in each condition (audiovisual/audio-only/written), we counted the number of occurrences of a given prediction at each critical word and converted these counts to probabilities (i.e., such that the prediction distribution summed to one). For LLMs, we extracted the logits for each critical word and converted them to probabilities using a softmax function. We verified that our findings were robust to the choice of entropy metric by demonstrating similar results using two alternative metrics that include bias correction for small sample sizes and unobserved data: nonparametric Chao-Shen estimation (89) and Bayesian Nemenman-Shafee-Bialek (NSB) estimation (90).

Kullback–Leibler (KL) divergence. We used Kullback–Leibler (KL) divergence to quantify the alignment of LLM prediction distributions with the prediction distributions generated by our populations of human participants. Importantly, KL divergence is not symmetric and therefore allows us to quantify how well LLM prediction distributions reflected the human prediction distributions. While LLM prediction distributions are computed over all words within its vocabulary, human prediction distributions are limited to the set of words predicted across all participants. To compare human and LLM distributions, we therefore limited the LLM prediction distributions to the unique words predicted by human participants (in all conditions) and normalized the truncated distribution. We then calculated the KL divergence between the LLM distributions and human distributions for each of the three conditions (audiovisual/audio-only/written).

Prosody Annotations. We used the Wavelet Prosody Analyzer Toolkit (37) to extract word-level prosodic prominence labels–the relative auditory salience of a linguistic unit (here, words)–for each stimulus. The toolkit uses a continuous wavelet transform (CWT) to estimate prosodic prominences as a composite signal of three important features within the auditory spectrogram. Specifically, the CWT combines the fundamental frequency (f_0), energy, and duration of a linguistic unit and hierarchically contextualizes these signals within the surrounding auditory context. Following prior work, we parameterized the CWT model using a weighted multiplication of prosodic features (weights: f_0, 1.0; energy, 0.5; duration, 1.0). This approach yields two values–prosodic prominence and prosodic boundaries–that are estimated from the same underlying signal. In this work, we focused on prosodic prominence given that it often signals the importance a speaker places on each word within a given utterance (37). To this end, we calculated the average prosodic prominence of the five words preceding a prediction. While prosodic boundaries provide additional signals that may serve to direct the listener's attention to particular intonational phrases, we focused on prosodic prominence given its well-studied role in conveying the speaker's intentions. Given the relationship between the two, as well as the fact that our averaging method was naive to prosodic boundaries, our results likely underestimate the role of prosody in informing predictions of language.

Careful Whisper Details. *Careful Whisper* is a multimodal language model that leverages yoked sensory and language tokens to perform next-word prediction. We developed *Careful Whisper* with the goal of testing hypotheses about language development–namely, whether access to sensory information during language learning improves model performance.

Model architecture. The architecture of *Careful Whisper* is inspired by the speech recognition model *Whisper* (91) and uses a causal language modeling objective. We selected this architecture as the cross attention mechanism allows interpretability of the interactions between sensory and linguistic information. In contrast to *Whisper*, we implemented a causal mask within the cross attention layers–effectively limiting the model to leverage past and present sensory information to predict upcoming language–to be commensurate with the task we gave human participants.

We evaluated four variants of our model that represented different conditions of language acquisition. Our first model–*sensory deprived (SD)*–used the architecture of *GPT-2* to evaluate language acquisition without access to any sensory information. Our second model–*prosody access (PA)*–used the *Careful Whisper* architecture to evaluate how prosodic information in particular facilitates language acquisition. Our third model–*auditory context (AC)*–used the *Careful Whisper* architecture to evaluate how auditory information more generally facilitates language acquisition. Our final model–*audiovisual context (AVC)*–used the *Careful Whisper* architecture to evaluate how integrated audiovisual information promotes language acquisition.

All models were instantiated with the same randomly initialized weights. All models used an embedding dimensionality of 1024*D*, absolute positional embeddings, and 12 residual attention blocks–where each multihead attention mechanism had 16 attention heads. Both forms of *Careful Whisper* included positional embeddings for the context (e.g., *prosody*, *auditory*, *audiovisual*) and an embedding dropout rate of 0.1.

Datasets. We leveraged three openly available audiovisual speech datasets to train and evaluate our models (introduced from smallest to largest). In this section, we provide the dataset sizes prior to preprocessing (see next section for preprocessing steps).

LRS3 (92) is an audiovisual speaker identification dataset with videos sourced from TED and TEDx videos. The dataset contains over 400 h of audiovisual speech recordings from over 5,000 speakers. All videos in this dataset were in English, and transcripts were provided for each video.

AVSpeech (93) is a large-scale audiovisual speech dataset sourced from TED talks and how-to-videos from YouTube. The dataset contains a total of 4,700 h of video clips of diverse, unconstrained audiovisual speech samples across multiple languages and speakers.

VoxCeleb2 (94) is a large-scale audiovisual speaker identification dataset. The dataset contains over 2,442 h of audiovisual data from 6,112 speakers (all celebrities). The dataset comprises YouTube videos featuring a wide range of speakers with varying linguistic backgrounds, facial expressions, and recording conditions.

Preprocessing. Given that the *AVSpeech* and *VoxCeleb2* datasets were multilingual, we first identified the language of each video using an open-source language classifier (95). We filtered videos based on two criteria: 1) the majority of segments being in English (>75%) and 2) all clips having high-confidence language classifications (>95%). Transcripts were then extracted using *Whisper*. For all datasets, audio was resampled to 16,000 Hz to match *wav2vec2* (c.f., *Model architecture*). We submitted each audio-transcript pair to the Montreal Forced Aligner (96) to obtain word-level alignment times, and generated language tokens using the *GPT-2* tokenizer (vocabulary size: 50,257).

Prosody tokens were created by applying the Wavelet Prosody Toolkit (37) to word-level transcripts, yielding a 1*D* prominence value per word. We then used a static linear projection to map this 1*D* value to a 1024*D* vector that matched the embedding dimensionality of the model. Therefore, each audio sample was also described with a $S \times D$ prosodic features matrix, where each prosody token corresponds to a given language token.

Auditory tokens were constructed by segmenting each audio sample into word-level clips based on alignment times. For each word, we used *wav2vec2* (97) to extract auditory features from the word's corresponding audio clip. Importantly, segmenting each word before extracting embeddings prevents the model from integrating information across the speech sample, which would allow unwanted leakage of future information into predictions. We estimated a single auditory representation for the word by performing mean pooling over all audio embeddings for the word segment, creating a single 1024*D* representation of the word. Thus, we obtained a $S \times D$ auditory features matrix, where each auditory token corresponds to a given language token.

Audiovisual tokens were created in two steps. First, visual tokens were extracted by segmenting each video sample into word-level clips based on alignment times. We then extracted visual features using *data2vec* (pretrained on ImageNet-1K), followed by mean pooling to obtain 1024*D* representations per word. Second, auditory and visual tokens were fused using a two-layer perceptron that compressed concatenated 2048*D* inputs into 1024*D* representations,

trained with MSE loss to reconstruct the original auditory and visual vectors from the compressed, audiovisual vector. This resulted in an $S \times D$ audiovisual feature matrix aligned with the language tokens.

All sequences used for training and evaluation contained between 4 and 128 words. Dataset sizes following preprocessing are reported in *SI Appendix*, Table S1.

Training. We trained each of our four models (*sensory deprived*, *prosody access*, *auditory context*, *audiovisual context*) separately on each of the three datasets. All models were instantiated with the same seed, ensuring that weight initialization and randomization of training data happened consistently across training runs. We used a batch size of 32 samples and a gradient accumulation of 4 steps, resulting in an effective batch size of 128 samples.

We trained models for a maximum of 15 epochs using the AdamW optimizer (98) with a learning rate of $2.5e^{-5}$ and weight decay of 0.1, using a scheduler that reduced the learning rate if a plateau of learning was observed for 2 consecutive epochs. We additionally employed an early stopping criterion based on the validation accuracy with patience of 3 epochs.

Evaluation. For each dataset, we evaluated the trained models on a held-out test set using two common metrics for causal language modeling. First, we scored the *one-shot accuracy* for each word in the sequence to determine the correctness of each prediction. Second, we evaluated the *perplexity* of each sequence–the exponentiated loss–to understand the model prediction uncertainty. We used a batch size of 32 samples to estimate each of these metrics

Control models. For each dataset, we created a "sensory control" model for each form of sensory information (i.e., prosody, audio, audiovisual). Specifically, we held the linguistic input constant (i.e., the order of tokens) while randomizing the dimensions of the sensory modality independently for each sample and token. This manipulation holds the input dimensionality and temporal correspondence of tokens constant while destroying the representational structure of the sensory information and its congruence with the linguistic information (tokens). All other parameters of the training were held constant.

Assessing equivalence points across models. We last aimed to understand whether language acquisition is more efficient within multimodal environments. To this end, we retrained each of our four models (*sensory deprived*, *prosody access*, *auditory context*, *audiovisual context*) on subsets of our largest dataset (*VoxCeleb2*) that totaled around 750 h. Specifically, we sampled the dataset in both a) logarithmically spaced intervals spanning 2% to 25% and b) 10% intervals spanning 30% to 100%. Importantly, data within each subset (A_1) appeared in the larger intervals ($A_1 \subseteq A_2 ... \subseteq A_{20}$), ensuring consistency across model runs in the data used for training. We then evaluated model performance–accuracy and perplexity–for all models (total 72 models) on the unseen test dataset.

At each subset, we compared the one-shot accuracy and perplexity (a measure of uncertainty about a sequence) of model predictions across our four model types. For each of the multimodal models (*prosody access*, *auditory context*, *audiovisual context*), we used these results to fit a curve that predicted the training subset (e.g., number of hours of language data) from the model's average accuracy or perplexity. We used a robust regression approach (Huber loss) to stabilize curve fits. For all models, accuracy curves were best fit by an exponential decay function and perplexity curves were best fit a power law function (all $R^2 > 0.99$), demonstrating that models exhibited diminishing returns on performance with larger amounts of data. We then used these curves to estimate the magnitude of performance benefits provided by multimodal information over the *sensory deprived* model. Specifically, for each training subset, we supplied the average accuracy of the *sensory deprived* model and predicted the amount of data required to achieve equivalent performance. We defined the equivalence of model performance as the relative amount of data needed for a multimodal model to achieve similar performance to the *sensory deprived* model (e.g., $\frac{MM}{SD}$).

Statistical Analysis. Inferential statistics were performed in *R* using generalized linear mixed-effects models, where each model was computed with the maximal random effects structure permitted by our data (99).

For analyses of participant-level accuracy, we included random effects of subject, story, and predicted word. All analyses of population-level human behavior (e.g., audiovisual, audio-only, and written; humans vs. LLMs) included random effects of story and predicted word. In analyses that compared humans to LLMs, we also included a random effect of model (e.g., *GPT-2*) to control for differences in behavior across tested LLMs. All analyses of binary accuracy used logistic mixed-effects models.

Across analyses, we used linear contrasts to compare human performance across modalities (*Audiovisual* > *Audio* > *Written*) and to compare human performance with LLM performance (*Audiovisual* > *Audio* > *Written* > *CLM* > *MLM*). While our statistical analyses used data across the three stories, we visualize each story separately to demonstrate the consistency of the observed effects across different stimuli.

We used linear mixed-effects models to conduct inferential statistics on the results of multimodal modeling. We compared model performance using a linear contrast (*Audiovisual Context* > *Auditory Context* > *Prosody Access* > *Sensory Deprived*). We included random effects of batch number and dataset when analyzing the full model performance (e.g., using all data from a given dataset). When comparing models based on varying amount of training data, we predicted performance (accuracy/perplexity) from the log-transformed amount of data.

All *P* values result from two-sided tests and are corrected for multiple comparisons using the Benjamini–Hochberg method for *FDR* correction.

Data, Materials, and Software Availability. Stimuli and behavioral data (raw, cleaned, and analyzed) are available at the following link: https://osf.io/x4akz/. The behavioral experiment was programmed in HTML, CSS, and Javascript with JSPsych (100). Preprocessing, analyses, and LLM training were conducted in Python. Statistical analysis was conducted in R. All scripts are publicly available at: github.com/thefinnlab/pfka.git.

ACKNOWLEDGMENTS. T.L.B. and E.S.F. were supported by the Neukom Institute for Computational Science at Dartmouth College. We would like to thank Sushmita Sadhukha, Eshin Jolly, Erica Busch, Clara Sava-Segal, Kathryn O'Nell, and Soroush Vosoughi for feedback on the project.

1. W. T. Fitch, *The Evolution of Language* (Cambridge University Press, ed. 1, 2010).
2. M. D. Hauser, N. Chomsky, W. T. Fitch, The faculty of language: What is it, who has it, and how did it evolve? *Science* **298**, 1569–1579 (2002).
3. P. K. Kuhl, Early language acquisition: Cracking the speech code. *Nat. Rev. Neurosci.* **5**, 831–843 (2004).
4. S. A. Gelman, Learning from others: Children's construction of concepts. *Annu. Rev. Psychol.* **60**, 115–140 (2009).
5. P. Hagoort, L. Hald, M. Bastiaansen, K. M. Petersson, Integration of word meaning and world knowledge in language comprehension. *Science* **304**, 438–441 (2004).
6. J. H. McDermott, The cocktail party problem. *Curr. Biol.* **19**, R1024–R1027 (2009).
7. P. K. Kuhl, A. N. Meltzoff, The bimodal perception of speech in infancy. *Sci. N. Y.* **218**, 1138–1141 (1982).
8. J. Gillette, H. Gleitman, L. Gleitman, A. Lederer, Human simulations of vocabulary learning. *Cognition* **73**, 135–176 (1999).
9. C. D. Manning, K. Clark, J. Hewitt, U. Khandelwal, O. Levy, Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 30046–30054 (2020).
10. T. B. Brown *et al.*, Language models are few-shot learners. arXiv [Preprint] (2020). https://arxiv.org/abs/2005.14165 (Accessed 14 September 2022).
11. T. Linzen, M. Baroni, Syntactic structure from deep learning. *Annu. Rev. Linguist.* **7**, 195–212 (2021).
12. K. Mahowald *et al.*, Dissociating language and thought in large language models. *Trends Cogn. Sci.* **28**, 517–540 (2024).
13. J. W. A. Strachan *et al.*, Testing theory of mind in large language models and humans. *Nat. Hum. Behav.* **8**, 1285–1295 (2024).
14. M. Kosinski, Evaluating large language models in theory of mind tasks. *Proc. Natl. Acad. Sci. U.S.A.* **121**, e2405460121 (2024).
15. T. Webb, K. J. Holyoak, H. Lu, Emergent analogical reasoning in large language models. *Nat. Hum. Behav.* **7**, 1526–1541 (2023).
16. L. Ruis *et al.*, Procedural knowledge in pretraining drives reasoning in large language models. arXiv [Preprint] (2024). https://arxiv.org/abs/2411.12580 (Accessed 13 December 2024).
17. E. Fedorenko, S. T. Piantadosi, E. A. F. Gibson, Language is primarily a tool for communication rather than thought. *Nature* **630**, 575–586 (2024).
18. M. Shanahan, Talking about large language models. arXiv [Preprint] (2023).https://arxiv.org/abs/2212.03551 (Accessed 11 December 2024).

19. M. C. Frank, Bridging the data gap between children and large language models. *Trends Cogn. Sci.* **27**, 990–992 (2023).
20. N. Chomsky, Rules and representations. *Behav. Brain Sci.* **3**, 1–15 (1980).
21. M. C. Frank, M. Braginsky, D. Yurovsky, V. A. Marchman, *Variability and Consistency in Early Language Learning: The Wordbank Project* (MIT Press, Cambridge, Massachusetts London, England, 2021).
22. A. Warstadt, S. R. Bowman, "What artificial neural networks can tell us about human language acquisition" in *Algebraic Structures in Natural Language*, S. Lappin, J.-P. Bernardy, Eds. (CRC Press, 2022), pp. 17–60.
23. I. Sucholutsky *et al*., Getting aligned on representational alignment. arXiv [Preprint] (2023). https://arxiv.org/abs/2310.13018 (Accessed 27 October 2023).
24. M. Mitchell, D. C. Krakauer, The debate over understanding in AI's large language models. *Proc. Natl. Acad. Sci. U.S.A.* **120**, e2215907120 (2023).
25. C. Cuskley, R. Woods, M. Flaherty, The limitations of large language models for understanding human language and cognition. *Open Mind* **8**, 1058–1083 (2024).
26. E. Pavlick, Symbols and grounding in large language models. *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.* **381**, 20220041 (2023).
27. B. M. Lake, G. L. Murphy, Word meaning in minds and machines. *Psychol. Rev.* **130**, 401–431 (2023).
28. S. T. Piantadosi, *Modern Language Models Refute Chomsky's Approach to Language* (Language Science Press, 2024).
29. A. Goldstein *et al*., Shared computational principles for language processing in humans and deep language models. *Nat. Neurosci.* **25**, 369–380 (2022).
30. A. R. Vaidya, J. Turek, A. G. Huth, Humans and language models diverge when predicting repeating text. arXiv [Preprint] (2023). https://arxiv.org/abs/2310.06408 (Accessed 12 October 2023).
31. B. Shlegeris, F. Roger, L. Chan, E. McLean, Language models are better than humans at next-token prediction. arXiv [Preprint] (2022). https://arxiv.org/abs/2212.11281 (Accessed 29 November 2023).
32. P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, Enriching word vectors with subword information. arXiv [Preprint] (2017). https://arxiv.org/abs/1607.04606 (Accessed 1 February 2024).
33. H. H. Clark, G. L. Murphy, Audience design in meaning and reference. *Adv. Psychol.* **9**, 287–299 (1982).
34. G. J. Stephens, L. J. Silbert, U. Hasson, Speaker-listener neural coupling underlies successful communication. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 14425–14430 (2010).
35. R. Schmälzle, F. E. K. Häcker, C. J. Honey, U. Hasson, Engaged listeners: Shared neural processing of powerful political speeches. *Soc. Cogn. Affect. Neurosci.* **10**, 1137–1143 (2015).
36. C. H. C. Chang, S. A. Nastase, A. Zadbood, U. Hasson, How a speaker herds the audience: Multibrain neural convergence over time during naturalistic storytelling. *Soc. Cogn. Affect. Neurosci.* **19**, nsae059 (2024).
37. A. Suni, J. Šimko, D. Aalto, M. Vainio, Hierarchical representation and estimation of prosody using continuous wavelet transform. *Comput. Speech Lang.* **45**, 123–136 (2017).
38. G. M. Mason, M. H. Goldstein, J. A. Schwade, The role of multisensory development in early language learning. *J. Exp. Child Psychol.* **183**, 48–64 (2019).
39. L. E. Bahrick, R. Lickliter, Intersensory redundancy guides early perceptual and cognitive development. *Adv. Child Dev. Behav.* **30**, 153–187 (2002).
40. L. B. Smith, S. Jayaraman, E. Clerkin, C. Yu, The developing infant creates a curriculum for statistical learning. *Trends Cogn. Sci.* **22**, 325–336 (2018).
41. M. H. Goldstein, A. P. King, M. J. West, Social interaction shapes babbling: Testing parallels between birdsong and speech. *Proc. Natl. Acad. Sci. U.S.A.* **100**, 8030–8035 (2003).
42. A. J. DeCasper, M. J. Spence, Prenatal maternal speech influences newborns' perception of speech sounds. *Infant Behav. Dev.* **9**, 133–150 (1986).
43. G. Dehaene-Lambertz, S. Dehaene, L. Hertz-Pannier, Functional neuroimaging of speech perception in infants. *Science* **298**, 2013–2015 (2002).
44. S. R. Speer, K. Ito, Prosody in first language acquisition - acquiring intonation as a tool to organize information in conversation. *Lang. Linguist. Compass* **3**, 90–110 (2009).
45. B. Mampe, A. D. Friederici, A. Christophe, K. Wermke, Newborns' cry melody is shaped by their native language. *Curr. Biol.* **19**, 1994–1997 (2009).
46. N. Abboub, T. Nazzi, J. Gervain, Prosodic grouping at birth. *Brain Lang.* **162**, 46–59 (2016).
47. E. J. Tenenbaum *et al*., Attention to the mouth and gaze following in infancy predict language development. *J. Child Lang.* **42**, 1173–1190 (2015).
48. S. Dehaene, L. Cohen, J. Morais, R. Kolinsky, Illiterate to literate: Behavioural and cerebral changes induced by reading acquisition. *Nat. Rev. Neurosci.* **16**, 234–244 (2015).
49. Z. M. Saygin *et al*., Connectivity precedes function in the development of the visual word form area. *Nat. Neurosci.* **19**, 1250–1255 (2016).
50. J. Kaplan *et al*., Scaling Laws for Neural Language. arXiv [Preprint] (2020). https://arxiv.org/abs/2001.08361 (Accessed 27 March 2026).
51. A. Aghajanyan *et al*., Scaling laws for generative mixed-modal language models. arXiv [Preprint] (2023). https://arxiv.org/abs/2301.03728 (Accessed 20 August 2025).
52. T. I. Regev *et al*., The time scale of redundancy between prosody and linguistic context. arXiv [Preprint] (2025). https://arxiv.org/abs/2503.11630 (Accessed 24 March 2025).
53. L. Wolf *et al*., Quantifying the redundancy between prosody and text. arXiv [Preprint] (2023). https://arxiv.org/abs/2311.17233 (Accessed 1 December 2023).
54. M. Huh, B. Cheung, T. Wang, P. Isola, The platonic representation hypothesis. arXiv [Preprint] (2024). https://arxiv.org/abs/2405.07987 (Accessed 15 May 2024).
55. W. K. Vong, W. Wang, A. E. Orhan, B. M. Lake, Grounded language acquisition through the eyes and ears of a single child. *Science* **383**, 504–511 (2024).
56. C. E. Shannon, Prediction and entropy of printed English. *Bell Syst. Tech. J.* **30**, 50–64 (1951).
57. L. B. Levitin, Z. Reingold, Entropy of natural languages: Theory and experiment. *Chaos Solitons Fractals* **4**, 709–743 (1994).
58. P. F. Brown, S. A. Della Pietra, V. J. Della Pietra, J. C. Lai, R. L. Mercer, An estimate of an upper bound for the entropy of English. *Comput. Linguist.* **18**, 31–40 (1992).
59. M. Kutas, S. A. Hillyard, Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science* **207**, 203–205 (1980).
60. M. Besson, F. Faita, C. Czternasty, M. Kutas, What's in a pause: Event-related potential analysis of temporal disruptions in written and spoken sentences. *Biol. Psychol.* **46**, 3–23 (1997).
61. S. F. Ehrlich, K. Rayner, Contextual effects on word perception and eye movements during reading. *J. Verbal Learn. Verbal Behav.* **20**, 641–655 (1981).
62. A. Staub, The effect of lexical predictability on distributions of eye fixation durations. *Psychon. Bull. Rev.* **18**, 371–376 (2011).
63. A. Staub, Predictability in language comprehension: Prospects and problems for surprisal. *Annu. Rev. Linguist.* **11**, 17–34 (2025).
64. C. Shain, C. Meister, T. Pimentel, R. Cotterell, R. Levy, Large-scale evidence for logarithmic effects of word predictability on reading time. *Proc. Natl. Acad. Sci. U.S.A.* **121**, e2307876121 (2024).
65. W. L. Taylor, "Cloze Procedure": A new tool for measuring readability. *Journal. Q.* **30**, 415–433 (1953).
66. T. Ullman, Large language models fail on trivial alterations to theory-of-mind tasks. arXiv [Preprint] (2023). https://arxiv.org/abs/2302.08399 (Accessed 25 February 2023).
67. A. Clark, Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behav. Brain Sci.* **36**, 181–204 (2013).
68. J. E. Kosie, M. L. Nencheva, J. A. Jungé, C. Lew-Williams, Models of human learning should capture the multimodal complexity and communicative goals of the natural learning environment. *Lang. Dev. Res.* **5**, 188–220 (2025).
69. M. L. H. Vo, J. M. Henderson, Does gravity matter? Effects of semantic and syntactic inconsistencies on the allocation of attention during scene perception. *J. Vis.* **9**, 24.1–24.15 (2009).
70. M. I. Coco, F. Keller, G. L. Malcolm, Anticipation in real-world scenes: The role of visual context and visual memory. *Cogn. Sci.* **40**, 1995–2024 (2016).
71. D. C. Richardson, R. Dale, Looking to understand: The coupling between speakers' and listeners' eye movements and its relationship to discourse comprehension. *Cogn. Sci.* **29**, 1045–1060 (2005).
72. A. Bangerter, Using pointing and describing to achieve joint focus of attention in dialogue. *Psychol. Sci.* **15**, 415–419 (2004).
73. R. Ryskin, M. S. Nieuwland, Prediction during language comprehension: What is next? *Trends Cogn. Sci.* **27**, 1032–1052 (2023).
74. M. Schrimpf *et al*., The neural architecture of language: Integrative modeling converges on predictive processing. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2105646118 (2021).
75. A. Goldstein *et al*., A unified acoustic-to-speech-to-language embedding space captures the neural basis of natural language processing in everyday conversations. *Nat. Hum. Behav.* **9**, 1041–1055 (2025).
76. G. A. Castellucci, C. K. Kovach, M. A. Howard, J. D. W. Greenlee, M. A. Long, A speech planning network for interactive language use. *Nature* **602**, 117–122 (2022).
77. M. Włodarczak, M. Heldner, Breathing in conversation. *Front. Psychol.* **11**, 575566 (2020).
78. S. A. Nastase *et al*., The "Narratives" fMRI dataset for evaluating models of naturalistic language comprehension. *Sci. Data* **8**, 250 (2021).
79. A. LeBel *et al*., A natural language fMRI dataset for voxelwise encoding models. *Sci. Data* **10**, 555 (2023).
80. P. Boersma, PRAAT, a system for doing phonetics by computer. *Glot Int.* **5**, 341–345 (2001).
81. M. C. Frank, Openly accessible LLMs can help us to understand human cognition. *Nat. Hum. Behav.* **7**, 1825–1827 (2023).
82. T. Wolf *et al*., HuggingFace's transformers: State-of-the-art natural language processing. arXiv [Preprint] (2020). https://arxiv.org/abs/1910.03771 (Accessed 1 February 2024).
83. A. A. Ivanova, How to evaluate the cognitive abilities of LLMs. *Nat. Hum. Behav.* **9**, 230–233 (2025).
84. Y. Yeshurun, M. Nguyen, U. Hasson, Amplification of local changes along the timescale processing hierarchy. *Proc. Natl. Acad. Sci. U.S.A.* **114**, 9475–9480 (2017).
85. T. I. Regev *et al*., Neural populations in the language network differ in the size of their temporal receptive windows. *Nat. Hum. Behav.* **8**, 1924–1942 (2024).
86. M. Kumar *et al*., Bayesian surprise predicts human event segmentation in story listening. *Cogn. Sci.* **47**, e13343 (2023).
87. R. Schaeffer, B. Miranda, S. Koyejo, Are Emergent Abilities of Large Language Models a Mirage? arXiv [Preprint] (2023). https://arxiv.org/abs/2304.15004 (Accessed 1 February 2024).
88. N. J. Smith, R. Levy, The effect of word predictability on reading time is logarithmic. *Cognition* **128**, 302–319 (2013).
89. A. Chao, T. J. Shen, Nonparametric estimation of Shannon's index of diversity when there are unseen species in sample. *Environ. Ecol. Stat.* **10**, 429–443 (2003).
90. I. Nemenman, F. Shafee, W. Bialek, Entropy and inference, revisited. arXiv [Preprint] (2002). https://arxiv.org/abs/physics/0108025 (Accessed 6 April 2026).
91. A. Radford *et al*., Robust speech recognition via large-scale weak supervision. arXiv [Preprint] (2022). https://arxiv.org/abs/2212.04356 (Accessed 7 February 2025).
92. T. Afouras, J. S. Chung, A. Zisserman, LRS3-TED: A large-scale dataset for visual speech recognition. arXiv [Preprint] (2018). https://arxiv.org/abs/1809.00496 (Accessed 26 March 2025).
93. A. Ephrat *et al*., Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. *ACM Trans. Graph.* **37**, 1–11 (2018).
94. J. S. Chung, A. Nagrani, A. Zisserman, VoxCeleb2: Deep speaker recognition. arXiv [Preprint] (2018). https://arxiv.org/abs/1806.05622 (Accessed 26 March 2025).
95. V. Pratap *et al*., Scaling Speech Technology to 1,000+ Languages. arXiv [Preprint] (2023). https://arxiv.org/abs/2305.13516 (Accessed 9 April 2025).
96. M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, M. Sonderegger, "Montreal forced aligner: Trainable text-speech alignment using Kaldi" in *Interspeech 2017* (ISCA,2017), pp. 498–502.
97. A. Baevski *et al*., data2vec: A general framework for self-supervised learning in speech. arXiv [Preprint] (2022). https://arxiv.org/abs/2202.03555 (Accessed 26 March 2025).
98. I. Loshchilov, F. Hutter, Decoupled weight decay regularization. arXiv [Preprint] (2019). https://arxiv.org/abs/1711.05101 (Accessed 18 November 2024).
99. D. J. Barr, R. Levy, C. Scheepers, H. J. Tily, Random effects structure for confirmatory hypothesis testing: Keep it maximal. *J. Mem. Lang.* **68**, 255–278 (2013).
100. J. R. De Leeuw, R. A. Gilbert, B. Luchterhandt, jsPsych: Enabling an open-source collaborative ecosystem of behavioral experiments. *J. Open Source Softw.* **8**, 5351 (2023).