Toward Cognitive Alignment in Large Language Models: Integrating Linguistic Theory and Human Data

Wajdi Zaghouani
Northwestern University in Qatar
wajdi.zaghouani@northwestern.edu

Abstract

Large language models (LLMs) have reshaped natural language processing and reopened foundational questions about the relationship between computational systems, linguistic theory, and human cognition. While recent work shows that neural language models correlate with human behavioral and neural measures of language processing, predictive performance alone does not constitute cognitive adequacy. This position paper argues for a balanced and programmatic integration of cognitive science, linguistics, and NLP centered on the concept of cognitive alignment. We examine both the empirical promise and conceptual limitations of current LLM claims about human language processing. We then outline how experimental paradigms from psycholinguistics, formal insights from linguistic theory, and large-scale computational modeling can be combined to produce more interpretable, data-efficient, and cognitively grounded systems. We propose four concrete initiatives: a Cognitive Alignment Benchmark Suite, hybrid neuro-symbolic modeling challenges, population-level language simulation infrastructures, and interdisciplinary training programs. Together, these steps move toward a unified science of language grounded in empirical rigor, theoretical depth, and computational scalability.

Keywords: Cognitive modeling, computational psycholinguistics, large language models, linguistic theory, interdisciplinary research

1. Introduction

The rapid development of large language models has renewed a longstanding question: can computational systems serve as explanatory models of human language processing? While transformer-based architectures (Vaswani et al., 2017) achieve remarkable predictive performance, the relationship between predictive success and cognitive adequacy remains under-theorized. This tension reflects deeper questions about the nature of linguistic knowledge and the criteria for explanatory adequacy in cognitive science.

Cognitive science, linguistics, and NLP have historically pursued partially independent research programs. Cognitive science investigates representational format, learning constraints, and processing mechanisms. Linguistics develops formal accounts of structure, compositionality, and cross-linguistic variation. NLP prioritizes scalability, empirical coverage, and engineering performance. The emergence of LLMs creates both an opportunity and a risk: the opportunity to reconnect these traditions, and the risk of collapsing explanation into benchmark performance.

This paper advances a balanced thesis. Neural language models provide unprecedented tools for modeling large-scale linguistic structure and predicting human behavioral data. However, predictive correlation alone does not constitute mechanistic explanation. A principled research program must integrate experimental constraints, formal theory, and computational scale. We draw on foundational work in computational psycholinguistics (Levy, 2008), cognitive architectures (Anderson et al., 2004), and formal semantics (Montague, 1973) to articulate criteria for cognitive alignment that go beyond surface-level behavioral matching.

The cognitive modeling community is uniquely positioned to lead this integration. Unlike purely engineering-driven NLP research, cognitive modeling has always maintained commitment to psychological plausibility and theoretical grounding. Unlike purely theoretical linguistics, it embraces computational implementation and empirical validation. The challenge now is to scale these commitments to the era of large language models while preserving their scientific integrity. Importantly, substantial work already exists at the intersection of NLP and cognitive science, including shared tasks on eye-tracking prediction organized at previous editions of the CMCL workshop (Hollenstein et al., 2021, 2022) and large-scale multilingual reading corpora (Siegelman et al., 2022). The present paper builds on these foundations to propose a more systematic integration.

The structure of this paper proceeds as follows. Section 2 defines cognitive alignment as a multi-level constraint spanning behavioral, representational, and developmental dimensions. Section 3 offers a measured critique of current LLM claims. Section 4 outlines a programmatic integration across disciplines. Section 5 proposes concrete infrastructure initiatives. Section 6 discusses challenges and limitations. Section 7 concludes with reflections on the path forward.

Proceedings of the 15th Workshop on Cognitive Modeling and Computational Linguistics (CMCL) @ LREC 2026, pages 347–358
16 May 2026.

2. Cognitive Alignment as a Multi-Level Constraint

We define cognitive alignment not as superficial behavioral similarity, but as convergence across three explanatory levels. This tripartite framework draws on Marr's levels of analysis (Marr, 1982) while extending it to address the specific challenges posed by neural language models. True cognitive alignment requires satisfaction of constraints at multiple levels simultaneously, not merely optimization of any single criterion.

2.1. Behavioral Alignment

Surprisal derived from probabilistic models correlates robustly with reading times and EEG responses (Hale, 2001; Levy, 2008). The surprisal theory of sentence processing posits that processing difficulty at each word is proportional to its negative log probability given context. This relationship has been confirmed across multiple languages and experimental paradigms (Smith and Levy, 2013; Frank et al., 2015; Wilcox et al., 2023).

Neural language model representations capture predictive neural processing signals. Studies using magnetoencephalography (MEG) and functional magnetic resonance imaging (fMRI) demonstrate that transformer activations correlate with neural responses in language-selective cortical regions (Schrimpf et al., 2021; Goldstein et al., 2022). Caucheteux and King demonstrate that GPT-2 activations predict brain activity with remarkable precision (Caucheteux and King, 2022). These findings establish that distributional models approximate certain aspects of incremental expectation.

However, such correlations remain level-neutral. As Lake et al. (2017) argue, human-like cognition requires structured inductive biases that support generalization beyond observed distributions. Correlation with processing difficulty does not entail isomorphism of representation. A model might achieve high brain-score correlations through mechanisms entirely unlike those deployed by the human brain. Furthermore, behavioral alignment on naturalistic stimuli does not guarantee alignment on carefully controlled experimental materials designed to isolate specific linguistic phenomena.

Recent work by Wilcox et al. (2020) demonstrates that while neural language models capture broad patterns of syntactic expectation, they fail on specific constructions that require hierarchical structure sensitivity. Similarly, Futrell et al. (2019) show that models trained on realistic language data are less robust, less sensitive, and less selective than humans in processing garden-path sentences, which are known to cause significant processing difficulty for human readers as well. Linzen et al. (2016) introduced systematic tests of subject-verb agreement across intervening material, revealing that models often fail where humans succeed. These findings underscore the need for fine-grained behavioral testing beyond aggregate correlation metrics.

Crucially, recent studies have revealed that the relationship between model scale and cognitive plausibility is more nuanced than previously assumed. Kuribayashi et al. (2024) show that instruction-tuned LLMs often yield worse psychometric predictive power for human reading behavior than base LLMs, indicating that alignment with human preferences does not entail alignment with human processing. Furthermore, the inverse scaling effect, in which surprisal from larger pre-trained language models provides a poorer fit to human reading times, represents a significant challenge to the assumption that better language models are necessarily more cognitively plausible (Oh et al., 2025). Kuribayashi et al. (2025) offer a more nuanced picture, demonstrating that when internal layers of larger LMs are examined rather than only final-layer outputs, larger models align with human sentence processing data as well as or better than smaller ones. This finding holds across behavioral (self-paced reading, gaze durations, maze processing times) and neurophysiological (N400 potentials) measures, suggesting that the cognitive plausibility of larger LMs has been underestimated.

2.2. Representational Alignment

Formal linguistics provides explicit hypotheses about the structure of linguistic representations. Construction Grammar emphasizes form-meaning pairings at multiple levels of abstraction (Goldberg, 1995, 2006). Abstract Meaning Representation offers graph-structured semantic representations that capture predicate-argument structure (Banarescu et al., 2013). FrameNet encodes frame-semantic relations that link linguistic expressions to conceptual structures (Baker et al., 1998).

While neural architectures may implicitly encode similar information, interpretability research remains essential. Mechanistic analysis of transformer components reveals attention heads specialized for specific syntactic relations (Clark et al., 2019; Voita et al., 2019). Causal probing techniques enable controlled intervention studies that distinguish correlation from causation (Vig et al., 2020; Elazar et al., 2021). However, systematic mapping to formal constructs remains incomplete.

The question of whether transformers implement symbolic computation implicitly or represent a fundamentally different computational paradigm remains open. Smolensky (1990) proposed tensor product representations as a bridge between connectionist and symbolic computation. More recent work explores whether attention mechanisms im-

plement variable binding (Manning et al., 2020). Hewitt and Manning (2019) introduced structural probes to test whether syntax trees are recoverable from neural representations, finding evidence for implicit hierarchical encoding.

Critically, representational alignment requires more than demonstrating that linguistic information is present in model activations. It requires showing that this information is organized in ways consistent with theoretical proposals and, crucially, that the model uses this information in the same way humans do during processing. This functional criterion remains largely unaddressed in current probing research. Ravfogel et al. (2020) demonstrate that probing accuracy can be high even when the probed information is not causally implicated in model behavior, highlighting the need for causal rather than merely correlational evidence.

From the perspective of brain data, recent advances using fMRI and representational similarity analysis techniques have begun mapping whether LLM embeddings are structurally aligned with brain activations. Doerig et al. (2025) show that LLM embeddings of scene captions successfully characterize brain activity evoked by viewing natural scenes, revealing alignment between LLM representations and high-level visual processing in the human brain. Using RSA on 23 mainstream LLMs, Ren et al. (2025) find that LLM-brain similarity correlates positively with pre-training data size and model scaling, and that alignment training can significantly improve LLM-brain correspondence. These studies suggest genuine points of convergence between representational formats in artificial and biological systems.

Furthermore, representations must support the productivity and systematicity characteristic of human language (Fodor and Pylyshyn, 1988). A model might encode syntactic structure in a way that supports parsing but fails to support novel combination. Testing for systematic generalization thus provides a crucial constraint on representational adequacy.

2.3. Developmental Alignment

Children acquire language from sparse, structured, socially embedded input. Bayesian models of word learning demonstrate that children leverage cross-situational statistics efficiently (Xu and Tenenbaum, 2007; Smith et al., 2014). Theory-driven frameworks for human-like learning highlight the role of inductive bias and active hypothesis testing (Lake et al., 2017; Tenenbaum et al., 2011).

Web-scale training regimes differ radically from child-directed input. The CHILDES database documents that children hear approximately 10 to 15 million words by age 10 (MacWhinney, 2000), orders of magnitude less than LLM training corpora. Furthermore, child-directed speech exhibits distinctive prosodic, lexical, and syntactic properties optimized for learning (Soderstrom, 2007).

Recent work has begun to explore whether models trained on developmentally realistic data can achieve human-like generalization. Warstadt et al. (2020) introduce the BLiMP benchmark for evaluating grammatical knowledge, finding that models require substantially more data than children to acquire comparable competence. The BabyLM Challenge (Warstadt et al., 2023) explicitly addresses this gap by constraining training data to approximate child language exposure, revealing both the difficulty of data-efficient learning and promising architectural directions.

Bridging the developmental gap requires curriculum design and learning constraints informed by developmental research. Work on curriculum learning (Bengio et al., 2009) and meta-learning (Finn et al., 2017) offers computational frameworks, but integration with developmental theory remains nascent. Understanding how children leverage multimodal, interactive, and socially scaffolded learning environments may prove essential for achieving data-efficient language acquisition in artificial systems.

The social dimension deserves particular emphasis. Children learn language in interaction with caregivers who provide contingent feedback, joint attention, and communicative scaffolding (Tomasello, 2003). Current LLMs learn from static text corpora stripped of these interactive dynamics. Active learning and curiosity-driven exploration suggest alternative training regimes grounded in intrinsic motivation rather than passive ingestion (Gottlieb et al., 2013; Oudeyer et al., 2007). Children do not simply observe language; they actively probe their environment through questions and experimentation. Incorporating such mechanisms may enhance data efficiency and generalization, and may be necessary for achieving human-like acquisition trajectories.

3. A Measured Critique of LLM Hype

Public discourse often equates LLM fluency with understanding. This conflation introduces both conceptual and scientific risks. A balanced assessment must acknowledge genuine achievements while maintaining appropriate skepticism about overclaims.

3.1. Prediction versus Explanation

Predictive models can approximate behavioral patterns without instantiating explanatory mechanisms. Shanahan (2024) caution against anthropomorphic descriptions that obscure mechanistic uncertainty. When we say a model “understands” or “knows,”

we risk importing unjustified assumptions about its internal states.

The distinction between prediction and explanation has deep roots in philosophy of science (Hempel, 1965). A model might predict human judgments with high accuracy while implementing mechanisms entirely unlike human cognition. Conversely, a model with lower predictive accuracy might better capture the causal structure of human language processing.

Recent debates about whether LLMs possess genuine semantic understanding illustrate these tensions. Bender and Koller (2020) argue that models trained solely on form cannot acquire meaning. Piantadosi and Hill (2022) counter that distributional information may suffice for certain semantic tasks. Resolving this debate requires precise operationalization of "understanding" and empirical tests that distinguish genuine comprehension from sophisticated pattern matching.

We suggest that the cognitive modeling community can contribute by developing precise behavioral signatures that differentiate alternative hypotheses about model cognition. Rather than debating understanding in the abstract, we should identify specific predictions that competing accounts make and design experiments to adjudicate between them.

3.2. Scale as Surrogate for Theory

LLMs achieve performance through scale. Yet scale cannot substitute for representational clarity. Baroni (2022) argues that grammatical abstraction remains indispensable for systematic compositionality. Typological diversity studies demonstrate that distributional similarity alone does not guarantee cross-linguistic generalization (Bender, 2011).

The emergence of apparently complex behaviors from simple learning objectives at scale has been termed "emergence" (Wei et al., 2022). However, recent work questions whether true qualitative transitions occur or whether apparent emergence reflects smooth scaling obscured by measurement artifacts (Schaeffer et al., 2023). Distinguishing genuine cognitive capacities from brittle shortcuts requires careful experimental design.

Furthermore, scaling laws that hold for English may not generalize to morphologically rich or low-resource languages. Wu and Dredze (2020) demonstrate substantial performance variation across languages even for multilingual models. This concern is equally relevant for cognitively oriented evaluation: the majority of reading time corpora and psycholinguistic benchmarks are in English, limiting the generalizability of cognitive alignment claims. Wilcox et al. (2023) address this gap by testing the predictions of surprisal theory across 11 languages from five language families, finding that surprisal is consistently predictive of reading times crosslinguistically, that contextual entropy also contributes, and that the surprisal-reading time relationship is linear. While these results offer the most robust cross-linguistic support for surprisal theory to date, they also highlight that model performance varies substantially across languages and scripts.

The cognitive modeling perspective emphasizes that human language acquisition is remarkably uniform across languages despite surface typological diversity, suggesting that human learners possess inductive biases that current models lack. Achieving truly general cognitive alignment may require architectural innovations beyond simple scaling, and evaluation across a broad typological sample is essential before claims of cognitive plausibility can be considered robust.

3.3. Benchmark Overfitting and Ecological Validity

Many evaluations rely on static benchmarks rather than dynamic processing measures. Eye-tracking corpora such as Dundee (Kennedy et al., 2003), GECO (Cop et al., 2017), and self-paced reading corpora such as Natural Stories (Futrell et al., 2021) provide ecologically valid constraints that go beyond accuracy metrics. These resources capture the temporal dynamics of language processing, revealing phenomena invisible to end-state accuracy measures. The Maze task has recently emerged as a complementary paradigm offering superior localization of processing effects, as demonstrated by Boyce and Levy (2023) who adapted the method for naturalistic passages using the Natural Stories corpus.

Benchmarks risk becoming targets rather than measures (Goodhart, 1984). As models are optimized for specific test sets, performance may inflate without corresponding improvements in genuine linguistic competence. McCoy et al. (2019) demonstrate that models achieving high accuracy on natural language inference benchmarks rely on superficial heuristics rather than robust reasoning.

Process-level evaluation using reading time prediction, N400 amplitude modeling, and garden-path recovery times offers more stringent tests of cognitive alignment. These measures constrain not just what the model gets right but how it processes linguistic input incrementally. The Provo corpus (Luke and Christianson, 2018) and the UCL corpus (Frank and Bod, 2011) provide valuable resources for such evaluation, but their adoption in NLP evaluation remains limited.

3.4. Societal and Ethical Dimensions

LLMs inherit biases from training data. Comprehensive auditing frameworks document systematic dis-

parities across demographic groups (Blodgett et al., 2020). Implicit association paradigms adapted from psychology reveal encoded stereotypes (Caliskan et al., 2017). These concerns extend naturally to the cross-linguistic dimension. Work on multilingual author profiling (Zaghouani and Charfi, 2018) reveals how language, dialect, and demographic variation interact in ways that both NLP and cognitive science must account for.

Research with low-resource languages raises ethical considerations regarding data sovereignty, community consent, and equitable benefit distribution (Bird, 2020). The computational resources required for LLM training concentrate power in wealthy institutions, potentially exacerbating existing inequalities in NLP research capacity. A balanced stance recognizes both empirical promise and conceptual limits while attending to broader societal implications. Cognitive alignment research can contribute by ensuring that models not only perform well but do so through mechanisms that are interpretable, fair, and aligned with human values.

4. Toward a Programmatic Integration

Building on the preceding critique, we outline a research program that leverages the complementary strengths of cognitive science, linguistics, and NLP.

4.1. Cognitive Science as Constraint Provider

Psycholinguistic paradigms offer fine-grained process-level data. Reading time studies using eye-tracking provide millisecond-resolution measures of incremental processing difficulty (Rayner, 2009). Event-related potential studies reveal distinct neural signatures for semantic versus syntactic violations (Kutas and Hillyard, 1980; Osterhout and Holcomb, 1992). These paradigms constrain models not just to produce correct outputs but to exhibit appropriate processing dynamics.

Lesioning studies of neural language models demonstrate how artificial systems can simulate selective impairments, creating bidirectional hypotheses between computational and neurocognitive research (Lakretz et al., 2021). By selectively ablating components and observing behavioral consequences, researchers can test causal hypotheses about representational organization.

Active learning and curiosity-driven exploration suggest alternative training regimes grounded in intrinsic motivation rather than passive ingestion (Gottlieb et al., 2013; Oudeyer et al., 2007). Children do not simply observe language; they actively probe their environment through questions and experimentation. Incorporating such mechanisms may enhance data efficiency and generalization.

Visual world paradigm studies provide real-time measures of referential processing (Tanenhaus et al., 1995). These paradigms could be adapted to evaluate whether model predictions align with human incremental interpretation in grounded contexts. The tight temporal coupling between linguistic input and visual attention provides a rich source of constraints on incremental processing that goes beyond what text-only measures can capture.

Memory and attention constraints from cognitive psychology also deserve consideration. Human working memory limitations shape sentence processing (Gibson, 1998), and models that incorporate similar constraints may better approximate human behavior. The interplay between memory retrieval and predictive processing remains an active area of investigation that computational models can help elucidate (Lewis and Vasishth, 2005).

4.2. Linguistics as Representational Scaffold

Formal theory supplies structured hypotheses about compositionality and semantic roles. Embedding structured representations such as AMR (Banarescu et al., 2013) or frame semantics (Baker et al., 1998) into neural architectures may enhance interpretability without sacrificing coverage.

Dependency grammar provides cross-linguistically applicable representations of syntactic structure (Tesnière, 1959; Mel'čuk, 1988). Universal Dependencies (Nivre et al., 2016) offers a standardized annotation framework enabling systematic cross-linguistic comparison.

Typological databases such as WALS (Dryer and Haspelmath, 2013) encode structural variation across languages, providing constraints for evaluating whether models capture universal versus language-specific properties. Models claiming cognitive alignment should perform consistently across typologically diverse languages, not just English. Construction grammar offers a framework for representing form-meaning pairings at multiple levels of abstraction (Goldberg, 1995), potentially offering better alignment with gradient human judgments and productivity patterns.

Discourse and pragmatic theory provide additional constraints often neglected in NLP evaluation. Coherence relations (Asher and Lascarides, 2003), information structure (Krifka, 2007), and pragmatic inference (Grice, 1975) all shape human language processing in ways that current benchmarks rarely assess.

4.3. NLP as Hypothesis Generator

Computational models can surface distributional regularities at scales beyond manual analysis.

Corpus-based approaches reveal statistical patterns that may inform theoretical proposals (Gries, 2006). The distributional hypothesis suggests that semantic similarity correlates with distributional similarity (Harris, 1954), a claim that neural embeddings operationalize at unprecedented scale.

Iterated learning frameworks demonstrate how population-level transmission shapes structure (Kirby et al., 2014). Mechanistic interpretability tools enable controlled intervention at representational levels (Geiger et al., 2021). Neuro-symbolic integration offers a path toward combining the strengths of neural and symbolic computation (Garcez and Lamb, 2020). Hybrid architectures that incorporate explicit symbolic reasoning may better capture the compositional structure of human language while retaining neural flexibility.

5. Infrastructure for Cognitive Modeling

We propose four initiatives aligned with cognitive and computational linguistics research priorities:

5.1. Cognitive Alignment Benchmark Suite

A comprehensive benchmark suite should integrate multiple data sources: eye-tracking corpora including Dundee (Kennedy et al., 2003) and GECO (Cop et al., 2017); self-paced reading data from the Natural Stories corpus (Futrell et al., 2021) and the UCL corpus (Frank and Bod, 2011); Maze task data providing superior localization (Boyce and Levy, 2023); ERP datasets targeting specific linguistic phenomena (Fedorenko et al., 2016); and child-directed speech corpora from CHILDES (MacWhinney, 2000). Recent large-scale multilingual eye-tracking resources, such as the Multilingual Eye-movement Corpus (MECO) covering 13 languages (Siegelman et al., 2022) and the ongoing MultiplEYE initiative (Jakobi et al., 2025), should also be incorporated to ensure cross-linguistic coverage.

Evaluation should span multiple metrics: surprisal correlation with reading times; psychometric predictive power for individual differences (Kuribayashi et al., 2024); representational similarity analysis comparing model activations to neural data; and incremental commitment measures capturing garden-path effects. Critically, the benchmark should include adversarial test sets designed to probe specific theoretical predictions, not just naturalistic data.

The CMCL shared tasks on eye-tracking prediction (Hollenstein et al., 2021, 2022) offer valuable precedents for standardized evaluation, and future benchmarks should build on these by incorporating a wider range of behavioral paradigms and languages.

5.2. Hybrid Modeling Shared Task

A shared task should require explicit integration of structured semantic representations with neural architectures. Participants would receive AMR-annotated data (Banarescu et al., 2013) and FrameNet annotations (Baker et al., 1998) alongside raw text.

Evaluation criteria should include compositional generalization on COGS (Kim and Linzen, 2020) and SCAN (Lake and Baroni, 2018) benchmarks; interpretability assessments using probing tasks and attention analysis; and behavioral alignment with human processing data. Submissions should include documentation of how structured knowledge is incorporated and used during inference. Such a shared task would incentivize development of architectures that bridge neural and symbolic paradigms rather than treating them as competing approaches. It would also generate valuable data on which integration strategies prove most effective for different aspects of cognitive alignment.

5.3. Multi-Agent Language Evolution Platform

Grounded in iterated learning theory (Kirby et al., 2014), a simulation platform should enable controlled experiments on language emergence and change. Agents would communicate to accomplish shared goals, with language emerging from interaction rather than being imposed externally.

The platform should support manipulation of population structure, transmission fidelity, and environmental complexity. Emergent communication systems could be analyzed using standard linguistic tools, enabling comparison with natural language typology. Such simulations provide controlled environments for testing hypotheses about language universals and historical change. Integration with reinforcement learning frameworks would enable exploration of how communicative utility shapes linguistic structure (Lazaridou et al., 2020). The platform could also support investigation of how social structure and network topology affect language evolution.

5.4. Cross-Disciplinary Training Programs

Sustainable progress requires researchers fluent in multiple traditions. Joint doctoral programs bridging computational linguistics, cognitive psychology, and formal linguistics would address this need. Core curricula should include probability theory and statistical modeling; formal syntax and semantics;

experimental methods in psycholinguistics; and machine learning and neural network architectures.

Shared repositories linking computational implementations, experimental materials, and theoretical analyses would lower barriers to interdisciplinary collaboration. Summer schools and workshops bringing together researchers from different traditions can foster communication and identify productive research directions. Funding agencies should incentivize interdisciplinary proposals, and review panels should include expertise spanning the relevant fields.

6. Challenges and Limitations

Several challenges complicate the proposed agenda. First, the three levels of alignment may conflict: a model achieving strong behavioral alignment might violate theoretical constraints on representation, or vice versa. Resolving such conflicts requires principled criteria for adjudicating trade-offs. We suggest that when conflicts arise, they should be treated as scientifically informative rather than merely problematic, potentially revealing inadequacies in current theories. Multi-objective optimization across potentially conflicting criteria requires careful methodological development, and the field has yet to establish widely accepted frameworks for weighing behavioral fit against representational fidelity or developmental plausibility.

Second, human language processing itself remains incompletely understood. Models cannot align with a target that is poorly specified. Ongoing empirical research must continue refining our understanding of human cognition alongside computational modeling efforts. The interaction between modeling and empirical research should be bidirectional, with models generating predictions that drive new experiments and experimental results informing model design.

Third, computational and experimental resources remain unevenly distributed. Many of the proposed initiatives require substantial investment in data collection, infrastructure, and interdisciplinary training. Ensuring equitable access and participation will require deliberate effort. In particular, the concentration of reading time datasets on English and a handful of European languages limits our ability to assess whether cognitive alignment generalizes across the world's languages. Recent efforts toward multilingual eye-tracking data collection, such as MECO (Siegelman et al., 2022) covering 13 languages and the MultiplEYE initiative (Jakobi et al., 2025) collecting data across numerous additional languages, represent important steps toward addressing this gap. However, many language families, scripts, and morphological typologies remain unrepresented. The concentration of NLP research and psycholinguistic experimentation on English continues to limit our understanding of how well current approaches generalize, and the computational resources required for LLM training further concentrate power in wealthy institutions.

Fourth, evaluation metrics for cognitive alignment remain underdeveloped. While we have proposed multiple criteria, integrating them into a coherent framework and validating that framework against human cognition presents ongoing challenges. Standardized evaluation protocols would enable systematic comparison across models and facilitate meta-analysis, but the diversity of data types (reading times, ERP amplitudes, fMRI patterns, acquisition trajectories) makes unified scoring non-trivial.

Finally, the relationship between cognitive alignment and practical utility is unclear. Systems optimized for engineering performance may differ substantially from cognitively aligned systems. Whether cognitive alignment enhances robustness, interpretability, and alignment with human values remains an empirical question. We hypothesize that cognitively aligned systems may exhibit better generalization and more predictable failure modes, but this requires systematic investigation.

7. Conclusion

Large language models mark a significant empirical advance. They demonstrate that distributional learning at scale can capture substantial linguistic structure. Correlations with human behavioral and neural data suggest genuine points of contact between artificial and natural language processing.

However, cognitive modeling requires more than predictive performance. A principled research program must integrate experimental constraint, formal theory, and computational scalability. Each tradition contributes essential perspectives: cognitive science provides process-level data and learning constraints; linguistics supplies representational hypotheses and cross-linguistic generalizations; NLP offers computational tools for modeling and analysis.

By pursuing cognitive alignment across behavioral, representational, and developmental levels, the field can move beyond hype toward cumulative scientific progress. The initiatives proposed here, including benchmark suites, shared tasks, simulation platforms, and training programs, offer concrete steps toward this goal.

The ultimate aim is not merely to build better language models but to understand language itself: how it is represented, processed, learned, and used. This understanding has implications beyond basic science, informing applications in education, clinical intervention, and human-computer interac-

tion. A cognitively aligned approach ensures that progress in language technology remains grounded in and accountable to our knowledge of human cognition. The path forward requires sustained collaboration across disciplinary boundaries, and we hope this position paper contributes to ongoing dialogue and inspires concrete initiatives advancing the vision of a unified science of language.

8. Limitations

This paper proposes a research agenda and is necessarily constrained in several ways. As a position paper, it does not include empirical validation of the proposed framework. The tripartite alignment scheme (behavioral, representational, developmental) is presented as a guiding framework rather than a formally specified evaluation metric, and operationalizing each level into concrete, standardized measures remains an open challenge.

The paper focuses primarily on text-based language processing and does not extensively address multimodal aspects of human language use, including gesture, prosody, and visual grounding, which are increasingly recognized as central to human communication. Human language processing is fundamentally situated in embodied, social contexts, and a complete account of cognitive alignment would need to incorporate these dimensions.

Furthermore, while we advocate for cross-linguistic evaluation, the examples and resources discussed in this paper still skew toward well-resourced languages, particularly English and European languages. Extending cognitive alignment research to typologically diverse languages, including morphologically rich languages, tonal languages, and languages with non-Latin scripts, is essential for testing the universality of proposed alignment criteria but faces significant practical barriers in terms of both data availability and computational infrastructure. The collection of psycholinguistic processing data (e.g., eye-tracking corpora, self-paced reading data, ERP recordings) in underrepresented languages remains a crucial bottleneck.

9. Ethics Statement

This paper proposes a research agenda and does not involve new human subjects data. The proposed research directions raise ethical considerations regarding data collection from human participants, which should follow established protocols including informed consent and appropriate compensation. Research with low-resource languages should adhere to community-engaged practices respecting data sovereignty and ensuring equitable benefit distribution.

Acknowledgments

This work was made possible by the National Priorities Research Program grant NPRP14C-0916-210015 from the Qatar National Research Fund (QNRF), part of the Qatar Research, Development and Innovation Council (QRDI). The author also acknowledge the Artificial Intelligence and Media Lab (AIM Lab) at Northwestern University in Qatar (NU-Q) and the MARSAD Lab for providing valuable resources and support that contributed to this research.

10. Bibliographical References

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of the mind. *Psychological Review*, 111(4), 1036–1060.

Asher, N., & Lascarides, A. (2003). *Logics of Conversation*. Cambridge University Press.

Baker, C. F., Fillmore, C. J., & Lowe, J. B. (1998). The Berkeley FrameNet Project. In *Proceedings of COLING-ACL*, 86–90.

Banarescu, L., Bonial, C., Cai, S., Georgescu, M., Griffitt, K., Hermjakob, U., Knight, K., Koehn, P., Palmer, M., & Schneider, N. (2013). Abstract meaning representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop*, 178–186.

Baroni, M. (2022). On the proper role of linguistically oriented deep net analysis in linguistic theorising. In *Algebraic Structures in Natural Language*, 1–28. CRC Press.

Bender, E. M., & Koller, A. (2020). Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of ACL*, 5185–5198.

Bender, E. M. (2011). On achieving and evaluating language-independence in NLP. *Linguistic Issues in Language Technology*, 6(3), 1–26.

Bengio, Y., Louradour, J., Collobert, R., & Weston, J. (2009). Curriculum learning. In *Proceedings of ICML*, 41–48.

Bird, S. (2020). Decolonising speech and language technology. In *Proceedings of COLING*, 3504–3519.

Blodgett, S. L., Barocas, S., Daumé III, H., & Wallach, H. (2020). Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of ACL*, 5454–5476.

Boyce, V., & Levy, R. (2023). A-maze of Natural Stories: Comprehension and surprisal in the Maze task. *Glossa Psycholinguistics*, 2(1).

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334), 183–186.

Caucheteux, C., & King, J. R. (2022). Brains and algorithms partially converge in natural language processing. *Communications Biology*, 5(1), 134.

Clark, K., Khandelwal, U., Levy, O., & Manning, C. D. (2019). What does BERT look at? An analysis of BERT's attention. In *Proceedings of the BlackboxNLP Workshop*, 276–286.

Cop, U., Dirix, N., Drieghe, D., & Duyck, W. (2017). Presenting GECO: An eyetracking corpus of monolingual and bilingual sentence reading. *Behavior Research Methods*, 49(2), 602–615.

Doerig, A., Kietzmann, T. C., Allen, E., Wu, Y., Naselaris, T., Kay, K., & Charest, I. (2025). High-level visual representations in the human brain are aligned with large language models. *Nature Machine Intelligence*, 7(8), 1220–1234.

Dryer, M. S., & Haspelmath, M. (Eds.). (2013). *WALS Online*. Max Planck Institute for Evolutionary Anthropology.

Elazar, Y., Ravfogel, S., Jacovi, A., & Goldberg, Y. (2021). Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the ACL*, 9, 160–175.

Fedorenko, E., Scott, T. L., Brunner, P., Coon, W. G., Pritchett, B., Schalk, G., & Kanwisher, N. (2016). Neural correlate of the construction of sentence meaning. *Proceedings of the National Academy of Sciences*, 113(41), E6256–E6262.

Finn, C., Abbeel, P., & Levine, S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of ICML*, 1126–1135.

Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28(1-2), 3–71.

Frank, S. L., & Bod, R. (2011). Insensitivity of the human sentence-processing system to hierarchical structure. *Psychological Science*, 22(6), 829–834.

Frank, S. L., Otten, L. J., Galli, G., & Vigliocco, G. (2015). The ERP response to the amount of information conveyed by words in sentences. *Brain and Language*, 140, 1–11.

Futrell, R., Wilcox, E., Mober, T., Qian, P., Ballesteros, M., Dyer, C., Gibson, E., & Levy, R. (2019). Neural language models as psycholinguistic subjects: Representations of syntactic state. In *Proceedings of NAACL-HLT*, 32–42.

Futrell, R., Gibson, E., Tily, H. J., Blank, I., Vishnevetsky, A., Piantadosi, S. T., & Fedorenko, E. (2021). The Natural Stories corpus: A reading-time corpus of English texts containing rare syntactic constructions. *Language Resources and Evaluation*, 55(1), 63–77.

Garcez, A. d'A., & Lamb, L. C. (2020). Neurosymbolic AI: The 3rd wave. *arXiv preprint arXiv:2012.05876*.

Geiger, A., Lu, H., Icard, T., & Potts, C. (2021). Causal abstractions of neural networks. In *Advances in Neural Information Processing Systems*, 34, 9574–9586.

Gibson, E. (1998). Linguistic complexity: Locality of syntactic dependencies. *Cognition*, 68(1), 1–76.

Goldberg, A. E. (1995). *Constructions: A Construction Grammar Approach to Argument Structure*. University of Chicago Press.

Goldberg, A. E. (2006). *Constructions at Work: The Nature of Generalization in Language*. Oxford University Press.

Goldstein, A., et al. (2022). Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience*, 25(3), 369–380.

Goodhart, C. A. (1984). Problems of monetary management: The UK experience. In *Monetary Theory and Practice*, 91–121. Springer.

Gottlieb, J., Oudeyer, P. Y., Lopes, M., & Baranes, A. (2013). Information-seeking, curiosity, and attention: Computational and neural mechanisms. *Trends in Cognitive Sciences*, 17(11), 585–593.

Grice, H. P. (1975). Logic and conversation. In *Syntax and Semantics 3: Speech Acts*, 41–58. Academic Press.

Gries, S. T. (2006). Corpus-based methods and cognitive semantics: The many senses of to run. In *Corpora in Cognitive Linguistics*, 57–99. Mouton de Gruyter.

Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of NAACL*, 1–8.

Harris, Z. S. (1954). Distributional structure. *Word*, 10(2-3), 146–162.

Hempel, C. G. (1965). *Aspects of Scientific Explanation*. Free Press.

Hewitt, J., & Manning, C. D. (2019). A structural probe for finding syntax in word representations. In *Proceedings of NAACL-HLT*, 4129–4138.

Hollenstein, N., Chersoni, E., Jacobs, C. L., Oseki, Y., Prévot, L., & Santus, E. (2021). CMCL 2021 shared task on eye-tracking prediction. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, 72–78.

Hollenstein, N., Chersoni, E., Jacobs, C. L., Oseki, Y., Prévot, L., & Santus, E. (2022). CMCL 2022 shared task on multilingual and crosslingual prediction of human reading behavior. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*.

Jakobi, D. N., Stegenwallner-Schütz, M., Hollenstein, N., et al. (2025). MultiplEYE: Creating a multilingual eye-tracking-while-reading corpus. In *Proceedings of the 2025 Symposium on Eye Tracking Research and Applications*, 1–11. ACM.

Kennedy, A., Hill, R., & Pynte, J. (2003). The Dundee corpus. In *Proceedings of the 12th European Conference on Eye Movement*.

Kim, N., & Linzen, T. (2020). COGS: A compositional generalization challenge based on semantic interpretation. In *Proceedings of EMNLP*, 9087–9105.

Kirby, S., Griffiths, T., & Smith, K. (2014). Iterated learning and the evolution of language. *Current Opinion in Neurobiology*, 28, 108–114.

Krifka, M. (2007). Basic notions of information structure. In *Interdisciplinary Studies on Information Structure*, 6, 13–55.

Kuribayashi, T., Oseki, Y., & Baldwin, T. (2024). Psychometric predictive power of large language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, 1983–2005.

Kuribayashi, T., Oseki, Y., Ben Taieb, S., Inui, K., & Baldwin, T. (2025). Large language models are human-like internally. *Transactions of the Association for Computational Linguistics*, 13, 1743–1766.

Kutas, M., & Hillyard, S. A. (1980). Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science*, 207(4427), 203–205.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40, e253.

Lake, B. M., & Baroni, M. (2018). Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *Proceedings of ICML*, 2873–2882.

Lakretz, Y., Desbordes, T., King, J. R., Crabbé, B., Oquab, M., & Dehaene, S. (2021). Can RNNs learn recursive nested subject-verb agreements? In *Proceedings of NAACL-HLT*, 3558–3569.

Lazaridou, A., Potapenko, A., & Tieleman, O. (2020). Multi-agent communication meets natural language: Synergies between functional and structural language learning. In *Proceedings of ACL*, 7663–7674.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, 106(3), 1126–1177.

Lewis, R. L., & Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science*, 29(3), 375–419.

Linzen, T., Dupoux, E., & Goldberg, Y. (2016). Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the ACL*, 4, 521–535.

Luke, S. G., & Christianson, K. (2018). The Provo Corpus: A large eye-tracking corpus with predictability norms. *Behavior Research Methods*, 50(2), 826–833.

MacWhinney, B. (2000). *The CHILDES Project: Tools for Analyzing Talk* (3rd ed.). Lawrence Erlbaum Associates.

Manning, C. D., Clark, K., Hewitt, J., Khandelwal, U., & Levy, O. (2020). Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences*, 117(48), 30046–30054.

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. W.H. Freeman.

McCoy, R. T., Pavlick, E., & Linzen, T. (2019). Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of ACL*, 3428–3448.

Mel'čuk, I. (1988). *Dependency Syntax: Theory and Practice*. SUNY Press.

Montague, R. (1973). The proper treatment of quantification in ordinary English. In *Approaches to Natural Language*, 221–242. Springer.

Nivre, J., et al. (2016). Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of LREC*, 1659–1666.

Oh, B.-D., Zhu, H., & Schuler, W. (2025). The inverse scaling effect of pre-trained language model surprisal is not due to data leakage. In *Findings of the Association for Computational Linguistics: ACL 2025*, 1820–1827.

Osterhout, L., & Holcomb, P. J. (1992). Event-related brain potentials elicited by syntactic anomaly. *Journal of Memory and Language*, 31(6), 785–806.

Oudeyer, P. Y., Kaplan, F., & Hafner, V. V. (2007). Intrinsic motivation systems for autonomous mental development. *IEEE Transactions on Evolutionary Computation*, 11(2), 265–286.

Piantadosi, S. T., & Hill, F. (2022). Meaning without reference in large language models. *arXiv preprint arXiv:2208.02957*.

Ravfogel, S., Elazar, Y., Gonen, H., Twiton, M., & Goldberg, Y. (2020). Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of ACL*, 7237–7256.

Rayner, K. (2009). Eye movements and attention in reading, scene perception, and visual search. *The Quarterly Journal of Experimental Psychology*, 62(8), 1457–1506.

Ren, Y., Jin, R., Zhang, T., & Xiong, D. (2025). Do large language models mirror cognitive language processing? In *Proceedings of the 31st International Conference on Computational Linguistics (COLING)*, 2988–3001.

Schaeffer, R., Miranda, B., & Koyejo, S. (2023). Are emergent abilities of large language models a mirage? In *Advances in Neural Information Processing Systems*, 36.

Schrimpf, M., et al. (2021). The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45), e2105646118.

Shanahan, M. (2024). Talking about large language models. *Communications of the ACM*, 67(2), 68–79.

Siegelman, N., Schroeder, S., Acartürk, C., et al. (2022). Expanding horizons of cross-linguistic research on reading: The Multilingual Eye-movement Corpus (MECO). *Behavior Research Methods*, 54(6), 2843–2863.

Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3), 302–319.

Smith, L. B., Suanda, S. H., & Yu, C. (2014). The unrealized promise of infant statistical word-referent learning. *Trends in Cognitive Sciences*, 18(5), 251–258.

Smolensky, P. (1990). Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial Intelligence*, 46(1-2), 159–216.

Soderstrom, M. (2007). Beyond babytalk: Re-evaluating the nature and content of speech input to preverbal infants. *Developmental Review*, 27(4), 501–532.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, 268(5217), 1632–1634.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science*, 331(6022), 1279–1285.

Tesnière, L. (1959). *Éléments de syntaxe structurale*. Klincksieck.

Tomasello, M. (2003). *Constructing a Language: A Usage-Based Theory of Language Acquisition*. Harvard University Press.

Vaswani, A., et al. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*, 30, 5998–6008.

Vig, J., et al. (2020). Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems*, 33, 12388–12401.

Voita, E., Talbot, D., Moiseev, F., Sennrich, R., & Titov, I. (2019). Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of ACL*, 5797–5808.

Warstadt, A., et al. (2020). BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the ACL*, 8, 377–392.

Warstadt, A., et al. (2023). Findings of the BabyLM Challenge: Sample-efficient pretraining on developmentally plausible corpora. In *Proceedings of the BabyLM Challenge at CoNLL*, 1–34.

Wei, J., et al. (2022). Emergent abilities of large language models. *Transactions on Machine Learning Research*.

Wilcox, E. G., Gauthier, J., Hu, J., Qian, P., & Levy, R. (2020). On the predictive power of neural language models for human real-time comprehension behavior. In *Proceedings of CogSci*, 1707–1713.

Wilcox, E. G., Pimentel, T., Meister, C., Cotterell, R., & Levy, R. P. (2023). Testing the predictions of surprisal theory in 11 languages. *Transactions of the Association for Computational Linguistics*, 11, 1451–1470.

Wu, S., & Dredze, M. (2020). Are all languages created equal in multilingual BERT? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, 120–130.

Xu, F., & Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychological Review*, 114(2), 245–272.

Zaghouani, W., & Charfi, A. (2018). AraP-Tweet: A large multi-dialect twitter corpus for gender, age and language variety identification. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.