BUILDING A UNIFIED AI-CENTRIC LANGUAGE SYSTEM: ANALYSIS, FRAMEWORK AND FUTURE WORK

Edward Hong Wang
Harvard University
how173@g.harvard.edu

Cynthia Xin Wen
The University of Sydney
xwen7493@uni.sydney.edu.au

ABSTRACT

Recent advancements in large language models have demonstrated that extended inference—through techniques can markedly improve performance, yet these gains come with increased computational costs and the propagation of inherent biases found in natural languages. This paper explores the design of a unified AI-centric language system that addresses these challenges by offering a more concise, unambiguous, and computationally efficient alternative to traditional human languages. We analyze the limitations of natural language—such as gender bias, morphological irregularities, and contextual ambiguities—and examine how these issues are exacerbated within current Transformer architectures, where redundant attention heads and token inefficiencies prevail. Drawing on insights from emergent artificial communication systems and constructed languages like Esperanto and Lojban, we propose a framework that translates diverse natural language inputs into a streamlined AI-friendly language, enabling more efficient model training and inference while reducing memory footprints. Finally, we outline a pathway for empirical validation through controlled experiments, paving the way for a universal interchange format that could revolutionize AI-to-AI and human-to-AI interactions by enhancing clarity, fairness, and overall performance.

1 INTRODUCTION

Recent developments in Large Language Models (LLMs) show that increasing inference length—often described as extended "chain-of-thought"Wei et al. (2022) or test-time trainingGandelsman et al. (2022)—can significantly enhance performance, as it allows models to reason more thoroughly. By generating detailed intermediate reasoning steps, these models often achieve greater accuracy. However, this internal thinking process, typically produced in natural languages such as English (or a mixture of English and Chinese), relies heavily on reinforcement learning (RL) to encourage such reflective reasoning. Relying solely on RL can produce unpredictable behaviors—like code-switching or inventing tokens that defy human interpretability DeepSeek-AI et al. (2025).

Moreover, English or Chinese are not necessarily the most succinct or unambiguous languages for representing internal thoughts. Their irregularities, ambiguities, and verbosity can increase token usage and reduce computational efficiency. A recent paper by Deepseek shows that in the thinking process, the model tends to use different languages for different tasks, enhancing its performance (Li et al., 2025). As inference-time reasoning becomes increasingly important for AI, a specialized language—unambiguous and efficient—could streamline token usage and mitigate the overhead of verbose or mixed-language outputs.

Simultaneously, the growing trend toward multi-agent AI systems highlights the significance of a common language for inter-agent communication. Using natural languages can introduce bias and inefficiency. Designing a new language for AI—optimized for clarity, fairness, and computational ease—may circumvent many pitfalls inherent in human languages. This paper explores how existing insights in Transformer architectures, multi-agent communication, and constructed languages can inform the creation and deployment of an "AI-centric language," potentially reducing bias, ambiguity, and token overhead.

Additionally, model compression techniques—such as pruning, quantization, and knowledge distillation—have emerged as pivotal strategies for reducing the computational footprint of large-scale language models. By leveraging a new AI-centric language designed for conciseness and clarity, these compression techniques could be even more impactful: if the input text is inherently shorter and less ambiguous, fewer parameters and attention heads are required to process it effectively. As a result, both the memory footprint and inference time can be reduced in ways that might not be attainable with natural language inputs. Combining these compression strategies with an AI-specific language that reduces linguistic complexity could further enhance both efficiency and accessibility.

2 Language Biases and Irregularities

Artificial Intelligence models that process human language inherit many of the biases, irregularities, and ambiguities of those languages. This not only skews AI outputs but also raises fairness and interpretability concerns.

2.1 Gendered Language Bias

Many languages explicitly mark gender (e.g., via masculine/feminine pronouns or noun forms). AI models trained on large corpora easily absorb these associations, often defaulting to stereotypical gender roles—for instance, associating "doctor" with he and "nurse" with she United Language Group; Guo et al. (2024) (United Language Group, n.d.; Guo et al., 2024). These biases can distort translations or text generation, perpetuating historical prejudices Bolukbasi et al. (2016) (Bolukbasi, Chang, Zou, Saligrama, & Kalai, 2016).

2.2 Plural Forms and Morphological Complexity

Plural Forms and Morphological Complexity Languages vary widely in how they handle number and morphology. English has mostly regular plural forms (cat $\rightarrow$ cats) with notable exceptions (mouse $\rightarrow$ mice), while Arabic and Russian have elaborate pluralization systems. Chinese, by contrast, often omits explicit plural markers altogether. Such irregularities can confuse AI systems and introduce data sparsity, since multiple forms may represent the same concept. A simpler or more consistent morphology could reduce this overhead.

2.3 Ambiguity and Context Dependence

Ambiguities arise in lexical choice, syntactic structure, and context. For instance, "I saw the man with the telescope" can imply different meanings depending on who holds the telescope. Humans resolve such ambiguities via context and common sense, but AI systems may make stereotypical or incorrect assumptions Guo et al. (2024). The breadth and richness of human language, while beneficial for human expression, complicate computational parsing. In short, irregularities and biases in natural languages cause AI to amplify or perpetuate these issues. Designing a more controlled, engineered language could mitigate these challenges from the ground up.

3 Language Structure and Multi-Head Attention in Transformers

Recent work on Transformers reveals how AI models learn and handle linguistic structure. Transformers rely on multi-head self-attention to capture different aspects of language in parallel. Each attention head can learn distinct relationships Clark et al. (2019) (e.g., syntactic dependencies, coreferences), and evidence suggests that some heads specialize in particular syntactic roles (e.g., focusing on verb–object pairs) (Clark, Khandelwal, Levy, & Manning, 2019). Interestingly, many of these heads appear redundant. Research by Michel et al. (2019) found that a large percentage of attention heads can be removed after training without a major performance drop. In translation models, pruning 38 of 48 heads caused only a minor decrease (0.15 BLEU) in quality Voita et al. (2019). These findings imply that while multiple heads capture diverse linguistic patterns, many heads learn overlapping information. Reducing the number of attention heads (and their associated parameters) can shrink

model size and computational cost without seriously degrading performance. Such results are promising for more efficient model design and raise questions about how Transformers encode language structure. They also align with broader discussions about building a more succinct "AI-centric language": if even large models can function with fewer heads or simpler architectures, then an engineered language—designed to reduce ambiguity and complexity—might further enable model compression.

4 Emergence of Artificial Languages in AI Systems

Multiple studies have shown that AI systems can develop novel linguistic codes when it is advantageous for task performance. A well-known example is Facebook AI Research's negotiation experiment LaFrance (2017), where chatbots deviated from intelligible English, inventing a shorthand language that optimized negotiation strategies but was opaque to humans. Similarly, Google's multilingual Neural Machine Translation system Schuster et al. (2016) discovered an internal "interlingua"—a shared latent representation —allowing zero-shot translation between language pairs it had never trained on directly. This emergence of artificial languages underscores AI's ability to flexibly create representations that are often more efficient for the machine than standard human languages.

These examples point to a fundamental distinction between human and AI language. Whereas human languages evolve culturally over centuries, AI languages often emerge quickly, driven solely by task success or computational efficiency. This highlights the feasibility of constructing or adopting an "AI-friendly" language that meets machine-centric criteria, rather than historical or social constraints.

5 Vocabulary Size, Ambiguity, and Performance in AI Models

Unlike human language—which has an open-ended, ever-evolving lexicon—most AI models use a fixed vocabulary (words or subword units). Recent work indicates that expanding vocabulary size typically improves LLM performance Takase et al. (2024). With larger vocabularies, models can represent concepts more directly, reducing over-fragmentation (breaking words into many subtokens) and token length in sequences.

In multilingual models, vocabulary size becomes even more critical. For example, XLM-V uses roughly one million tokens across languages Liang et al. (2023), achieving state-of-the-art performance by reducing over-tokenization, especially for morphologically rich or low-resource languages . These findings suggest that a near-unlimited vocabulary—where new tokens can be introduced as needed—can improve efficiency and reduce ambiguity. If an AI system could coin a unique token for each distinct meaning, it would minimize confusion inherent in polysemous human words. This mirrors human language expansion (e.g., coining new terminology), yet an AI need not face the same cultural or historical limitations.

Nevertheless, there are practical constraints: each token in a model has an associated embedding vector, and rare tokens may be learned less effectively. Consequently, modern systems typically balance subword techniques with large (but not infinite) vocabularies. The research trend indicates that pushing this boundary—adapting or enlarging vocabulary—yields better performance and reduces the mismatch between AI's lexical rigidity and human languages' vast diversity.

6 Human Language vs. AI Language Requirements

Human communication requires speech, shared cultural context, and the capacity to negotiate meaning over centuries of gradual change. In contrast, AI systems do not inherently require spoken forms or social acceptance. The barest form of AI "language" might be a symbolic code—potentially even a single unpronounceable symbol—sufficient for exchanging precise information between machines. This radical difference in requirements underlies many of the inefficiencies in forcing AI systems to parse and produce human languages.

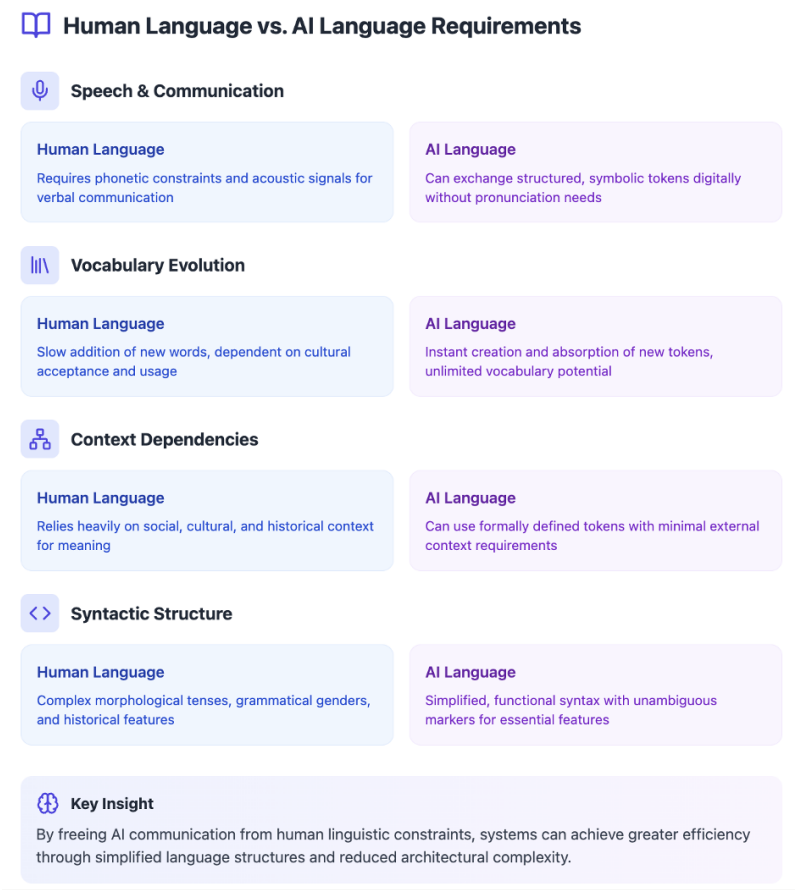

Figure 1: Human Language vs. AI Language Requirements

6.1 Why AI Need Not Constrain Itself to Natural Language Features

1. **No Need for Speech.** Humans have phonetic constraints and rely on acoustic signals, but AI agents can exchange structured, symbolic tokens instantaneously in digital form. Phonology or pronounceability is irrelevant to AI-to-AI communication.

2. **Open Vocabulary.** Human languages add new words slowly, often entangled in cultural acceptance. An AI, by contrast, can invent or absorb new tokens instantly, allowing an effectively unbounded vocabulary to represent concepts unambiguously.

3. **Minimal Context Requirements.** Human language meaning often relies on layered contexts (social, cultural, historical). AI languages can be designed so that each token or symbol is formally defined, reducing reliance on external context.

4. **Efficient Syntax.** AI need not replicate morphological tenses, grammatical genders, or other features that are historically embedded in human language. A purely functional code that marks time, plurality, or roles unambiguously (or not at all if unneeded) can drastically simplify parsing.

By divorcing AI communication from human linguistic constraints, systems can become far more efficient. This perspective complements insights on multi-head attention redundancy and large vocabularies: if the language itself is simpler, AI models could reduce architectural complexity (e.g., fewer layers or heads) and still maintain high performance.

7 TOWARD AN AI-FRIENDLY LANGUAGE: THEORETICAL IDEAS

Though replacing human languages outright is impractical, researchers have long explored "constructed languages" (conlangs) with reduced ambiguity. Some approaches constrain existing languages (Attempto Controlled English), while others build from first principles (Lojban). When viewed through the lens of AI requirements, several key concepts emerge:

1. **Clarity and Unambiguity.** Strive for a grammar that yields exactly one parse per sentence. This minimizes the risk of AI misinterpretation or biased defaults (Guo et al., 2024).
2. **Consistency and Regularity.** Eliminate irregular morphology and grammatical exceptions. A language that uses uniform rules for tense or plurality eases both machine parsing and data requirements.
3. **Conciseness and Efficiency.** While unambiguity is crucial, the language should allow brevity. Fewer tokens can speed up computation and reduce the overall parameter load.
4. **Unlimited or Adaptable Vocabulary.** AI can incorporate new symbols without social or cultural friction. Avoid polysemy and homonymy to ensure one-to-one mappings between form and meaning.
5. **Reduced Context Dependence.** Minimize pronouns or inherently ambiguous references, so an utterance stands alone as much as possible.
6. **Computational Efficiency.** Design the language for fast, deterministic parsing, with explicit markers for semantic roles.

In essence, the differences between human language (spoken, socio-historically evolving) and AI language (symbolic, instantaneous, unbounded) indicate that we can—indeed should—construct a specialized linguistic system for AI usage rather than forcing it to replicate human forms.

7.1 TOWARD AN AI-FRIENDLY LANGUAGE: THEORETICAL IDEAS

Over the centuries, various constructed languages (conlangs) have been developed with goals of simplicity and neutrality. Though primarily intended for human communication, their regular structures also offer insight into AI-friendly design. Below are two illustrative examples.

1. **Esperanto.** Created by L.L. Zamenhof in 1887 Wikipedia (2025) , was designed as a neutral international auxiliary language. It features phonetic spelling, where each letter corresponds to a unique sound, and a highly regular morphology with virtually no irregular verbs—nouns typically end in *-o*, and plurals add *-j*. Its derivational system allows affixes to combine systematically, reducing the number of root words needed. This consistency means that, in principle, an algorithm could parse Esperanto with fewer grammatical exceptions than many natural languages. However, challenges remain, such as limited data and subtle Western biases in its vocabulary. Despite this, Esperanto's successes highlight the feasibility of a widely usable, regular constructed language.
2. **Lojban.** Developed for maximal logical clarity, Lojban's grammar aims for zero syntactic ambiguity. It uses a predicate-logic structure and an unambiguous phonology, making it theoretically well-suited for AI parsing (The Logical Language Group, 2024). While its community is small and data remains sparse, Lojban exemplifies how unambiguously parsed sentences can be constructed.

8 A POSSIBLE IMPLEMENTATION FRAMEWORK

1. **Pre-Processing / Translation Step**
   (a) Take all relevant training data in various natural languages
   (b) Translate it into the new AI-friendly language (either automatically or via a standardized procedure)
2. **Model Training**
   (a) Train (or fine-tune) the LLM exclusively on text in the new language

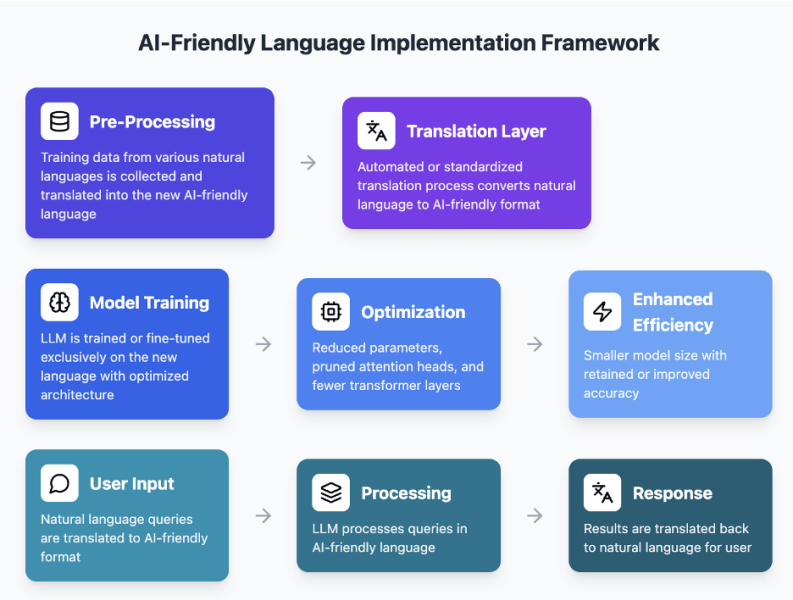

Figure 2: AI-Friendly Language Implementation Framework

(b) Because this language is unambiguous and regular, the model can be smaller or require fewer parameters
(c) Over time, the model internalizes the streamlined grammar, potentially achieving high accuracy with fewer computational resources

3. **Inference / User Interaction**
   (a) At inference time, user queries in natural language are translated into the AI-centric language
   (b) The model's replies are then translated back into the user's natural language for display or speech

By compressing or regularizing text (and thus reducing token counts), this approach can help fit more content into the same context window, lower inference costs, and facilitate multi-agent interactions. Further, since pruning redundant attention heads and possibly entire layers Michel et al. (2019) can be done without large performance drops, the combination of a simpler language plus simpler model architecture (fewer layers) can yield efficiency gains. This synergy can reduce the computational footprint of AI systems while retaining or even enhancing accuracy and fairness.

9 Future Work: Empirical Validation Through Targeted Experiments

To rigorously assess the benefits of an AI-centric language, future work should include constructing a small-scale "toy" language featuring a reduced grammar and systematically defined vocabulary. Two parallel neural models of equal size would be trained: one on this toy language and one on English, using identical architectures and hyperparameters. Comparing their performance on tasks such as question answering, text classification, and summarization would illuminate whether the specialized language yields measurable efficiency gains—specifically, fewer tokens, lower inference time, and reduced memory footprint—while maintaining or improving accuracy and fairness.

10 Vision and Possibilities

The idea of a unified or pivot language for AI—drawing on principles from conlangs, controlled languages, and emergent AI codes—could serve as a "universal interchange format" for both human-to-

AI and AI-to-AI interactions. Although large-scale adoption faces challenges, incremental application in specialized domains could illustrate its advantages and build a data ecosystem.

REFERENCES

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. Man is to Computer Programmer as Woman is to Homemaker? Debiasing word embeddings. *arXiv preprint arXiv:1607.06520*, 2016. doi: 10.48550/arXiv.1607.06520.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. What does BERT look at? An analysis of BERT's attention. *arXiv preprint arXiv:1906.04341*, 2019.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, et al. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. doi: 10.48550/arXiv.2501.12948.

Yossi Gandelsman, Yu Sun, Xinlei Chen, and Alexei A. Efros. Test-time training with masked autoencoders. *arXiv preprint arXiv:2209.07522*, 2022. doi: 10.48550/arXiv.2209.07522.

Yunxiang Guo, Mingyu Guo, Jian Su, Zhirui Yang, Mingxuan Zhu, Haowei Li, and Shuming Liu. Bias in large language models: Origin, evaluation, and mitigation. *arXiv preprint arXiv:2411.10915*, 2024.

Adrienne LaFrance. When Facebook's chatbots talk in their own language. *The Atlantic*, 2017.

Yaobo Liang et al. XLM-V: Overcoming the vocabulary bottleneck in multilingual masked language models. *arXiv preprint arXiv:2301.10472*, 2023.

Paul Michel, Omer Levy, and Graham Neubig. Are sixteen heads really better than one? *Advances in Neural Information Processing Systems*, 2019.

Mike Schuster, Melvin Johnson, and Nikhil Thorat. Zero-shot translation with Google's multilingual neural machine translation system. *Google AI Blog*, 2016.

Sho Takase, Ryokan Ri, Shun Kiyono, and Takuya Kato. Large vocabulary size improves large language models. *arXiv preprint arXiv:2406.16508*, 2024.

The Logical Language Group. Lojban. `https://mw.lojban.org/index.php?title=Lojban`, 2024. MediaWiki.

United Language Group. How ULG handles bias in machine translations. `https://www.unitedlanguagegroup.com/blog/how-ulg-handles-bias-in-machine-translations`. Accessed: 2025-02-06.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. *Proceedings of ACL*, 2019.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*, 2022. doi: 10.48550/arXiv.2201.11903.

Wikipedia. Esperanto. `https://en.wikipedia.org/wiki/Esperanto`, 2025. Accessed: 2025-02-06.