Title:
Hypothesis: Reducing Vocabulary Complexity in Large Language Models via Simple-Word Compound Substitution

Jubo Zhang
digbigAI, Independent Research Lab, Newark,CA
jerryzhang668@gmail.com

Abstract:
Large Language Models (LLMs) exhibit strong performance across a range of language tasks, but their extensive vocabulary sizes—often exceeding 100,000 tokens—contribute significantly to computational and memory costs. This paper explores a hypothesis: that replacing complex or low-frequency words with semantically equivalent compounds made from a fixed set of common words may reduce vocabulary size while preserving or even enhancing expressivity. By limiting the core vocabulary to around 20,000 frequently used words and constructing compounds from them, it may be possible to build more efficient, interpretable, and generalizable LLMs. While this idea remains untested, we outline its potential benefits, implementation strategies, and the challenges that must be addressed in future empirical studies.

1. Introduction

Large Language Models (LLMs) such as GPT and PaLM typically rely on expansive vocabularies, sometimes comprising more than 100,000 tokens. These large vocabularies improve the ability to model linguistic nuances but also contribute significantly to resource consumption, especially in embedding and softmax layers [1]. In this paper, we propose a hypothetical strategy to reduce vocabulary complexity by replacing complex words with simple-word compounds formed from a small, fixed vocabulary.

This approach aligns with the observation that human language, despite its vastness, relies on the combinatorial power of a relatively smaller set of simpler, high-frequency used words [2]. For instance, “innovate” could potentially be replaced by “create new,” or “photosynthesis” by “plant light make energy.” We hypothesize that using these compound representations could reduce vocabulary size without critically degrading performance, and may even lead to improved generalization or interpretability.

2. Background and Motivation

Several established areas of research inspire this hypothesis:

- **Subword Tokenization** methods like Byte-Pair Encoding (BPE) reduce vocabulary size by breaking words into common substrings, but often lose semantic transparency [3].
- **Controlled Natural Languages** have explored restricted vocabularies to improve accessibility and comprehension, showing that complex meaning can still be conveyed with limited vocabulary [4].
- **Semantic Decomposition** in linguistics posits that word meanings can often be expressed through smaller conceptual units [5].

This paper extends these ideas by proposing a semantic-level simplification approach that has not yet been validated through experiments but holds potential for investigation.

3. Hypothesized Method

3.1 Vocabulary Restriction

We hypothesize that a fixed core vocabulary of ~20,000 high-frequency English words could serve as the foundation for all input and output expressions in a language model. Words outside this set would be substituted by compounds of core words [6].

3.2 Compound Substitution Mechanism

Complex words could be decomposed using tools such as:

- Lexical databases (e.g., WordNet) [7]
- Semantic similarity models [8]
- Manually curated mappings or paraphrase corpora [9]

Examples of hypothesized substitutions:

Complex Word	Hypothetical Compound
Skeptical	doubt feeling

Photosynthesis	plant light energy make
Innovate	create new idea

This transformation could be applied before tokenization (during input processing) and reversed during output generation.

---

4. Hypothesized Benefits

If successful, the method may offer several benefits:

- **Reduced Model Size:** Smaller vocabulary means smaller embedding layers [1].
- **Faster Training and Inference:** Less computational overhead per token [1].
- **Lower Memory Footprint:** Enables LLMs to run on limited-resource devices [6].
- **Improved Generalization:** Models may develop more compositional reasoning over compound phrases [10].
- **Enhanced Multilingual Capabilities:** Shared concepts can make cross-lingual extensions more tractable [11].

These benefits remain theoretical and require validation through empirical testing.

---

5. Open Challenges

There are several open challenges and limitations associated with this hypothesis:

- **Semantic Ambiguity:** Simple compounds may be too imprecise [12].
- **Compositional Demands:** Models must robustly interpret compound meanings [13].
- **Identification and Substitution at Scale:** Requires scalable, accurate substitution pipelines [8].

- **Longer Sequence Lengths:** More tokens per idea may increase attention cost [14].
- **Reconstruction Difficulty:** Converting compound tokens back to fluent natural language can be awkward or ambiguous [12].

Any practical implementation must carefully address these issues, possibly through dynamic context-aware substitutions or hybrid systems.

---

6. Research Agenda

Future work is needed to evaluate this hypothesis rigorously. Suggested steps include:

1. **Dataset Creation:** Corpora with aligned complex-to-compound translations.
2. **Model Prototyping:** Small-scale LLMs using only a 20k vocabulary.
3. **Evaluation:** Benchmarks comparing fluency and generalization.
4. **Human Judgment Studies:** Assess output interpretability and clarity.

---

7. Conclusion

This paper presents a hypothesis that LLMs could operate more efficiently and transparently by replacing complex vocabulary with simple-word compounds drawn from a small, fixed vocabulary. While promising in theory, this method has yet to be validated experimentally. We invite further exploration into the linguistic, computational, and cognitive implications of vocabulary simplification for large-scale language models.

References

[1] J. Devlin, M. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL*, 2019.

[2] G. Zipf. Human Behavior and the Principle of Least Effort. *Addison-Wesley*, 1949.

[3] R. Sennrich, B. Haddow, and A. Birch. Neural Machine Translation of Rare Words with Subword Units. In *ACL*, 2016.

[4] A. Ogden. Basic English: A General Introduction with Rules and Grammar. *Paul Treber*, 1930.

[5] A. Wierzbicka. Semantics: Primes and Universals. *Oxford University Press*, 1996.

[6] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient Estimation of Word Representations in Vector Space. *arXiv:1301.3781*, 2013.

[7] G. A. Miller. WordNet: A Lexical Database for English. *Communications of the ACM*, 38(11):39–41, 1995.

[8] D. Cer, Y. Yang, S.-Y. Kong, et al. Universal Sentence Encoder. *arXiv:1803.11175*, 2018.

[9] E. Pavlick et al. The Paraphrase Database 2.0. In *LREC*, 2015.

[10] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein. Learning to Compose Neural Networks for Question Answering. In *NAACL*, 2016.

[11] P. Schwenk and M. Douze. Learning Joint Multilingual Sentence Representations with Neural Machine Translation. In *ACL*, 2017.

[12] D. Jurafsky and J. H. Martin. Speech and Language Processing, 3rd ed. *Prentice Hall*, 2023.

[13] B. Lake and M. Baroni. Generalization without Systematicity: A Challenge for Emergent Compositionality. In *ICLR*, 2018.

[14] A. Vaswani et al. Attention is All You Need. In *NeurIPS*, 2017.