Canonical Vocabulary Compression: A Synonym-Mapping Approach to LLM Training Efficiency

Table of Contents

Author: Theodore Tennant

Date Published: November 1, 2025

Canonical vocabulary compression presents a novel proposed paradigm for LLM efficiency by preprocessing both training data and user inputs to eliminate lexical redundancy. This paper outlines a comprehensive methodology for collapsing synonym variants into canonical representations, hypothesizing improvements in accuracy, fine-tuning efficiency, and inference robustness. The approach directly addresses the vocabulary bottleneck identified in recent multilingual language model research while introducing challenges in meaning preservation and stylistic variation that require careful mitigation and extensive empirical validation.

1. Introduction

Modern LLMs process vocabularies containing 50,000+ tokens, with substantial redundancy from synonymous expressions. Words like "big," "large," "huge," and "enormous" occupy separate embedding spaces despite overlapping semantics. This redundancy increases model size, training time, and potential for inconsistent representations, creating a vocabulary bottleneck that limits both efficiency and performance.

Traditional approaches address vocabulary size through subword tokenization (BPE, WordPiece) or parameter sharing, but these methods don't explicitly target semantic redundancy. We propose canonical vocabulary compression (CVC), a novel paradigm that preprocesses both training data and user inputs by mapping synonym clusters to single representatives, forcing models to learn unified representations for semantically equivalent terms while ensuring all runtime inputs conform to the model's learned vocabulary.

1.1 Motivation

The hypothesis underlying CVC is that eliminating lexical variation at the input stage reduces the model's burden of learning that "big" and "large" are functionally equivalent. By collapsing synonyms before training and applying the same transformation at inference, we theoretically achieve: reduced vocabulary size (fewer unique tokens decrease embedding matrix dimensions), improved semantic consistency (models learn stronger representations for canonical forms), higher task accuracy (concentrated training on canonical forms reduces confusion between synonymous variants), enhanced fine-tuning efficiency (gradient updates concentrate on fewer parameters with stronger signals), robust inference (input normalization ensures models never encounter unfamiliar synonyms), and faster convergence (less redundant information may accelerate learning).

Important Note: These benefits represent theoretical predictions based on the proposed methodology. Empirical testing across diverse benchmarks, model architectures, and task types is essential to validate whether this paradigm delivers measurable improvements in real-world applications.

2. Methodology

2.1 Stage 1: Building the Synonym-to-Canonical Mapping

Construct a bidirectional dictionary mapping synonyms to chosen canonical forms. Sources include WordNet [1] which provides synonym sets (synsets) organized by sense, with frequency analysis to identify the most common variant as the canonical representative, and expert curation for critical terms. The mapping establishes relationships like "large," "huge," "enormous," and "gigantic" all pointing to "big" as the canonical form.

2.2 Stage 2: Scoring Meaning Retention

Not all synonym substitutions preserve meaning equally. Implement validation metrics using BERTScore [2] to compare embeddings of original and transformed sentences, computing F1 scores between contextualized token representations. Sentence-BERT [3] provides cosine similarity measurements between sentence embeddings before and after substitution with thresholds set at 0.90 for acceptable transformations. Human evaluation samples transformed sentences for judgment on meaning preservation, establishing ground truth for automated metrics. Context-sensitive filtering rejects substitutions that significantly alter meaning for polysemous terms.

2.3 Stage 3: Preprocessing Training Data

Transform training corpora through systematic synonym replacement by tokenizing text using standard NLP pipelines, checking each token against the synonym dictionary and replacing with canonical form if present, maintaining grammatical structure and morphological variants, applying meaning retention scores to validate transformations, and tracking replacement frequency and vocabulary reduction metrics.

2.4 Stage 4: Training with Input Normalization

Train models on the preprocessed corpus with runtime input handling. Build tokenizer vocabulary from transformed text, excluding non-canonical synonyms. Use standard next-token prediction with cross-entropy loss on preprocessed data. Critically, implement inference-time input normalization: when users submit queries containing words the model never encountered during training (non-canonical synonyms), automatically transform the input using the same synonym-to-canonical mappings before processing. This ensures zero out-of-vocabulary synonyms (all variants map to known canonical forms), consistent model behavior (model always receives inputs in expected canonical vocabulary), transparent preprocessing (users can type naturally using any synonym variant), and guaranteed comprehension (model understands all synonym variants through mapping).

Example Inference Pipeline:

```
User input: "Show me enormous buildings"
→ Preprocessing: "Show me big buildings"
→ Model processing: Handles familiar canonical form
→ Output generation: Natural response
```

3. Theoretical Advantages

3.1 Vocabulary Compression

If a typical LLM vocabulary contains 50,000 tokens with approximately 20% semantic redundancy, CVC could reduce effective vocabulary to 40,000 tokens, yielding embedding matrix reduction from (50000 × d) to (40000 × d) parameters (where d is embedding dimension), faster inference through smaller softmax over vocabulary at output layer, and reduced memory footprint for deployment. These compression benefits must be empirically measured against actual vocabulary statistics from representative corpora to determine realistic reduction rates.

3.2 Improved Task Accuracy

CVC hypothetically enhances model accuracy through concentrated training signal. When training examples containing "big," "large," "huge," and "enormous" all present as "big," the model receives 4× the training signal for this semantic concept, producing stronger embeddings with lower variance, reduced confusion where model doesn't waste capacity learning subtle distinctions between true synonyms, and better generalization where unified understanding transfers more reliably to downstream tasks. Elimination of synonym ambiguity prevents sentiment classifiers from learning different weights for "big" vs. "huge" despite identical semantic contribution. Focused attention mechanisms concentrate on semantic distinctions rather than distributing attention across lexical variations.

We hypothesize that CVC models will achieve 2-5% higher accuracy on classification tasks and 3-7% improvement in semantic similarity tasks compared to baseline models with equivalent parameter counts. These predictions require rigorous testing across multiple benchmarks to validate.

3.3 Enhanced Fine-Tuning Efficiency

CVC could dramatically improve fine-tuning characteristics through gradient concentration. During fine-tuning, gradient updates for synonym-related concepts concentrate on single canonical embeddings rather than dispersing across multiple synonym tokens, producing faster convergence (fewer steps required to adapt embeddings to new domain), stronger updates (higher magnitude gradients for canonical forms), and better stability (reduced gradient variance across training batches).

Lower data requirements emerge because each canonical form receives concentrated updates; a concept appearing 100 times across various synonyms provides unified signal equivalent to 100 occurrences of one form rather than 20-30 occurrences each of multiple forms. Reduced overfitting risk occurs with smaller effective vocabulary meaning fewer parameters to overfit during fine-tuning on small datasets, particularly valuable for specialized domains. Pre-trained canonical representations may transfer more effectively because they're more densely trained and less sensitive to lexical variation in target domains.

We expect CVC models to achieve target fine-tuning performance 30-50% faster (in terms of training steps) and require 20-30% less fine-tuning data compared to standard models. Controlled experiments are necessary to confirm these efficiency gains.

3.4 Addressing the Vocabulary Bottleneck

Recent research has identified the vocabulary bottleneck as a fundamental limitation in multilingual and large-scale language models [4]. CVC addresses this bottleneck from a different angle than existing approaches by reducing semantic redundancy rather than optimizing tokenization strategies. By collapsing synonym variants, we potentially reduce the effective vocabulary size while maintaining or improving semantic coverage.

3.5 Semantic Density and Inference Robustness

By forcing models to encounter only canonical forms during training, we increase frequency of canonical representations. A word appearing 10,000 times across synonyms now appears as one canonical form 10,000 times, producing stronger embeddings with more training signal, better generalization with unified understanding of semantic concepts, and reduced ambiguity through clearer distinction between truly different concepts. The bidirectional application of CVC (training preprocessing + input normalization) ensures models handle vocabulary diversity without performance degradation, guaranteeing comprehension through preprocessing that maps all variants to canonical forms, maintaining consistent performance regardless of user's lexical choices, and eliminating synonym confusion where model behavior remains identical whether users type "big," "large," or "enormous."

4. Implementation Details

4.1 Input Preprocessing Pipeline

The inference-time normalization system operates as follows: tokenize user input into tokens, query the synonym-to-canonical mapping dictionary for each token, replace tokens existing in the mapping with canonical forms while preserving originals otherwise, reassemble normalized text maintaining original structure, and feed normalized input to LLM for generation. Preprocessing adds minimal latency (approximately 1-2ms for typical queries), dictionary lookup achieves O(1) complexity with hash table implementation, can be implemented as lightweight preprocessing layer, and requires no model architecture changes. Maintaining cache of previously normalized queries eliminates repeated preprocessing overhead for common inputs.

5. Challenges and Limitations

5.1 Polysemy and Context Sensitivity

Many synonyms aren't truly interchangeable. "Big" and "large" may be equivalent for size, but "big" carries additional connotations ("big deal," "big brother"). Context-insensitive replacement risks meaning distortion in idiomatic expressions becoming nonsensical, register mismatches in formal contexts requiring formal vocabulary, and connotation loss where subtle emotional or cultural associations disappear. Mitigation requires implementing context-aware replacement using dependency parsing or sense disambiguation algorithms while maintaining whitelists of protected phrases.

5.2 Loss of Stylistic Variation

Human language uses synonyms for rhetorical effect, emphasis, and stylistic diversity. Collapsing vocabulary may produce repetitive generation with models outputting monotonous text lacking lexical variety, reduced expressiveness preventing capture of nuanced distinctions between near-synonyms, and style degradation where generated text feels mechanical or unnatural. Mitigation involves hybrid approaches where core training uses canonical forms but fine-tuning stages reintroduce controlled variation for style.

5.3 Granularity Selection

Determining appropriate semantic granularity for synonym clusters is non-trivial. Over-aggressive clustering maps semantically distinct terms (e.g., "warm" and "hot") losing important distinctions. Under-clustering retains too many variants defeating the purpose. Domain-specific requirements mean technical domains may need finer distinctions than general language. Mitigation uses hierarchical clustering with multiple granularity levels allowing domain-specific tuning.

5.4 Validation Requirements

Critical Testing Needed: The proposed CVC paradigm must undergo extensive empirical validation including accuracy measurements on standard benchmarks (GLUE, SuperGLUE), fine-tuning efficiency studies with controlled data budgets, generalization testing on out-of-domain tasks, cross-lingual applicability assessments, human evaluation of generation quality, and ablation studies on synonym cluster granularity.

6. Experimental Validation Framework

Note: The following framework represents a proposed experimental protocol. Actual implementation and results are required to validate the CVC paradigm.

6.1 Baseline Comparison and Evaluation Metrics

Train identical architectures on original vs. preprocessed corpora, control for training steps, batch sizes, and hyperparameters, and measure both standard and CVC models on diverse synonym-containing inputs. Evaluation metrics include classification accuracy comparing performance on tasks with synonym-rich inputs, semantic similarity measuring consistency of embeddings for synonym variants, robustness testing evaluating performance degradation with non-canonical vocabulary, convergence speed measuring steps required to reach target validation performance, data efficiency showing minimum dataset size needed for successful adaptation,

stability metrics tracking gradient variance and training stability, perplexity measuring language modeling performance, GLUE/SuperGLUE scores for natural language understanding, generation quality through human evaluation of coherence and naturalness, inference speed measuring tokens processed per second including preprocessing overhead, and memory usage tracking model size and runtime memory requirements.

6.2 Ablation Studies and Statistical Rigor

Vary synonym cluster granularity, test different canonical selection strategies, compare meaning retention thresholds, evaluate impact of input normalization vs. no normalization, and measure accuracy gains from vocabulary concentration. All comparative results must include multiple random seeds for training stability, confidence intervals for performance metrics, statistical significance tests (t-tests, bootstrap), and analysis of variance across different model sizes and architectures.

7. Related Work

Vocabulary Pruning prunes rare tokens to reduce vocabulary size but doesn't target semantic redundancy or provide input normalization. Knowledge Distillation [5] compresses models through distillation while maintaining full vocabulary. Subword Tokenization using BPE [6] and WordPiece [7] reduces vocabulary through character-level decomposition without explicitly handling synonymy. Semantic Embeddings from Word2Vec [8] and GloVe [9] learn that synonyms are similar but don't collapse representations. Input Normalization through spelling correction and text normalization preprocesses inputs without systematically addressing synonym variation.

Vocabulary Bottleneck Research [4] on multilingual models identifies vocabulary limitations as key constraints on model performance, exploring solutions through vocabulary expansion rather than compression. CVC differs by combining preprocessing of both training data and inference inputs, creating end-to-end vocabulary consistency while maintaining natural user interaction. However, unlike existing validated approaches, CVC remains a proposed paradigm requiring empirical confirmation of its theoretical benefits.

8. Future Directions

Multilingual extension applies CVC across languages, mapping cross-lingual synonyms to shared canonical forms improving multilingual models by reducing language-specific redundancy. Dynamic vocabulary develops adaptive systems adjusting canonical mappings based on domain or task requirements while maintaining flexibility. Generative expansion trains separate modules converting canonical outputs back to stylistically appropriate synonyms during generation, preserving compression benefits during processing while maintaining output diversity. Integration with other compression combines CVC with quantization, pruning, and distillation achieving multiplicative compression effects. Domain-specific optimization creates specialized synonym mappings for technical domains (medical, legal, scientific) where terminology precision matters but synonym variation persists.

9. Conclusion

Canonical vocabulary compression presents a novel proposed paradigm for LLM efficiency by preprocessing both training data and user inputs to eliminate lexical redundancy. The bidirectional application of synonym-to-canonical mapping ensures models never encounter unfamiliar vocabulary variants while concentrating training and fine-tuning signals on canonical forms. This paper outlines a theoretical framework with hypothesized advantages; comprehensive empirical testing is essential to determine whether these benefits materialize in practice.

The proposed accuracy improvements stem from eliminating synonym-based confusion and concentrating model capacity on semantic distinctions rather than lexical variations. Fine-tuning benefits would arise from gradient concentration, reduced data requirements, and more stable optimization dynamics. The input normalization layer ensures these advantages extend to real-world deployment where users employ diverse vocabulary.

While the methodology faces challenges in preserving context-dependent meaning and stylistic variation, careful implementation of context-sensitive replacement and appropriate granularity selection could mitigate these concerns. The approach merits rigorous empirical investigation to quantify accuracy gains, fine-tuning speedups, and overall trade-offs across diverse tasks and model architectures.

Research Imperative: Future work must focus on large-scale empirical validation across multiple benchmarks, comparative studies against existing compression techniques, human evaluation of generation quality and naturalness, robustness testing across diverse domains and languages, and computational cost-benefit analysis. As LLMs continue growing in size and scope, techniques that reduce unnecessary complexity while potentially improving performance become increasingly valuable. CVC represents a promising but unproven direction that requires systematic investigation to determine its practical viability. Only through comprehensive testing can we establish whether this paradigm delivers measurable improvements or whether the theoretical benefits are outweighed by practical limitations. The field needs controlled experiments with statistical rigor to move this proposal from hypothesis to validated methodology.

References

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

[2] Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2020). BERTScore: Evaluating Text Generation with BERT. International Conference on Learning Representations (ICLR).

[3] Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 3982-3992.

[4] Liang, D., Gonen, H., Mao, Y., Hou, R., Goyal, N., Ghazvininejad, M., Schwartz, R., & Zettlemoyer, L. (2023). XLM-V: Overcoming the Vocabulary Bottleneck in Multilingual Masked Language Models. Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pp. 13996-14018.

[5] Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the Knowledge in a Neural Network. arXiv preprint arXiv:1503.02531.

[6] Sennrich, R., Haddow, B., & Birch, A. (2016). Neural Machine Translation of Rare Words with Subword Units. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 1715-1725.

[7] Schuster, M., & Nakajima, K. (2012). Japanese and Korean Voice Search. IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 5149-5152.

[8] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. arXiv preprint arXiv:1301.3781.

[9] Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global Vectors for Word Representation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532-1543.

This paper presents a proposed methodology requiring empirical validation. The authors encourage the research community to test these hypotheses and share findings to advance understanding of vocabulary optimization strategies for large language models.