Novelty Assessment Report

Paper: StochasTok: Improving Fine-Grained Subword Understanding in LLMs
PDF URL: https://openreview.net/pdf?id=gqCh1k0CEX
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2025-12-29

Abstract

Subword-level understanding is integral to numerous tasks, including understanding multi-digit numbers, spelling mistakes, abbreviations, rhyming, and wordplay. Despite this, current large language models (LLMs) still struggle disproportionally with seemingly simple subword-level tasks, like counting the number of 'r's in 'strawberry'. A key factor behind these failures is tokenization, which obscures the fine-grained structure of words. Current alternatives, such as character-level and dropout tokenization methods, significantly increase computational costs and provide inconsistent improvements. In this paper, we revisit tokenization and introduce StochasTok, a simple, efficient stochastic tokenization scheme that randomly splits tokens during training, allowing LLMs to 'see' their internal structure. Our experiments show that pretraining with StochasTok substantially improves LLMs' downstream performance across multiple subword-level language games, including character counting, substring identification, and math tasks. Furthermore, StochasTok's simplicity allows seamless integration at any stage of the training pipeline, and we demonstrate that post-training with StochasTok can instill improved subword understanding into existing pretrained models, thus avoiding costly pretraining from scratch. These dramatic improvements achieved with a minimal change suggest StochasTok holds exciting potential when applied to larger, more capable models.

Core Task Landscape

This paper addresses: **improving subword-level understanding in large language models**

A total of **50 papers** were analyzed and organized into a taxonomy with **16 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Tokenization Architecture and Representation Design**
- **Subword Tokenization Analysis and Evaluation**
- **Token-Level Optimization and Manipulation**
- **Token-Level Phenomena and Theoretical Foundations**

Complete Taxonomy Tree

- improving subword-level understanding in large language models Survey Taxonomy
- Tokenization Architecture and Representation Design
  - Character-Level and Byte-Level Modeling (3 papers)
  - [10] SpaceByte: Towards Deleting Tokenization from Large Language Modeling (Slagle, 2024) View paper
  - [28] T-FREE: Subword tokenizer-free generative LLMs via sparse representations for memory-efficient embeddings (BjÃ¶rn Deiseroth, 2024) View paper
  - [35] Word-level representation from bytes for language modeling (Lee, 2022) View paper
  - Hierarchical and Hybrid Tokenization Architectures (4 papers)
  - [8] Hierarchical Autoregressive Transformers: Combining Byte- and Word-Level Processing for Robust, Adaptable Language Models (Neitemeier, 2025) View paper
  - [18] Bridging the Gap between Subword and Character Segmentation in Pretrained Language Models (Shun Kiyono, 2023) View paper
  - [26] Dynamic token hierarchies: Enhancing large language models with a multi-tiered token processing framework (David Barbere, 2024) View paper
  - [39] Retrofitting Large Language Models with Dynamic Tokenization (Darius Feher, 2025) View paper
  - Stochastic and Dynamic Tokenization Methods ★ (1 papers)
  - [0] StochasTok: Improving Fine-Grained Subword Understanding in LLMs (Anon et al., 2026) View paper
  - Continuous and Embedding-Based Representations (2 papers)
  - [34] Llm pretraining with continuous concepts (Tack, 2025) View paper
  - [48] Auto-encoding morph-tokens for multimodal llm (Pan, 2024) View paper
- Subword Tokenization Analysis and Evaluation
  - Tokenization Impact on Model Capabilities (4 papers)
  - [2] The Strawberry Problem: Emergence of Character-level Understanding in Tokenized Language Models (Cosma, 2025) View paper
  - [3] Tokenization falling short: On subword robustness in large language models (Yekun Chai, 2024) View paper
  - [7] Enhancing llm character-level manipulation via divide and conquer (Xiong Zhen, 2025) View paper
  - [49] Tokenization Constraints in LLMs: A Study of Symbolic and Arithmetic Reasoning Limits (Zhang Xiang, 2025) View paper
  - Semantic and Linguistic Properties of Tokens (5 papers)
  - [6] How much semantic information is available in large language model tokens? (David A. Haslett, 2025) View paper
  - [19] From tokens to words: On the inner lexicon of LLMs (G.J. Kaplan, 2024) View paper

- [24] Tomato, Tomahto, Tomate: Do Multilingual Language Models Understand Based on Subword-Level Semantic Concepts? (Crystina Zhang, 2025) View paper
- [37] Understanding Subword Compositionality of Large Language Models (Peng Qiwei, 2025) View paper
- [47] Subword segmentation in LLMs: Looking at inflection and consistency (Marion Di Marco, 2024) View paper
- Tokenization Algorithm Comparison and Design (5 papers)
- [9] Effects of sub-word segmentation on performance of transformer language models (Jue Hou, 2023) View paper
- [16] Towards Linguistically-Aware and Language-Independent Tokenization for Large Language Models (LLMs) (Rahman, 2024) View paper
- [30] Evaluating subword tokenization: Alien subword composition and oov generalization challenge (Batsuren, 2024) View paper
- [41] Improving subword embeddings in large language models using morphological information (Lauren, 2024) View paper
- [45] Achieving Tokenizer Flexibility in Language Models through Heuristic Adaptation and Supertoken Learning (Shaurya Sharthak, 2025) View paper
- Tokenization Effects on Specialized Domains (2 papers)
- [4] Investigating Hierarchical Term Relationships in Large Language Models (Cai Guohui, 2025) View paper
- [43] Representation learning of structured data for medical foundation models (Dwivedi, 2024) View paper

- Token-Level Optimization and Manipulation
  - Token Pruning and Compression (6 papers)
  - [1] Tokenskip: Controllable chain-of-thought compression in llms (Xia, 2025) View paper
  - [17] VQToken: Neural Discrete Token Representation Learning for Extreme Token Reduction in Video Large Language Models (Zhang Haichao, 2025) View paper
  - [21] Dynamic token pruning for LLMs: leveraging task-specific attention and adaptive thresholds (Seyed Hossein Ahmadpanah, 2025) View paper
  - [22] {JENGA}: Enhancing {LLM}{Long-Context} Fine-tuning with Contextual Token Sparsity (T Wang, 2025) View paper
  - [31] Large language models with adaptive token fusion: A novel approach to reducing hallucinations and improving inference efficiency (Yuan-Yuan Fang, 2024) View paper
  - [36] Improve the accuracy and efficiency of large language models via dynamic token compression and adaptive layer pruning (Fabian Potkins, 2024) View paper
  - Token-Level Reward and Attention Mechanisms (3 papers)
  - [12] TGDPO: Harnessing Token-Level Reward Guidance for Enhancing Direct Preference Optimization (Zhu, 2025) View paper
  - [14] Critical Tokens Matter: Token-Level Contrastive Estimation Enhances LLM's Reasoning Capability (Lin Zi-cheng, 2024) View paper
  - [32] Question Tokens Deserve More Attention: Enhancing Large Language Models without Training through Step-by-Step Reading and Question Attention Recalibration (F Han, 2025) View paper
  - Token Representation and Semantic Analysis (5 papers)
  - [15] Interpreting token compositionality in LLMs: A robustness analysis (Nura Aljaafari, 2024) View paper
  - [25] Contextual gradient recomposition for sequential coherence preservation in large language model token generation (Annabel, 2025) View paper
  - [27] Semantic latency mapping of contextual vector embeddings in transformer-based models (Serena Bouzina, 2024) View paper
  - [33] Frame representation hypothesis: Multi-token llm interpretability and concept-guided text generation (Pedro H. V. Valois, 2025) View paper
  - [38] Token-level optimization for enhanced text generation: A prompt engineering framework with large language models (Damien Lococ, 2024) View paper
  - Token-Level Safety and Quality Control (3 papers)
  - [11] Toxic Subword Pruning for Dialogue Response Generation on Large Language Models (Lu, 2024) View paper
  - [20] Mitigating heterogeneous token overfitting in llm knowledge editing (Liu Tian-ci, 2025) View paper
  - [44] Token-Level Uncertainty Estimation for Large Language Model Reasoning (Shi, 2025) View paper

- Token-Level Phenomena and Theoretical Foundations
  - Duplicate and Near-Duplicate Token Effects (1 papers)
  - [13] On the effect of (near) duplicate subwords in language modelling (Hofmann, 2024) View paper
  - Information-Theoretic and Representational Foundations (2 papers)
  - [5] Stochastic resonance pathways for latent knowledge reassembly in large language models (Allan, 2025) View paper
  - [29] Forget BIT, It is All about TOKEN: Towards Semantic Information Theory for LLMs (Bai, 2025) View paper
  - Constrained Generation and Decoding Challenges (2 papers)
  - [40] Pitfalls, subtleties, and techniques in automata-based subword-level constrained generation (M Cognetta, 2025) View paper
  - [46] EFIM: Efficient Serving of LLMs for Infilling Tasks with Improved KV Cache Reuse (Guo Tianyu, 2025) View paper
  - Cross-Domain and Multimodal Token Applications (3 papers)
  - [23] Self-supervised user embedding alignment for cross-domain recommendations via multi-LLM co-training (Haotian Lyu, 2025) View paper
  - [42] Silent grammars in emergent language models: An exploratory study of latent instructional drift via stochastic scaffold morphogenesis (R Cattermole, 2025) View paper
  - [50] Efficiently Learning an Encoder That Classifies Token Replacements and Masked Permuted Network-Based BIGRU Attention Classifier for Enhancing Sentiment â¦ (MIU Haq, 2024) View paper

Narrative

Core task: improving subword-level understanding in large language models. The field addresses fundamental challenges in how LLMs process and represent text at the subword level, organizing research into four main branches. Tokenization Architecture and Representation Design explores novel ways to construct and organize token representations, including stochastic and dynamic methods that move beyond fixed vocabularies (e.g., StochasTok[0], Tokenskip[1]) as well as hierarchical and byte-level approaches (SpaceByte[10], Hierarchical Autoregressive Transformers[8]). Subword Tokenization Analysis and Evaluation investigates how existing tokenization schemes affect model behavior, examining issues like the "strawberry problem" (Strawberry Problem[2]) and compositional failures (Tokenization Falling Short[3], Subword Compositionality Understanding[37]). Token-Level Optimization and Manipulation focuses on techniques for pruning, compressing, or selectively modifying tokens during inference or training (Dynamic Token Pruning[21], Toxic Subword Pruning[11]). Token-Level Phenomena and Theoretical Foundations studies the underlying mechanisms and theoretical properties of subword processing, including attention patterns and semantic information flow.

Several active lines of work reveal key trade-offs in the field. One tension involves balancing flexibility with computational efficiency: dynamic tokenization methods promise better adaptability to diverse inputs but introduce overhead, while fixed vocabularies remain efficient yet brittle on edge cases. Another contrast appears between analysis-focused studies that diagnose tokenization failures (Tokenization Falling Short[3], Duplicate Subwords Effect[13]) and intervention-focused work that proposes architectural solutions. StochasTok[0] sits within the stochastic and dynamic tokenization cluster, emphasizing probabilistic token selection as a way to improve robustness. Compared to deterministic dynamic approaches like Tokenskip[1], which selectively skips tokens, StochasTok[0] introduces randomness to explore multiple segmentation possibilities. This positions it as exploring a middle ground between fully fixed tokenization and more radical architectural redesigns, aiming to enhance subword understanding through controlled stochasticity rather than complete vocabulary overhaul.

Related Works in Same Category

No sibling papers were found in the same taxonomy leaf. A taxonomy-subtopic-level comparison will be produced instead.

Taxonomy-Level Summary

The original leaf focuses on introducing randomness or adaptivity into tokenization processes to expose internal word structure, while its siblings explore alternative approaches to subword-level understanding. The siblings either bypass traditional tokenization entirely (character/byte-level, continuous representations) or combine multiple tokenization strategies (hierarchical/hybrid). Together, these subtopics represent different architectural choices for handling subword information in language models.

Similarities: - All subtopics aim to improve subword-level understanding beyond standard static tokenization - Each approach seeks to address limitations of fixed vocabulary tokenization schemes - All methods involve modifications to how models process linguistic units below the word level - Each subtopic excludes standard static tokenization and post-hoc token manipulation

Differences: - Stochastic/Dynamic methods modify existing tokenization with randomness/adaptivity, while Character/Byte-Level eliminates subword tokenization entirely - Continuous Representations replace discrete tokens with continuous embeddings, whereas Stochastic/Dynamic maintains discrete tokens with variable processing - Hierarchical/Hybrid combines multiple granularities simultaneously, while Stochastic/Dynamic focuses on single-level adaptive tokenization - Character/Byte-Level operates at the finest granularity, Continuous at the most abstract, Hierarchical/Hybrid at multiple levels, and Stochastic/Dynamic at variable subword levels - Stochastic/Dynamic introduces variability during training/inference, while Hierarchical/Hybrid uses fixed multi-level architectures

Suggested Search Directions: - Hybrid approaches combining stochastic tokenization with character-level fallback mechanisms - Dynamic tokenization methods that adapt granularity based on continuous representation similarity - Comparative studies on when randomness in tokenization versus hierarchical processing is more effective - Methods that use stochastic tokenization to learn better continuous representations

Sibling Subtopics

- **Character-Level and Byte-Level Modeling** (leaves: 1, papers: 3)
- Scope: Approaches that eliminate or reduce reliance on subword tokenization by operating at character or byte granularity.
- Exclude: Excludes hybrid methods that combine character and subword processing; see Hierarchical and Hybrid Tokenization Architectures.
- **Continuous and Embedding-Based Representations** (leaves: 1, papers: 2)
- Scope: Methods replacing discrete tokens with continuous concept representations or learned embeddings to bypass tokenization constraints.
- Exclude: Excludes discrete token modifications and standard embedding improvements; see Token-Level Optimization and Manipulation or Subword Tokenization Analysis and Evaluation.
- **Hierarchical and Hybrid Tokenization Architectures** (leaves: 1, papers: 4)
- Scope: Systems combining multiple granularities of tokenization or processing levels to balance efficiency and fine-grained understanding.
- Exclude: Excludes pure character-level or pure subword methods; see Character-Level and Byte-Level Modeling or Subword Tokenization Analysis and Evaluation.

Contributions Analysis

Overall novelty summary. The paper introduces StochasTok, a stochastic tokenization scheme that randomly splits tokens during training to expose internal word structure. Within the taxonomy, it occupies the 'Stochastic and Dynamic Tokenization Methods' leaf under 'Tokenization Architecture and Representation Design'. Notably, this leaf contains only one paper (the original work itself), indicating a relatively sparse research direction. The broader parent category includes four leaves addressing alternative tokenization architectures, suggesting that stochastic approaches represent a less-explored avenue compared to character-level modeling or hierarchical designs.

The taxonomy reveals neighboring work in character-level modeling (three papers), hierarchical architectures (four papers), and continuous representations (two papers). StochasTok diverges from these by maintaining subword tokenization while introducing controlled randomness, rather than abandoning tokens entirely or layering multiple granularities. The 'Subword Tokenization Analysis and Evaluation' branch (seventeen papers across four leaves) documents extensive empirical work on tokenization failures—the 'strawberry problem' and compositional breakdowns—that motivate StochasTok's design. The taxonomy's scope notes clarify that stochastic methods differ from post-hoc token manipulation (excluded to 'Token-Level Optimization') and static schemes (excluded to 'Subword Tokenization Analysis').

Among twenty-nine candidates examined, the contribution-level analysis shows varied novelty signals. The core StochasTok scheme (Contribution A) examined nine candidates with zero refutations, suggesting limited direct prior work on stochastic token splitting. The pretraining demonstration (Contribution B) examined ten candidates, also with zero refutations, indicating that empirical validation of stochastic tokenization on subword tasks appears underexplored. However, the post-training application (Contribution C) examined ten candidates and found one refutable match, suggesting that adapting pretrained models via tokenization modifications has some precedent in the limited search scope.

Based on the top-29 semantic matches and taxonomy structure, StochasTok appears to occupy a relatively novel position within stochastic tokenization methods. The single-paper leaf and low refutation rates across contributions suggest limited direct overlap, though the analysis does not cover exhaustive literature beyond these candidates. The taxonomy context indicates that while tokenization challenges are well-documented (seventeen analysis papers), stochastic solutions remain less developed compared to architectural alternatives like character-level or hierarchical approaches.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: StochasTok: A simple stochastic tokenization scheme

Description: The authors propose StochasTok, a stochastic tokenization method that randomly splits tokens into equivalent pairs of smaller tokens during training. This approach allows language models to observe the fine-grained morphological structure of words, improving subword-level understanding while maintaining compatibility with any base tokenizer.

This contribution was assessed against **9 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Distributional properties of subword regularization

URL: View paper

Brief Assessment

Subword Regularization Properties[51] analyzes distributional properties of existing stochastic tokenization methods (BPE-dropout, MaxMatch-dropout) and proposes uniform sampling from all possible tokenizations. StochasTok uses a different approach: randomly splitting tokens into equivalent pairs during training, which is architecturally distinct from the uniform sampling strategy analyzed in the candidate paper.

2. Linguistic features tokenization of text corpora of the Uzbek

URL: View paper

Brief Assessment

Uzbek Tokenization[58] focuses on linguistic feature tokenization for Uzbek language corpora using rule-based matching and machine learning. This is fundamentally different from StochasTok's approach of randomly splitting tokens during training to improve subword understanding in LLMs.

3. Self-supervision through random segments with autoregressive coding (randsac)

URL: View paper

Brief Assessment

Randsac[54] focuses on autoregressive visual representation learning through random segment prediction in vision transformers, not on stochastic tokenization for language models. The candidate operates on image patches/segments for self-supervised visual pretraining, while the original proposes a text tokenization method that randomly splits word tokens during language model training.

4. Improving Consistency in LLM Inference using Probabilistic Tokenization

URL: View paper

Brief Assessment

Probabilistic Tokenization Consistency[56] focuses on using probabilistic tokenizations to improve self-consistency in reasoning tasks for already-trained LLMs, not on a training-time stochastic tokenization scheme like StochasTok that randomly splits tokens during pretraining.

5. Stochastic tokenization with a language model for neural text classification

URL: View paper

Brief Assessment

Stochastic Tokenization[53] focuses on sampling different tokenizations during training for text classification in unsegmented languages (Chinese/Japanese), using a nested unigram language model. The original paper's StochasTok randomly splits existing tokens into equivalent pairs during training for any language, compatible with any base tokenizer. These are fundamentally different stochastic tokenization approaches with different mechanisms and goals.

6. Optimizing Biomedical Text Processing: A Comparative Analysis of Tokenization Methods and Context-Aware Representation Learning

URL: View paper

Brief Assessment

Biomedical Text Processing[55] focuses on biomedical domain-specific text processing and representation learning methods, not on general stochastic tokenization schemes for improving subword understanding in language models.

7. Improving Self Consistency in LLMs through Probabilistic Tokenization

URL: View paper

Brief Assessment

Self Consistency Probabilistic[57] focuses on using probabilistic tokenization to improve self-consistency in reasoning tasks through multiple tokenizations at inference time, while the original paper proposes StochasTok for training-time token splitting to improve subword understanding. These are fundamentally different applications and mechanisms of stochastic tokenization.

8. A Spitting Image: Superpixel Transformers

URL: View paper

Brief Assessment

Superpixel Transformers[59] focuses on vision transformers with superpixel-based spatial tokenization for images, not language model tokenization. The candidate addresses visual tokenization (partitioning images into irregular superpixels), while the original addresses text tokenization (splitting word tokens into subword units).

9. Magnet: Improving the multilingual fairness of language models with adaptive gradient-based tokenization

URL: View paper

Brief Assessment

Magnet[52] focuses on multilingual fairness through script-specific boundary predictors for equitable segmentation across languages, not on general stochastic tokenization for subword structure understanding. The technical approaches differ fundamentally: StochasTok randomly splits tokens into equivalent pairs during training, while Magnet routes sequences through language-script-specific predictors optimized with different compression rates.

Contribution 2: Demonstration of improved subword understanding through pretraining

Description: The authors demonstrate that pretraining language models with StochasTok leads to substantial improvements on various subword-level tasks such as character counting, substring identification, and multi-digit addition. Models pretrained with StochasTok achieve near-perfect accuracy on language game tasks and can grok mathematical operations.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. LUNA: language understanding with number augmentations on transformers via number plugins and pre-training

URL: View paper

Brief Assessment

LUNA[77] focuses on number understanding in transformers through number-specific tokenization and embeddings, not general subword-level tasks like character counting or substring identification that StochasTok addresses.

---

2. Language models trained to do arithmetic predict human risky and intertemporal choice

URL: View paper

Brief Assessment

Arithmetic Predict Choice[69] focuses on training language models on arithmetic tasks to predict human decision-making in risky and intertemporal choice, not on improving subword-level understanding through tokenization methods.

---

3. A survey of word embeddings based on deep learning

URL: View paper

Brief Assessment

Word Embeddings Survey[71] discusses general word embedding techniques and sub-word elements in language models, but does not address the specific pretraining methodology (StochasTok) or the particular subword-level tasks (character counting, substring identification, multi-digit addition) that are central to the original paper's contribution.

---

4. Neural Networks for Mathematical Reasoningâ□□Evaluations, Capabilities, and Techniques

URL: View paper

Brief Assessment

Mathematical Reasoning Survey[73] is a thesis focused on neural networks for mathematical reasoning tasks (abduction and induction), not on tokenization methods or subword-level understanding improvements in language models.

---

5. Improving numeracy by input reframing and quantitative pre-finetuning task

URL: View paper

Brief Assessment

Numeracy Input Reframing[76] focuses on improving numerical understanding through input notation changes (digit-based, scientific notation) and a comparing-numbers pre-finetuning task, not on general subword-level tasks like character counting or substring identification that StochasTok addresses.

---

6. Enabling High-Sparsity Foundational Llama Models with Efficient Pretraining and Deployment

URL: View paper

Brief Assessment

High Sparsity Llama[72] focuses on model sparsity and computational efficiency through pruning and sparse pretraining, not on improving subword-level understanding or tokenization methods for tasks like character counting and substring identification.

---

7. Refining Pre-trained Language Models for Domain Adaptation with Entity-Aware Discriminative and Contrastive Learning

URL: View paper

Brief Assessment

Entity Aware Adaptation[75] focuses on domain adaptation with entity-aware discriminative tasks for NLP applications like named entity recognition, not on general pretraining methods for improving subword-level understanding across diverse tasks like character counting or mathematical operations.

---

8. Cross-tokenizer distillation via approximate likelihood matching

URL: View paper

Brief Assessment

Cross Tokenizer Distillation[68] focuses on transferring knowledge between models with different tokenizers through distillation methods, not on pretraining methods that improve subword-level task performance like character counting or substring identification.

---

9. Scalable Influence and Fact Tracing for Large Language Model Pretraining

URL: View paper

Brief Assessment

Influence Fact Tracing[74] focuses on training data attribution methods for identifying influential examples in LLM pretraining, not on improving subword-level understanding through tokenization methods.

---

10. Llm the genius paradox: A linguistic and math expert's struggle with simple word-based counting problems

URL: View paper

Brief Assessment

Genius Paradox[70] focuses on analyzing why LLMs fail at character counting tasks and testing various conjectures (tokenization, training data, model size), rather than proposing a pretraining method to improve subword understanding. The candidate investigates failure modes while the original proposes StochasTok as a solution.

Contribution 3: Post-training application to existing pretrained models

Description: The authors show that StochasTok can be applied after pretraining to retrofit existing models with improved subword understanding. This continued pretraining approach allows models that were originally trained with deterministic tokenization to gain subword-level capabilities without requiring expensive retraining from scratch.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Sub-character tokenization for Chinese pretrained language models

URL: View paper

Brief Assessment

Sub Character Tokenization[64] focuses on a fundamentally different approach (sub-character encoding for Chinese) and does not discuss post-training application to existing pretrained models. The candidate paper's method requires encoding text at the sub-character level before tokenization, which is incompatible with retrofitting existing models.

---

2. Word and Character Semantic Fusion by Pretrained Language Models for Text Classification

URL: View paper

Brief Assessment

Word Character Fusion[63] focuses on fusing subword and character features through dual transformer encoders during model usage, not on post-training methods to retrofit existing models with improved subword understanding.

---

3. PMANet: Malicious URL detection via post-trained language model guided multi-level feature attention network

URL: View paper

Brief Assessment

PMANet[61] focuses on post-training for domain adaptation to URLs, not on instilling subword understanding into pretrained models for general language tasks.

---

4. Fast adaptation and robust quantization of multi-modal foundation models from associative memory: A case study in speechLM

URL: View paper

Brief Assessment

Associative Memory Adaptation[66] focuses on applying outlier-efficient Hopfield layers to pretrained models for speech-text tasks, not on instilling subword understanding through stochastic tokenization methods as in the original paper.

---

5. AlephBERT: Language model pre-training and evaluation from sub-word to sentence level

URL: View paper

Brief Assessment

AlephBERT[67] focuses on morphological extraction from pretrained models for Hebrew NLP tasks, not on post-training methods to instill subword understanding. The morphological extraction component operates on top of existing embeddings rather than modifying tokenization during continued pretraining.

---

6. OFA: A Framework of Initializing Unseen Subword Embeddings for Efficient Large-scale Multilingual Continued Pretraining

URL: View paper

Brief Assessment

OFA[62] focuses on vocabulary extension and continued pretraining for multilingual adaptation, not on instilling subword understanding into existing models through stochastic tokenization methods.

---

7. Breaking the Script Barrier in Multilingual Pre-Trained Language Models with Transliteration-Based Post-Training Alignment

URL: View paper

Brief Assessment

Transliteration Alignment[65] focuses on cross-lingual transfer between languages with different scripts using transliteration, not on instilling subword understanding into models through stochastic tokenization during post-training.

---

8. Toxic Subword Pruning for Dialogue Response Generation on Large Language Models

URL: View paper

Brief Assessment

Toxic Subword Pruning[11] focuses on pruning toxic subwords from vocabulary for safety purposes, not on instilling subword understanding capabilities into pretrained models through continued pretraining.

---

9. Retrofitting Large Language Models with Dynamic Tokenization

URL: View paper

Prior Art Analysis

Dynamic Tokenization Retrofitting[39] demonstrates that their method can be applied to existing pretrained models (GPT-2 and Mistral-7b) through continued pretraining with dynamic tokenization, achieving improved subword understanding without retraining from scratch. This directly parallels StochasTok's claim of post-training application. Both papers show that a small amount of continued pretraining with their respective tokenization methods can instill subword-level capabilities into deterministically-pretrained models. The candidate paper explicitly uses the term 'retrofitting' in its title and demonstrates application to pretrained models like GPT-2 and Mistral-7b, achieving performance improvements through continued pretraining.

Evidence

Evidence 1 - **Rationale**: Both papers explicitly claim the ability to apply their tokenization methods to existing pretrained models. The candidate uses 'retrofitting' in the title and abstract, while the original uses 'post-training' - both referring to the same concept of modifying pretrained models. - **Original**: we demonstrate that post-training with stochastokcan instill improved subword understanding into existing pretrained models, thus avoiding costly pretraining from scratch. - **Candidate**: we challenge the static design and propose retrofitting lms with dynamic tokenization: a way to dynamically decide on token boundaries based on the input text via a subwordmerging algorithm inspired by byte-pair encoding.

Evidence 2 - **Rationale**: Both papers frame their contribution as avoiding expensive retraining from scratch by applying their method to already-pretrained models. - **Original**: in this section, we therefore investigate whether stochastokcan be used to instill improved subword understanding into models that have already been pretrained with an alternative tokenization method, offering a more cost-effective alternative to full retraining from scratch. - **Candidate**: to address this issue, we challenge the static design and propose retrofitting lms with dynamic tokenization

---

10. Large Language Models for Data Discovery and Integration: Challenges and Opportunities.

URL: View paper

Brief Assessment

Data Discovery Integration[60] focuses on data discovery and integration challenges using LLMs, not on tokenization methods or post-training techniques for subword understanding.

Appendix: Text Similarity Detection

No high-similarity text segments were detected across any compared papers.

References

- [0] StochasTok: Improving Fine-Grained Subword Understanding in LLMs View paper
- [1] Tokenskip: Controllable chain-of-thought compression in llms View paper
- [2] The Strawberry Problem: Emergence of Character-level Understanding in Tokenized Language Models View paper
- [3] Tokenization falling short: On subword robustness in large language models View paper
- [4] Investigating Hierarchical Term Relationships in Large Language Models View paper
- [5] Stochastic resonance pathways for latent knowledge reassembly in large language models View paper
- [6] How much semantic information is available in large language model tokens? View paper
- [7] Enhancing llm character-level manipulation via divide and conquer View paper
- [8] Hierarchical Autoregressive Transformers: Combining Byte- and Word-Level Processing for Robust, Adaptable Language Models View paper
- [9] Effects of sub-word segmentation on performance of transformer language models View paper
- [10] SpaceByte: Towards Deleting Tokenization from Large Language Modeling View paper
- [11] Toxic Subword Pruning for Dialogue Response Generation on Large Language Models View paper
- [12] TGDPO: Harnessing Token-Level Reward Guidance for Enhancing Direct Preference Optimization View paper
- [13] On the effect of (near) duplicate subwords in language modelling View paper
- [14] Critical Tokens Matter: Token-Level Contrastive Estimation Enhances LLM's Reasoning Capability View paper
- [15] Interpreting token compositionality in LLMs: A robustness analysis View paper
- [16] Towards Linguistically-Aware and Language-Independent Tokenization for Large Language Models (LLMs) View paper
- [17] VQToken: Neural Discrete Token Representation Learning for Extreme Token Reduction in Video Large Language Models View paper
- [18] Bridging the Gap between Subword and Character Segmentation in Pretrained Language Models View paper
- [19] From tokens to words: On the inner lexicon of LLMs View paper
- [20] Mitigating heterogeneous token overfitting in llm knowledge editing View paper
- [21] Dynamic token pruning for LLMs: leveraging task-specific attention and adaptive thresholds View paper
- [22] {JENGA}: Enhancing {LLM}{Long-Context} Fine-tuning with Contextual Token Sparsity View paper
- [23] Self-supervised user embedding alignment for cross-domain recommendations via multi-LLM co-training View paper
- [24] Tomato, Tomahto, Tomate: Do Multilingual Language Models Understand Based on Subword-Level Semantic Concepts? View paper
- [25] Contextual gradient recomposition for sequential coherence preservation in large language model token generation View paper
- [26] Dynamic token hierarchies: Enhancing large language models with a multi-tiered token processing framework View paper
- [27] Semantic latency mapping of contextual vector embeddings in transformer-based models View paper
- [28] T-FREE: Subword tokenizer-free generative LLMs via sparse representations for memory-efficient embeddings View paper
- [29] Forget BIT, It is All about TOKEN: Towards Semantic Information Theory for LLMs View paper
- [30] Evaluating subword tokenization: Alien subword composition and oov generalization challenge View paper
- [31] Large language models with adaptive token fusion: A novel approach to reducing hallucinations and improving inference efficiency View paper
- [32] Question Tokens Deserve More Attention: Enhancing Large Language Models without Training through Step-by-Step Reading and Question Attention Recalibration View paper
- [33] Frame representation hypothesis: Multi-token llm interpretability and concept-guided text generation View paper
- [34] Llm pretraining with continuous concepts View paper
- [35] Word-level representation from bytes for language modeling View paper
- [36] Improve the accuracy and efficiency of large language models via dynamic token compression and adaptive layer pruning View paper
- [37] Understanding Subword Compositionality of Large Language Models View paper
- [38] Token-level optimization for enhanced text generation: A prompt engineering framework with large language models View paper
- [39] Retrofitting Large Language Models with Dynamic Tokenization View paper
- [40] Pitfalls, subtleties, and techniques in automata-based subword-level constrained generation View paper
- [41] Improving subword embeddings in large language models using morphological information View paper
- [42] Silent grammars in emergent language models: An exploratory study of latent instructional drift via stochastic scaffold morphogenesis View paper
- [43] Representation learning of structured data for medical foundation models View paper
- [44] Token-Level Uncertainty Estimation for Large Language Model Reasoning View paper
- [45] Achieving Tokenizer Flexibility in Language Models through Heuristic Adaptation and Supertoken Learning View paper
- [46] EFIM: Efficient Serving of LLMs for Infilling Tasks with Improved KV Cache Reuse View paper
- [47] Subword segmentation in LLMs: Looking at inflection and consistency View paper
- [48] Auto-encoding morph-tokens for multimodal llm View paper
- [49] Tokenization Constraints in LLMs: A Study of Symbolic and Arithmetic Reasoning Limits View paper
- [50] Efficiently Learning an Encoder That Classifies Token Replacements and Masked Permuted Network-Based BIGRU Attention Classifier for Enhancing Sentiment â View paper
- [51] Distributional properties of subword regularization View paper
- [52] Magnet: Improving the multilingual fairness of language models with adaptive gradient-based tokenization View paper
- [53] Stochastic tokenization with a language model for neural text classification View paper
- [54] Self-supervision through random segments with autoregressive coding (randsac) View paper
- [55] Optimizing Biomedical Text Processing: A Comparative Analysis of Tokenization Methods and Context-Aware Representation Learning View paper
- [56] Improving Consistency in LLM Inference using Probabilistic Tokenization View paper

- [57] Improving Self Consistency in LLMs through Probabilistic Tokenization View paper
- [58] Linguistic features tokenization of text corpora of the Uzbek View paper
- [59] A Spitting Image: Superpixel Transformers View paper
- [60] Large Language Models for Data Discovery and Integration: Challenges and Opportunities. View paper
- [61] PMANet: Malicious URL detection via post-trained language model guided multi-level feature attention network View paper
- [62] OFA: A Framework of Initializing Unseen Subword Embeddings for Efficient Large-scale Multilingual Continued Pretraining View paper
- [63] Word and Character Semantic Fusion by Pretrained Language Models for Text Classification View paper
- [64] Sub-character tokenization for Chinese pretrained language models View paper
- [65] Breaking the Script Barrier in Multilingual Pre-Trained Language Models with Transliteration-Based Post-Training Alignment View paper
- [66] Fast adaptation and robust quantization of multi-modal foundation models from associative memory: A case study in speechLM View paper
- [67] AlephBERT: Language model pre-training and evaluation from sub-word to sentence level View paper
- [68] Cross-tokenizer distillation via approximate likelihood matching View paper
- [69] Language models trained to do arithmetic predict human risky and intertemporal choice View paper
- [70] Llm the genius paradox: A linguistic and math expert's struggle with simple word-based counting problems View paper
- [71] A survey of word embeddings based on deep learning View paper
- [72] Enabling High-Sparsity Foundational Llama Models with Efficient Pretraining and Deployment View paper
- [73] Neural Networks for Mathematical Reasoningâ□□Evaluations, Capabilities, and Techniques View paper
- [74] Scalable Influence and Fact Tracing for Large Language Model Pretraining View paper
- [75] Refining Pre-trained Language Models for Domain Adaptation with Entity-Aware Discriminative and Contrastive Learning View paper
- [76] Improving numeracy by input reframing and quantitative pre-finetuning task View paper
- [77] LUNA: language understanding with number augmentations on transformers via number plugins and pre-training View paper