Specialized Attention Heads in Early Word Reconstruction

Anonymous ACL submission

Abstract

Large language models process text as sequences of tokens, raising the question of how coherent representations emerge from fragmented word pieces during inference. Prior work has proposed a *detokenization* stage in early layers, where tokens are recombined into larger linguistic units, but direct intervention-based evidence for the role of specific model components remains limited. This article investigates whether a small subset of attention heads (1/3/5%) contributes disproportionately to this process. Candidate heads are identified based on their preference for within-word attention, measured using attention weights and L_2 norms, and their contribution is evaluated through targeted ablation and activation patching in a word reconstruction task. Across languages (English, Turkish, Chinese), tokenization schemes (BPE, SentencePiece, byte-level), and model families, ablating these heads generally produces greater degradation in reconstruction performance than random ablations. Results further show that larger models exhibit increased specialization, characterized by steeper and more pronounced early-layer retrieval peaks. These findings provide additional evidence for the detokenization hypothesis and suggest that early layers rely on specialized attention heads to construct internal word-level representations from fragmented inputs.

1 Introduction

Large language models (LLMs) process text as sequences of discrete tokens rather than whole words (Mielke et al., 2021), which can split meaningful units into smaller fragments. For example, the GPT-2 tokenizer represents "Hemingway" as `['␣H', 'eming', 'way']`. The choice in tokenization scheme has been shown to have impact on downstream performance, especially in lower-resource languages and languages with non-Latin scripts (Ahia et al., 2023; Petrov et al., 2023). Despite this, models remain robust to input corruption (Cao et al., 2023), raising the question of how models recover coherent semantic units. One hypothesis, often called *detokenization* (Lad et al., 2024; Kaplan et al., 2025), suggests that early layers partially reverse tokenization by combining tokens into larger segments.

In this article, we investigate whether detokenization depends on a specific subset of attention heads and how the process varies across languages (Chinese, Turkish, and English), tokenization schemes, and model sizes. We examine two dimensions: differences between otherwise comparable models using BPE-, SentencePiece-, and byte-level tokenization, and differences in scale within a single model family. Using activation patching (Vig and Belinkov, 2019; Ghandeharioun et al., 2024), we evaluate performance on a word retrieval task in which a model reconstructs a full word from the representation of its final token. Retrieval performance is analyzed across layers under original, targeted ablation (1/3/5%), and random ablation conditions. This setup allows us to assess whether detokenization relies on specialized attention heads or whether later components compensate for disruptions introduced earlier in the network. We aim to clarify how tokenizer design and multilingual input influence the formation of intermediate word-level representations in large language models.

Prior work has identified attention heads associated with both token merging and higher-level semantic copying, but has largely not established whether these heads are necessary for forming coherent internal representations or whether later components can compensate for their disruption. For example, Feucht et al. (2025) study "concept attention heads" that copy semantic units in tasks such as translation and synonymy, while other work identifies heads involved in token-level structure, such as attending within word boundaries or

merging tokens (Ferrando and Voita, 2024; Lad et al., 2024; Kamoda et al., 2025). However, the causal importance of these heads, and the extent to which their roles generalize across tokenizers, languages, and model scales, has received limited empirical attention. In this work, we address these gaps by testing whether a subset of attention heads is critical for forming internal representations and by examining how their importance varies across tokenization schemes and languages.

2 Background

Modern LLMs represent text using tokenizers to split it into tokens. Different schemes exist, with compression-based algorithms such as BPE (Sennrich et al., 2016) and SentencePiece (Kudo and Richardson, 2018) being common choices, while some approaches instead use byte-level tokenizers that operate directly on raw bytes (Xue et al., 2022). The primary motivation for subword tokenization is to mitigate the out-of-vocabulary problems of earlier word-based approaches (Mielke et al., 2021). However, this process can produce non-optimal segmentations for low-resource languages which affect downstream performance (Ahia et al., 2023), possible non-canonical tokenizations and overly eager splitting when encountering minor typos. Despite this, modern LLMs remain surprisingly robust to typological perturbations (Cao et al., 2023) and non-canonical tokenizations (Jimenez Gutierrez et al., 2023; Zheng et al., 2026).

This robustness suggests that models reconstruct larger meaningful units from fragmented subtokens during inference, motivating analysis of their internal computations. Mechanistic interpretability studies how computations arise within transformer components, with Lad et al. (2024) proposing a hierarchical structure with four stages of inference inside LLMs: **detokenization**, where subtokens merge into lexical units; *feature engineering*, which constructs intermediate semantic features; *prediction ensembling*, where candidate next-token predictions emerge; and *residual sharpening*, which concentrates probability mass onto a small set of tokens. The detokenization stage, in particular, offers an explanation for the observed robustness by recombining fragmented subtokens into coherent lexical units in the early layers of the model. This stage-based view of inference is consistent with findings that LLM representations exhibit a structured progression across layers. For example, Wendler et al. (2024); Conneau et al. (2020) showing early and late layers being language-specific, while middle layers form a more language-agnostic/English conceptual space.

These layer-wise computations arise from the transformer architecture introduced by Vaswani et al. (2017), whose central component is the self-attention mechanism. Put simply, self-attention transforms each token representation in an input sequence to encode its relationship to other tokens. In a self-attention layer, three linear projections of an input sequence with dimensionality d produce the *Query*, *Key*, and *Value* vectors. These vectors Q, K, and V are then used to calculate attention scores:

$$\text{Att}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right)V \quad (1)$$

The output is a probability distribution over tokens that sums to 1. To allow the model to capture different types of relationships in parallel, transformers use *multi-head attention*. This involves applying Equation 1 many times in parallel, forming several *attention heads* which can attend to distinct parts of the sequence. Each head is calculated as such, where W_i represents the learnable weight matrices for the i-th attention head:

$$\text{head}_i(Q, K, V) = \text{Att}(QW_i^Q, KW_i^K, VW_i^V) \quad (2)$$

Because each head computes attention patterns independently, heads can become functionally specialized. Prior work has identified heads in BERT (Devlin et al., 2019), an encoder-only transformer, associated with syntactic structure (Clark et al., 2019). Liu et al. (2026) identified heads in LLMs corresponding to specific languages, and were able to steer the model towards specific languages by adjusting their activations. There are also more studies on more task-specific behaviors such as relational reasoning (Hanna et al., 2023) and acronym detection (García-Carrasco et al., 2024). Zhou et al. (2025) show that ablating even a single head in LLaMA2-7B leads to a 16x increase in harmful responses. At the same time, many heads can be removed with limited impact on performance (Voita et al., 2019; Michel et al., 2019), suggesting that only a subset contribute to some computations.

When discussing detokenization, Lad et al. (2024) argue that it functions as a continuation of the embedding process, concatenating nearby tokens. This builds on earlier work showing positional heads mainly attend to the previous token (Correia et al., 2019; Elhage et al., 2021). Elhage et al. (2021) create toy transformer models with single layers to form what is effectively a skip trigram model, with `if [source] ... [destination] -> [out]` pairs. For example, a tokenized sequence of "Cindy" without a space `["C", "indy"]` can be mapped back to the more typical form `["␣Cindy"]` through specific attention head patterns. Extending the identification of subword merging heads by Ferrando and Voita (2024), Lad et al. (2024) find heads in Pythia 2.8B with consistent attention patterns between the final subtoken of four-token words and a single-token control.

Kaplan et al. (2025) argue that attention is the earliest component of the detokenization mechanism, marking which elements are used in later stages. Rather than identifying individual heads, they measure mean attention weights across layers between single-token words and inside multi-token words, finding higher weights inside multi-token words. To assemble these attended components into a full internal representation, Kaplan et al. (2025) draw on findings that transformer feed-forward networks act as key/value memories encoding factual and linguistic knowledge (Geva et al., 2021; Meng et al., 2022), arguing that models use attention outputs to retrieve full internal representations in early-mid layers.

3 Data

The analysis uses the Universal Dependencies PUD (UD-PUD) treebanks v2.17 (Nivre et al., 2020; Akhundjanova et al., 2025) for Chinese, English, and Turkish. UD-PUD is a multilingual parallel corpus derived from Wikipedia and news text with syntactic dependency annotations, released under CC BY-SA 3.0. Its parallel structure enables cross-lingual comparison while minimizing domain variation. Throughout the paper, *word* refers to a syntactic word as annotated in UD.

Sequences are randomly shuffled and iterated over. For each sequence, up to two candidate words are selected under several constraints. Words with part-of-speech tags that often remain unchanged across languages (punctuation, proper nouns, symbols, and numerals) are excluded to avoid trivial matches. Words occurring within the first three positions of a sequence are also removed, since limited preceding context and positional artifacts (such as first-token attention sinks (Xiao et al., 2024)) may confound the analysis. For each language, we collect 250 unique words together with their preceding context from the PUD sequence. Different random seeds are chosen for identification and ablation.

4 Methodology

We identify candidate attention heads using boundary-based attention patterns inspired by Kaplan et al. (2025); Ferrando and Voita (2024). Since tokenization efficiency differs across languages and tokenizers, we avoid binary distinctions such as single-token versus multi-token words used in e.g. Kaplan et al. (2025). Instead, we compare two conditions: (a) *in-boundary* attention from the final subtoken of a word to its penultimate subtoken, and (b) *out-boundary* attention from the first subtoken of a word to the final subtoken of the preceding word. See Figure 1 for an example.

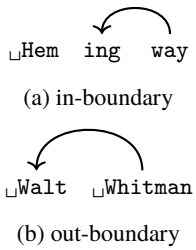


Figure 1: Examples of each condition in GPT-2: in-boundary (top) and out-boundary (bottom). Arrows denote attention from a source token to a target token.

For each layer l and head h, we then compute the mean attention difference between the candidates from each condition using:

$$\Delta_{l,h}^{(\text{attn})} = \mu_{l,h}^{(\text{in})} - \mu_{l,h}^{(\text{out})}. \quad (3)$$

Because attention weights alone may not reflect downstream contribution (Jain and Wallace, 2019; Serrano and Smith, 2019), we additionally measure the norm of each head output prior to output projection following Kobayashi et al. (2020):

$$\Delta_{l,h}^{(\text{norm})} = \mu_{l,h}^{(|\text{in}|)} - \mu_{l,h}^{(|\text{out}|)}. \quad (4)$$

Both metrics are z-normalized within each language/model configuration and summed to obtain

a final head score:

$$S_{l,h} = z_{l,h}^{(\text{attn})} + z_{l,h}^{(\text{norm})}. \quad (5)$$

To evaluate whether candidate heads contribute to word reconstruction, we combine zero-ablation (Michel et al., 2019) with Patchscopes (Ghandeharioun et al., 2024). Hidden states from a source prompt are patched into a retrieval prompt of the form:

`"Repeat the following word: X, X, X,"`

where `X` is replaced with the patched representation. Then the model generates 10 tokens using greedy decoding using the patched prompt, and it is measured whether or not the original word is in the generated output. The prompt is translated into Chinese[1] and Turkish[2] by native speakers.

We evaluate three conditions:

1. No ablation
2. Targeted ablation of the top-scoring heads under 1/3/5% ablation.
3. Random ablation of an equal number of heads from the same layers, excluding the heads selected by the corresponding targeted-ablation.

Experiments are conducted on TokSuite (Altıntas et al., 2025), a family of 1B Llama-2–style models trained using different tokenizers for a fixed token size (40B English, and 15B for Chinese, Turkish, Farsi, and Italian respectively), and Qwen2.5 models (Yang et al., 2024; Qwen et al., 2025). We evaluate GPT-2 (Radford et al., 2019), Aya (Dang et al., 2024), XGLM (Lin et al., 2022), and ByT5 (Xue et al., 2022) tokenizers to compare English-centric BPE, multilingual BPE, Unigram, and byte-level tokenization strategies in the TokSuite models. For brevity, models are referred to by tokenizer shorthand throughout the paper (e.g. the TokSuite model using the XGLM tokenizer is written as `xglm`).

Bytes per token (BPT $= \frac{\text{bytes}}{\text{tokens}}$), along with vocabulary size, are reported in Table 1 to capture tokenizer differences.

Under the detokenization hypothesis, ablating candidate heads should impair reconstruction performance more than random ablation. Candidate heads are identified separately for each language/model pair and restricted to the first fifteen layers, where detokenization is hypothesized to occur (Kaplan et al., 2025; Lad et al., 2024).

[1] `重复这个词语：X、X、X、`

[2] `Bu kelimeyi tekrarla: X, X, X,`

Tokenizer	Vocab size	BPT Score		
		English	Turkish	Chinese
`aya`	255k	4.68	4.35	4.05
`byt5`	256	1.00	1.00	1.00
`gpt2`	50k	4.63	2.30	1.36
`qwen2.5`	151k	4.52	3.17	3.67
`xglm`	256k	4.59	4.47	3.38

Table 1: Tokenizer statistics over UD-PUD: vocabulary size and byte/token (BPT) scores for English, Turkish, and Chinese. Higher BPT values indicate greater tokenization efficiency, with more bytes represented per token.

Experiments took ∼150 GPU hours to run, with the 7B models being run on an A40, the rest on V100s.

5 Results

Across nearly all models and languages, targeted ablation reduces retrieval performance more than random ablation, particularly at ablation rates of 3/5%. The strongest effect is observed for the `Qwen2.5` 7B model in English, where retrieval drops by 0.61 under 5% targeted ablation compared to random ablation. Similar patterns are observed across both TokSuite and Qwen2.5 models, suggesting that the identified heads contribute to representations involved in word reconstruction. In most cases, retrieval does not fully recover in later layers after targeted ablation, indicating that these effects are concentrated in early-stage processing. Final-layer retrieval results are summarized in Table 2 and Table 3.

The magnitude of the effect varies across tokenizers and languages. In the TokSuite models, English achieves the highest retrieval performance, followed by Turkish and Chinese. The strongest targeted-ablation effects occur for the `xglm` tokenizer in English, where retrieval drops from approximately 0.7 to 0.15 within the first quarter of layers under 5% ablation. In contrast, some tokenizer/language combinations exhibit relatively weak targeted-ablation effects. For `byt5`, targeted ablation produces little additional degradation beyond random ablation. Similarly, the `gpt2` tokenizer shows comparatively small targeted-ablation effects in Chinese and Turkish. Across Chinese and Turkish, differences between tokenizer variants are generally smaller than in English. Figure 3 presents cumulative retrieval across layers, and Figure 2a shows retrieval rates per layer.

Language	Model	Clean final ↑	Targeted Δ final			Random Δ final			Gap 5%
			1%	3%	5%	1%	3%	5%	
Chinese	aya	0.37	-0.04	-0.15	-0.18	0.00	0.02	0.00	-0.18
	gpt2	0.35	-0.01	-0.02	-0.02	0.00	0.01	-0.01	-0.01
	xglm	0.32	-0.09	-0.20	-0.21	0.00	0.01	0.01	-0.22
	byt5	0.17	-0.01	-0.01	0.00	0.00	0.00	0.01	-0.01
English	xglm	0.82	0.00	-0.21	-0.41	0.02	0.03	0.03	-0.45
	gpt2	0.45	-0.02	-0.09	-0.22	-0.01	-0.01	0.00	-0.22
	aya	0.27	0.04	-0.12	-0.22	0.00	0.00	0.00	-0.22
	byt5	0.06	-0.01	0.00	-0.04	-0.01	0.00	0.00	-0.04
Turkish	xglm	0.50	-0.02	-0.11	-0.20	0.00	0.01	0.02	-0.22
	gpt2	0.50	0.00	-0.06	-0.07	0.00	-0.01	0.00	-0.08
	aya	0.33	-0.04	-0.14	-0.21	0.00	0.01	0.00	-0.21
	byt5	0.02	0.00	-0.01	-0.01	0.00	-0.01	-0.01	0.00

Table 2: Summary of final-layer cumulative retrieval under increasing ablation strength for `toksuite`. Clean final is the cumulative retrieval rate at the final layer. Targeted and random deltas are measured relative to the clean condition. More negative values indicate stronger disruption after all layers have had the opportunity to compensate. Gap 5% is the targeted delta minus the random delta at the largest ablation level.

Language	Model	Clean final ↑	Targeted Δ final			Random Δ final			Gap 5%
			1%	3%	5%	1%	3%	5%	
Chinese	Qwen2.5-7B	0.28	-0.04	-0.19	-0.22	-0.01	-0.01	-0.01	-0.21
	Qwen2.5-1.5B	0.23	-0.03	-0.09	-0.08	0.00	-0.03	0.02	-0.10
	Qwen2.5-0.5B	0.16	-0.04	-0.13	-0.15	-0.02	-0.02	0.02	-0.17
	Qwen2.5-3B	0.15	-0.02	-0.05	-0.09	0.01	-0.01	0.02	-0.10
English	Qwen2.5-7B	0.81	-0.02	-0.74	-0.80	0.00	-0.11	-0.19	-0.61
	Qwen2.5-1.5B	0.74	-0.13	-0.34	-0.63	-0.02	-0.08	-0.18	-0.45
	Qwen2.5-3B	0.67	-0.19	-0.46	-0.53	0.03	-0.04	-0.07	-0.46
	Qwen2.5-0.5B	0.12	-0.04	-0.12	-0.12	0.00	0.01	0.01	-0.13
Turkish	Qwen2.5-7B	0.28	-0.02	-0.24	-0.27	-0.02	-0.07	-0.06	-0.21
	Qwen2.5-3B	0.16	-0.05	-0.07	-0.12	-0.02	-0.01	-0.03	-0.09
	Qwen2.5-1.5B	0.10	0.00	-0.07	-0.10	0.00	0.00	-0.04	-0.06
	Qwen2.5-0.5B	0.01	-0.01	0.00	0.00	0.00	0.00	0.01	-0.01

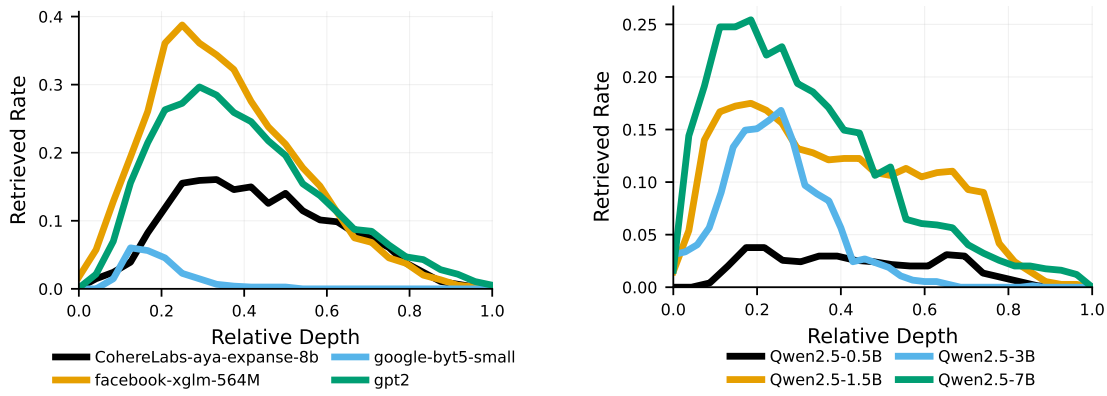
Table 3: Summary of final-layer cumulative retrieval under increasing ablation strength for `qwen2.5`. Clean final is the cumulative retrieval rate at the final layer. Targeted and random deltas are measured relative to the clean condition. More negative values indicate stronger disruption after all layers have had the opportunity to compensate. Gap 5% is the targeted delta minus the random delta at the largest ablation level.

The Qwen2.5 models exhibit clearer layer specialization as parameter size increases. The 7B and 3B models show pronounced early-layer retrieval peaks, as shown in Figure 2b, whereas smaller models distribute retrieval more evenly across layers. Targeted ablation effects also increase with model size: in the 7B model, retrieval drops sharply under 3/5% targeted ablation, while random ablation produces smaller reductions. Although random ablation has a larger effect in Qwen2.5 than in TokSuite overall, targeted ablation consistently remains more disruptive. Layerwise retrieval curves are shown in Figure 4.

6 Discussion

Targeted ablation generally causes larger retrieval degradation than random ablation, suggesting that the candidate-selection procedure successfully identifies retrieval-relevant attention heads. The effect is strongest in the TokSuite models under 5% ablation, while the `Qwen2.5` models show similar degradation between 3/5%. 1% ablation degradation is minimal in most cases. While all models are able to recover slightly from targeted ablation, they are not able to fully recover.

However, for some cases the difference between random/targeted ablation *increases* between 3% and 5% ablation. These findings suggests that the targeted candidate set captures many high-impact



(a) `toksuite` (b) `qwen2.5`

Figure 2: Retrieved rate over relative depth for each model family, pooled across languages.

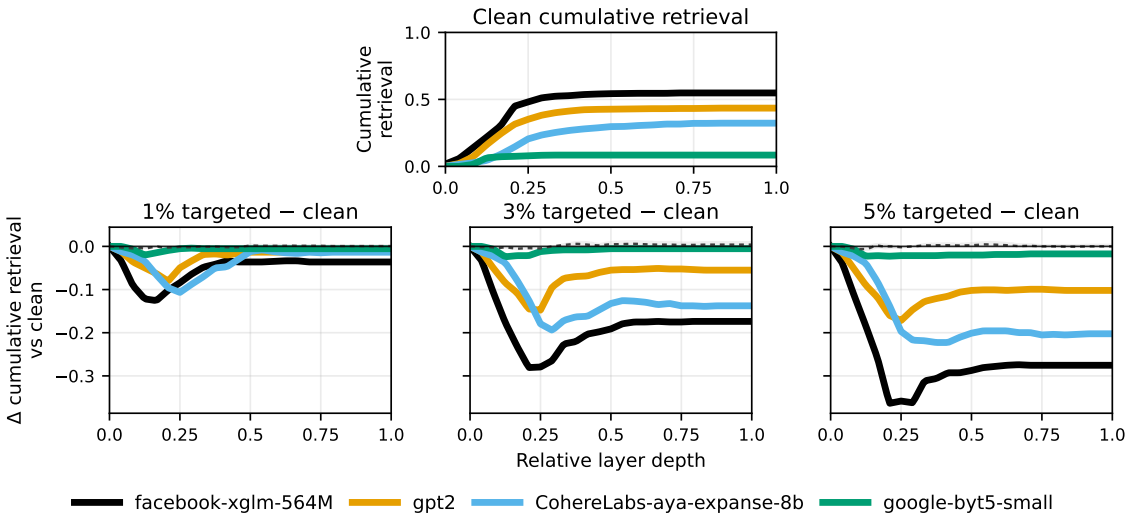


Figure 3: Cumulative retrieval for Toksuite under targeted ablation, pooled across languages. Top: clean retrieval. Bottom: targeted minus clean at increasing ablation levels. Colored lines show targeted ablation; dotted gray and shading show the random-ablation median and interquartile range.

heads, but not all of them. This would also be consistent with the limited impact of 1% ablation in many cases. More granular candidate-selection strategies may improve coverage. For example, Ferrando and Costa-Jussà (2024) identify candidate heads through logit-difference analysis between corrupted and clean runs combined with exhaustive sweeps over attention heads and components. Although substantially more computationally expensive, similar approaches may better recover detokenization-relevant heads that are not strongly expressed through token-boundary attention patterns alone.

The impact of tokenizer choice can be partly explained by differences in byte efficiency and the resulting training dataset size in TokSuite models. Because all models are trained on a fixed number of tokens, tokenizers with lower byte efficiency rates expose the model to less raw text. As shown in Table 1, `byt5` operates on 1 byte per token, meaning that it is effectively trained on less data than more efficient tokenizers such as `xglm` which gets ∼4 bytes per token. This reduced exposure may limit the model's ability to learn specialized attention patterns. Instead of relying on a small set of heads to reconstruct word-level representations, the model may develop more distributed mechanisms for combining token information, reducing the impact of targeted ablation.

However, byte efficiency alone does not explain the observed differences. For example, the `gpt2` tokenizer achieves a byte per token close to `byt5`

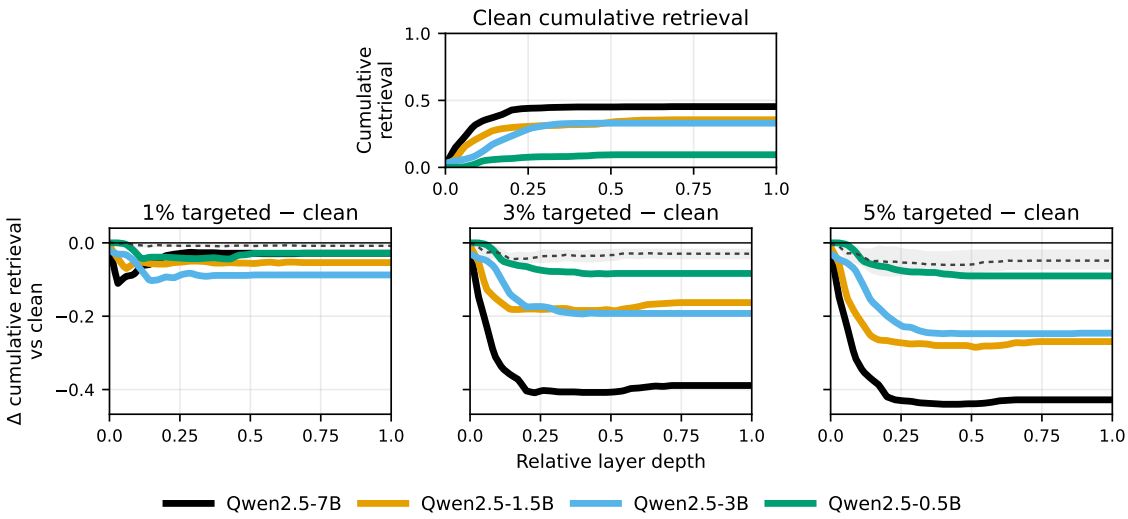


Figure 4: Cumulative retrieval for `qwen2.5` under targeted ablation, pooled across languages. Top: clean retrieval. Bottom: targeted minus clean at increasing ablation levels. Colored lines show targeted ablation; dotted gray and shading show the random-ablation median and interquartile range.

for Chinese (1.36 vs. 1.0), yet attains higher retrieval performance (0.35 vs. 0.17). Similarly, the `xglm` tokenizer markedly outperforms `gpt2` in English (0.82 vs. 0.45) despite slightly worse byte efficiency. Vocabulary size is also insufficient to explain the gap: `aya` and `xglm` have nearly identical vocabularies (256k vs. 255k) but exhibit different retrieval rates. These comparisons indicate that factors beyond byte efficiency, training dataset size, and vocabulary size play a role.

Cross-lingual transfer may be one such factor. Prior work shows that monolingual BERT models can learn abstractions that generalize across languages (Artetxe et al., 2020), and that embedding alignment correlates with transfer (Deshpande et al., 2022). The `gpt2`-based model may therefore benefit from representations learned from other languages when processing Chinese/Turkish, helping compensate for less efficient tokenization. In contrast, the explicitly multilingual `aya` tokenizer performs worse than `gpt2` in both English and Turkish despite more efficient tokenization. This is consistent with the "curse of multilinguality" (Chang et al., 2024): given fixed model capacity, increasing the number of languages can *reduce* per-language performance due to competition for shared representational resources. This apparent contradiction, that the less efficient `gpt2` tokenizer nevertheless yields stronger robustness and performance than explicitly multilingual tokenizers, may reflect the benefits of vocabulary overlap identified by (Kallini et al., 2025), who show that partially shared vocabularies can improve cross-lingual transfer relative to fully disjoint vocabularies, even in the presence of multilingual capacity trade-offs.

Tokenizer architecture itself may also play a role. The `xglm` tokenizer achieves the highest retrieval performance in English, outperforming even larger models such as 7B `Qwen2.5`, despite slightly worse byte efficiency than `gpt2`. This suggests that differences in tokenization strategy (e.g., Unigram vs. BPE) may influence how effectively embeddings and intermediate representations are learned. While prior work emphasizes the dominant role of the embedding layer as compared to the tokenizer itself (Wu et al., 2023), these results indicate that tokenizer design can still shape how efficiently embeddings are learned during training.

Retrieval performance generally increases with parameter size, with the 0.5B model generally performing worst and the 7B model performing best within the `Qwen2.5` family. This is consistent with prior findings regarding how embeddings emerge during training: embeddings of high-frequency words converge quickly during training, while low-frequency words take longer and retain alignment with the bias in their random initializations (Papadimitriou and Prince, 2025). However, this relationship does not strictly hold in the results in comparison to parameter size. The 1.5B model slightly outperforms the 3B model in English and Chinese. This suggests that increasing parameter size does not uniformly improve representation

quality across configurations. One interpretation is that scaling changes *how* representations are constructed within the model, where smaller models may rely on more distributed mechanisms, while larger models develop specialized components.

This shift is reflected in the layerwise retrieval patterns (Figure 2b). The 7B and 3B models exhibit qualitatively distinct early peaks (within the first 25% of layers), followed by a rapid decline, suggesting key representations are handled by specific components early, and then disregarded. In contrast, the 0.5B and 1.5B models show more gradual and distributed accumulation of retrieval across layers[3]. Similar patterns have been observed in larger models (Kaplan et al., 2025), as well as in studies showing that structured behaviors such as induction are less clearly localized in smaller models (Feucht et al., 2025). Within this framework, the 1.5B model may represent an intermediate regime in which layer specialization is emerging but not yet fully established. This could explain its performance relative to the 3B model despite more distributed behavior across layers, and is consistent with phase transition phenomena observed during training, where models abruptly shift toward more specialized structure after sufficient training (Elhage et al., 2022; Power et al., 2022). It would also be consistent with the results that the 3B model performs worse than the 0.5B model in Chinese, suggesting that the model is attempting to utilize specialization but is not yet effective at it.

7 Conclusion

This article investigated whether detokenization in large language models depends on a specialized subset of attention heads. Across languages, tokenizers, and model families, targeted ablation disrupted retrieval more than random ablation, suggesting that early reconstruction of fragmented tokens relies on a relatively concentrated set of components. Larger models also exhibited sharper early-layer retrieval peaks and stronger ablation effects, indicating increased functional specialization with scale.

The results further suggest that tokenizer choice influences how these representations are formed. While byte efficiency and training exposure contribute to performance differences, they do not fully explain the observed variation across tokenizers and languages. Instead, the findings point toward a combination of tokenizer architecture, multilingual transfer, and representational organization as contributing factors. At the same time, the candidate-selection procedure likely captures only part of the underlying circuitry, meaning that additional retrieval-relevant heads may not be identified by the current boundary-based metrics. The experiments provide additional causal evidence for the detokenization hypothesis and support the view that early transformer layers actively reconstruct coherent word-level representations from fragmented token inputs.

Future Work

There are several paths of future work. This study does not consider the impact ablation has on generation. It could be the case that ablation does not hinder the model's ability to generate good next-token predictions. One direction of future work could take inspiration from Basile et al. (2026) who managed to *steer* output using semantic attention head ablation, but redesign the task for detokenization. Another unexplored area is how the phenomena emerges during training. Körner et al. (2026) studied EuroLLM checkpoints to see how shared concept spaces emerged during training. This is unexplored for the area of detokenization.

Limitations

Ablation is an inherently interventional method and may induce non-typical model states. As noted by Hase et al. (2021), observed failures under ablation may reflect artifacts of the intervention rather than naturally occurring model behavior. While this study focuses on internal representations rather than downstream outputs, the possibility remains that the measured effects do not fully correspond to how the model operates under normal conditions.

The applicability of the findings may be limited by the limited model choices and dataset size. Increasing the dataset size, as well as increasing the parameter size to also include even larger models, could impact the results. This could not be investigated in this work due to computational limitations.

Another limitation concerns the interpretation of attention head importance. As argued by Wiegreffe and Pinter (2019), the apparent importance of specific components may be explained by cor-

[3]These effects are even more pronounced in English than in Turkish or Chinese.

related or redundant structures elsewhere in the model. Consequently, the results do not establish a strictly causal role for individual attention heads, but rather indicate their importance within the constraints of the ablation setup.

References

Orevaoghene Ahia, Sachin Kumar, Hila Gonen, Jungo Kasai, David R Mortensen, Noah A Smith, and Yulia Tsvetkov. 2023. Do all languages cost the same? Tokenization in the era of commercial language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9904–9923.

Arofat Akhundjanova, Furkan Akkurt, Bermet Chontaeva, Soudabeh Eslami, and Cagri Coltekin. 2025. Parallel Universal Dependencies treebanks for Turkic languages. In *Proceedings of the Eighth Workshop on Universal Dependencies (UDW, SyntaxFest 2025)*, pages 129–136, Ljubljana, Slovenia. Association for Computational Linguistics.

Gül Sena Altıntas, Malikeh Ehghaghi, Brian Lester, Fengyuan Liu, Wanru Zhao, Marco Ciccone, and Colin Raffel. 2025. Toksuite: Measuring the impact of tokenizer choice on language model behavior. *Preprint*, arXiv:2512.20757.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online. Association for Computational Linguistics.

Lorenzo Basile, Valentino Maiorca, Diego Doimo, Francesco Locatello, and Alberto Cazzaniga. 2026. Head pursuit: Probing attention specialization in multimodal transformers. *Advances in Neural Information Processing Systems*, 38:112351–112383.

Qi Cao, Takeshi Kojima, Yutaka Matsuo, and Yusuke Iwasawa. 2023. Unnatural error correction: GPT-4 can almost perfectly handle unnatural scrambled text. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8898–8913, Singapore. Association for Computational Linguistics.

Tyler A. Chang, Catherine Arnett, Zhuowen Tu, and Benjamin K. Bergen. 2024. When is multilinguality a curse? Language modeling for 250 high- and low-resource languages. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4074–4096, Miami, Florida, USA. Association for Computational Linguistics.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? An analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. Association for Computational Linguistics.

Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Emerging cross-lingual structure in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6022–6034, Online. Association for Computational Linguistics.

Gonçalo M. Correia, Vlad Niculae, and André F. T. Martins. 2019. Adaptively sparse Transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2174–2184, Hong Kong, China. Association for Computational Linguistics.

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, and 26 others. 2024. Aya expanse: Combining research breakthroughs for a new multilingual frontier. *Preprint*, arXiv:2412.04261.

Ameet Deshpande, Partha Talukdar, and Karthik Narasimhan. 2022. When is BERT multilingual? Isolating crucial ingredients for cross-lingual transfer. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3610–3623, Seattle, United States. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. 2022. Toy models of superposition. *Transformer Circuits Thread*. https://transformer-circuits.pub/2022/toy_model/index.html.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones,

Jackson Kernion, Liane Lovitt, Kamal Ndousse, and 6 others. 2021. A mathematical framework for Transformer circuits. *Transformer Circuits Thread*. https://transformer-circuits.pub/2021/framework/index.html.

Javier Ferrando and Marta R. Costa-Jussà. 2024. On the similarity of circuits across languages: a case study on the subject-verb agreement task. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10115–10125, Miami, Florida, USA. Association for Computational Linguistics.

Javier Ferrando and Elena Voita. 2024. Information flow routes: Automatically interpreting language models at scale. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17432–17445, Miami, Florida, USA. Association for Computational Linguistics.

Sheridan Feucht, Eric Todd, Byron C Wallace, and David Bau. 2025. The dual-route model of induction. In *Second Conference on Language Modeling*.

Jorge García-Carrasco, Alejandro Maté, and Juan Carlos Trujillo. 2024. How does GPT-2 predict acronyms? Extracting and understanding a circuit via mechanistic interpretability. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pages 3322–3330. PMLR.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5484–5495, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Asma Ghandeharioun, Avi Caciularu, Adam Pearce, Lucas Dixon, and Mor Geva. 2024. Patchscopes: A unifying framework for inspecting hidden representations of language models. *Preprint*, arXiv:2401.06102.

Michael Hanna, Ollie Liu, and Alexandre Variengien. 2023. How does GPT-2 compute greater-than?: Interpreting mathematical abilities in a pre-trained language model. In *Advances in Neural Information Processing Systems*, volume 36, pages 76033–76060. Curran Associates, Inc.

Peter Hase, Harry Xie, and Mohit Bansal. 2021. The out-of-distribution problem in explainability and search methods for feature importance explanations. In *Advances in Neural Information Processing Systems*, volume 34, pages 3650–3666. Curran Associates, Inc.

Sarthak Jain and Byron C. Wallace. 2019. Attention is not Explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3543–3556, Minneapolis, Minnesota. Association for Computational Linguistics.

Bernal Jimenez Gutierrez, Huan Sun, and Yu Su. 2023. Biomedical language models are robust to sub-optimal tokenization. In *Proceedings of the 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 350–362, Toronto, Canada. Association for Computational Linguistics.

Julie Kallini, Dan Jurafsky, Christopher Potts, and Martijn Bartelds. 2025. False Friends are not foes: Investigating vocabulary overlap in multilingual language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 21138–21154, Suzhou, China. Association for Computational Linguistics.

Go Kamoda, Benjamin Heinzerling, Tatsuro Inaba, Keito Kudo, Keisuke Sakaguchi, and Kentaro Inui. 2025. Weight-based analysis of detokenization in language models: Understanding the first stage of inference without inference. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 6339–6358, Albuquerque, New Mexico. Association for Computational Linguistics.

Guy Kaplan, Matanel Oren, Yuval Reif, and Roy Schwartz. 2025. From tokens to words: On the inner lexicon of LLMs. In *The Thirteenth International Conference on Learning Representations*.

Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. 2020. Attention is not only a weight: Analyzing transformers with vector norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7057–7075, Online. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Felicia Körner, Max Müller-Eberstein, Anna Korhonen, and Barbara Plank. 2026. When meanings meet: Investigating the emergence and quality of shared concept spaces during multilingual language model training. *Preprint*, arXiv:2601.22851.

Vedang Lad, Wes Gurnee, and Max Tegmark. 2024. The remarkable robustness of LLMs: Stages of inference? In *ICML 2024 Workshop on Mechanistic Interpretability*.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav

Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, and 2 others. 2022. Few-shot learning with multilingual language models. *Preprint*, arXiv:2112.10668.

Xin Liu, Qiyang Song, Qihang Zhou, Haichao Du, Shaowen Xu, Wenbo Jiang, Weijuan Zhang, and Xiaoqi Jia. 2026. Focusing on language: Revealing and exploiting language attention heads in multilingual large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 32195–32203.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35, pages 17359–17372. Curran Associates, Inc.

Paul Michel, Omer Levy, and Graham Neubig. 2019. Are sixteen heads really better than one? In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Sabrina J. Mielke, Zaid Alyafeai, Elizabeth Salesky, Colin Raffel, Manan Dey, Matthias Gallé, Arun Raja, Chenglei Si, Wilson Y. Lee, Benoît Sagot, and Samson Tan. 2021. Between words and characters: A brief history of open-vocabulary modeling and tokenization in NLP. *Preprint*, arXiv:2112.10508.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

Isabel Papadimitriou and Jacob Prince. 2025. Vocabulary embeddings organize linguistic structure early in language model training. *Preprint*, arXiv:2510.07613.

Aleksandar Petrov, Emanuele La Malfa, Philip Torr, and Adel Bibi. 2023. Language model tokenizers introduce unfairness between languages. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. 2022. Grokking: Generalization beyond overfitting on small algorithmic datasets. *Preprint*, arXiv:2201.02177.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, and 24 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Sofia Serrano and Noah A. Smith. 2019. Is attention interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2931–2951, Florence, Italy. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Jesse Vig and Yonatan Belinkov. 2019. Analyzing the structure of attention in a transformer language model. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 63–76, Florence, Italy. Association for Computational Linguistics.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. 2019. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5797–5808, Florence, Italy. Association for Computational Linguistics.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. Do llamas work in English? on the latent language of multilingual transformers. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15366–15394, Bangkok, Thailand. Association for Computational Linguistics.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 11–20, Hong Kong, China. Association for Computational Linguistics.

Zhengxuan Wu, Alex Tamkin, and Isabel Papadimitriou. 2023. Oolong: Investigating what makes transfer learning hard with controlled studies. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3280–3289, Singapore. Association for Computational Linguistics.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. 2024. Efficient streaming language models with attention sinks. In *The Twelfth International Conference on Learning Representations*.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2022. ByT5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics*, 10:291–306.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 40 others. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Brian Siyuan Zheng, Alisa Liu, Orevaoghene Ahia, Jonathan Hayase, Yejin Choi, and Noah A. Smith. 2026. Broken tokens? Your language model can secretly handle non-canonical tokenizations. *Preprint*, arXiv:2506.19004.

Zhenhong Zhou, Haiyang Yu, Xinghua Zhang, Rongwu Xu, Fei Huang, Kun Wang, Yang Liu, Junfeng Fang, and Yongbin Li. 2025. On the role of attention heads in large language model safety. In *The Thirteenth International Conference on Learning Representations*.