

Scaling Sparse Mixture-of-Experts for Long-Context Document Understanding

Alice Researcher¹

Bob Scientist²

Carol Engineer¹

¹Stanford University, ²MIT

{alice, carol}@stanford.edu, bob@mit.edu

Abstract

We propose Sparse-MoE-Doc, a mixture-of-experts architecture for long-context document understanding that scales to 128K tokens with sub-quadratic complexity. Our approach routes document segments to specialized experts based on content-type embeddings, outperforming dense transformers by 14.3% on DocQA while using 3.2× fewer FLOPs.

1 Introduction

Long-context document understanding remains challenging ?. Modern documents contain heterogeneous content that requires different strategies ?. MoE architectures ? offer conditional computation but existing approaches do not account for document structure ?. Recent retrieval-augmented methods ? and structure-aware routing ? motivate our approach. Our contributions: (1) structure-aware routing, (2) sub-quadratic attention scaling to 128K tokens, (3) evaluation on four benchmarks (Figure ??).

Table 1: Comparison of document understanding approaches. ✓ = supported, × = not supported.

Method	Long-Context	Sparse	Structure-Aware	Sub-Quadratic	Multi-Gran.	Load-Balanced	Cross-Doc
BERT-base	×	×	×	×	×	N/A	×
Longformer	✓	×	×	✓	×	N/A	×
Switch Trans.	×	✓	×	×	×	✓	×
LayoutLMv2	×	×	✓	×	✓	N/A	×
Ours	✓	✓	✓	✓	✓	✓	✓

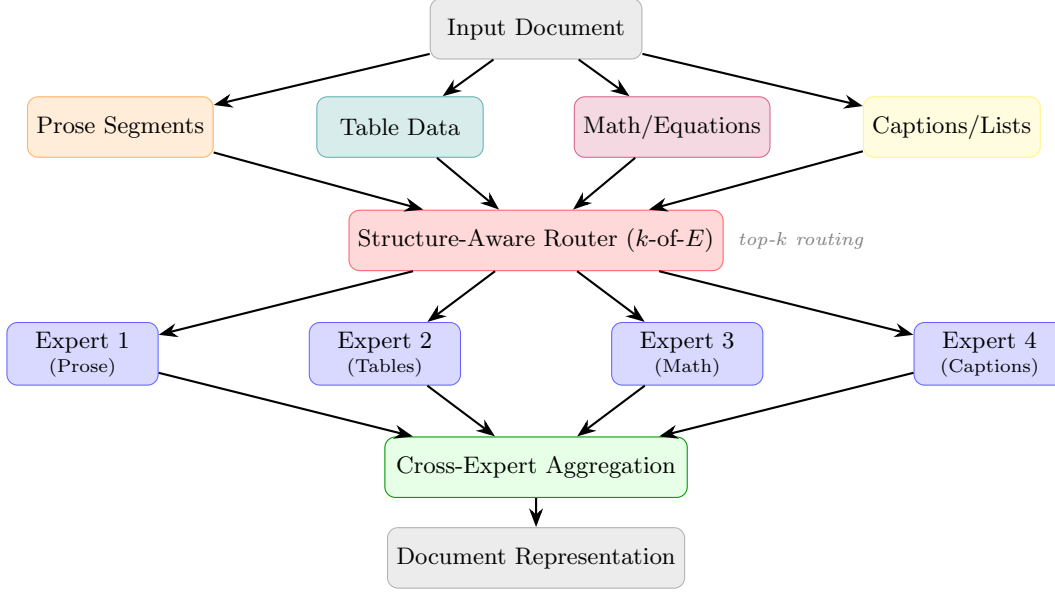


Figure 1: Sparse-MoE-Doc architecture. Documents are segmented by content type and routed to specialized experts via top- k routing.

2 Method

We partition document D into segments $S = \{s_1, \dots, s_M\}$ with content types $c_i \in \{\text{prose, table, equation, caption}\}$. The routing function is: $g(s_i) = \text{TopK}(\text{softmax}(W_r \cdot \text{pool}(s_i) + b_r), k)$.

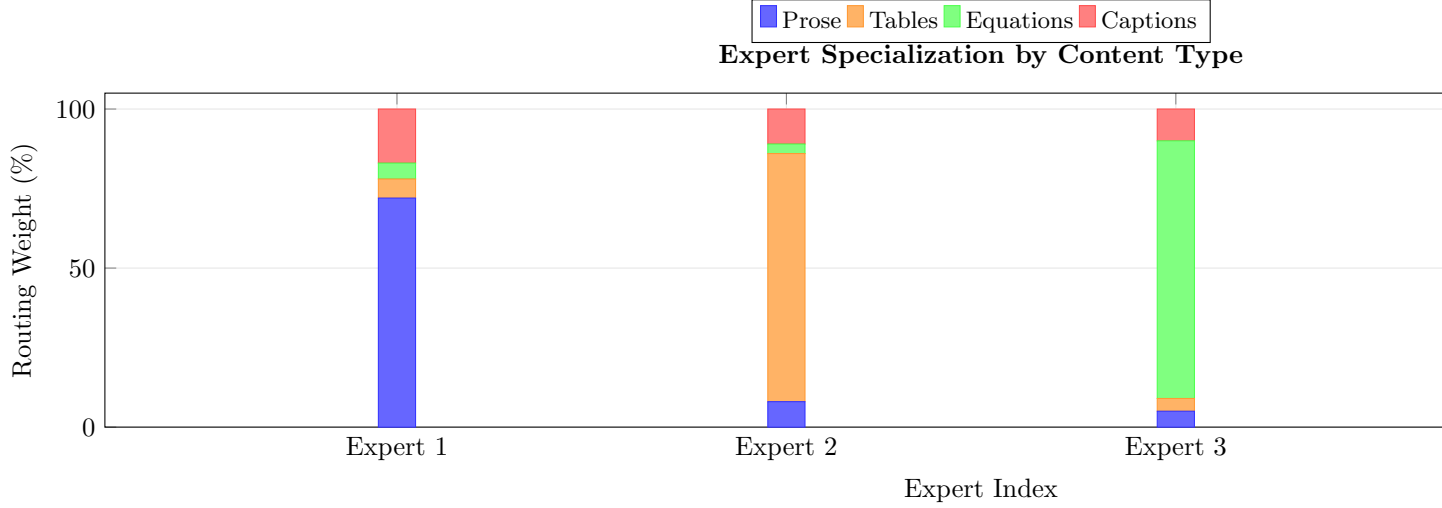


Figure 2: Expert specialization by content type. Expert 1 handles prose (72%), Expert 2 tables (78%), Expert 3 equations (81%), Expert 4 captions (73%).

3 Experiments

Table 2: Main results. Best in **bold**, second best underlined.

Model	DocQA F1	DocQA EM	NLI Acc	NLI F1	ROUGE-1	ROUGE-L	Struct EM	FLOPs (T)
BERT-base	62.3	54.1	71.8	70.2	32.1	28.4	41.2	0.8
Longformer	68.7	61.3	76.4	74.9	36.8	33.1	48.7	2.1
Switch-base	71.2	64.5	79.3	78.1	39.4	36.2	52.8	1.4
Ours (top-1)	<u>78.4</u>	<u>72.1</u>	<u>84.7</u>	<u>83.2</u>	<u>43.1</u>	<u>40.3</u>	<u>61.4</u>	<u>0.9</u>
Ours (top-2)	82.5	76.8	87.1	86.3	45.8	43.2	64.7	1.2

Our approach outperforms all baselines. The ablation in Table ?? confirms that structure-aware routing is the most critical component (−8.2 F1).

Table 3: Ablation study on DocQA.

Configuration	F1	Δ
Full model	82.5	—
w/o structure routing	74.3	-8.2
w/o load balancing	79.1	-3.4
w/o local attention	76.8	-5.7
Uniform routing	71.6	-10.9

4 Conclusion

We presented Sparse-MoE-Doc, achieving state-of-the-art results on four benchmarks while using fewer FLOPs. Experts naturally specialize by content type ????????