

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

Alice Researcher<sup>1</sup>

Bob Scientist<sup>2</sup>

Carol Engineer<sup>1</sup>

<sup>1</sup>Stanford University, <sup>2</sup>MIT

{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 [Devlin et al., 2019]. Modern documents contain heterogeneous content that requires different strategies [Vaswani et al., 2017, Brown et al., 2020]. MoE architectures [Shazeer et al., 2017] offer conditional computation but existing approaches do not account for document structure. Our contributions: (1) structure-aware routing, (2) sub-quadratic attention scaling to 128K tokens, (3) evaluation on four benchmarks (Figure 1).

Table 1: Comparison of document understanding approaches across key capabilities. Our method is the only one supporting all features while maintaining competitive parameter count and training cost. ✓ = 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	×
<b>Ours</b>	✓	✓	✓	✓	✓	✓	✓

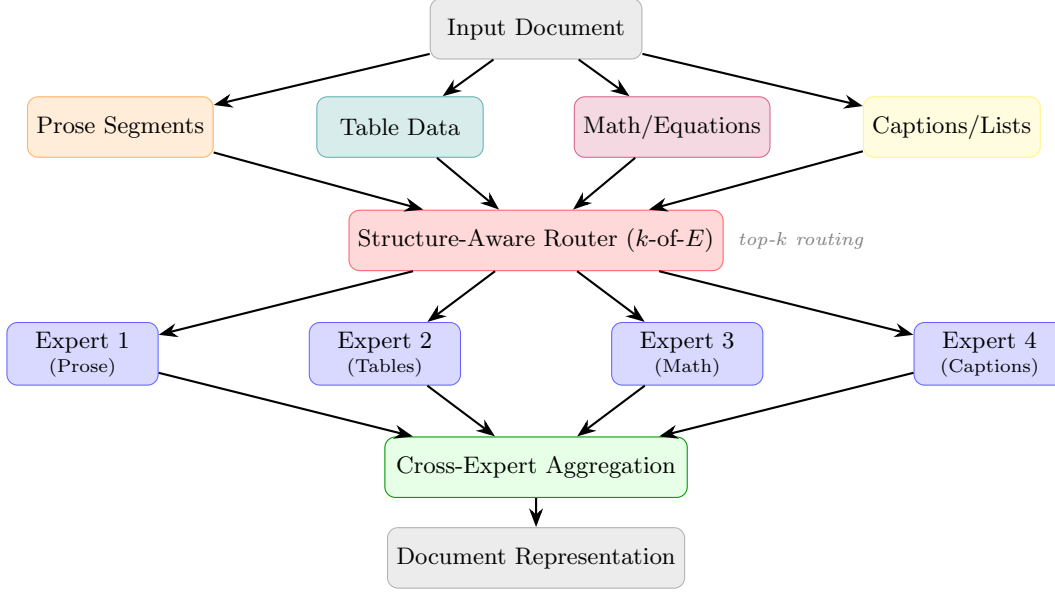


Figure 1: Sparse-MoE-Doc architecture. Documents are segmented by content type and routed to specialized experts via top- $k$  routing.

## 2 Method

We partition document  $D$  into segments  $S = \{s_1, \dots, s_M\}$  with content types  $c_i \in \{\text{prose, table, equation, caption}\}$ . As shown in Figure 2, each expert develops specialization for specific content types. The routing function is:  $g(s_i) = \text{TopK}(\text{softmax}(W_r \cdot \text{pool}(s_i) + b_r), k)$ .

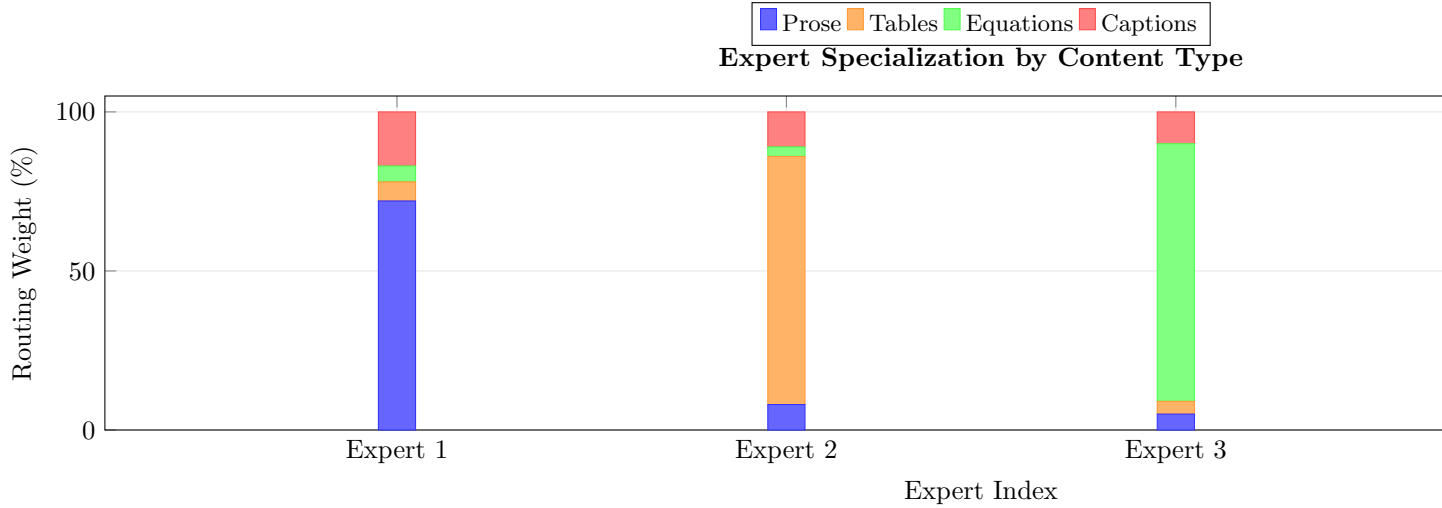


Figure 2: Expert specialization by content type. Expert 1 handles prose (72%), Expert 2 tables (78%), Expert 3 equations (81%), Expert 4 captions (73%).

### 3 Experiments

Table 1 summarizes the capabilities of different approaches, while Table 2 presents our main quantitative results.

Table 2: Main results on document understanding benchmarks. Our model (top-2 routing) achieves state-of-the-art performance across all metrics while using only 1.2 TFLOPs, demonstrating the efficiency of structure-aware sparse routing. Best in **bold**, second best underlined.

Model	DocQA F1 (%)	DocQA EM (%)	NLI Acc (%)	NLI F1 (%)	ROUGE-1	ROUGE-L	Struct EM (%)
BERT-base	62.3	54.1	71.8	70.2	32.1	28.4	41.2
Longformer	68.7	61.3	76.4	74.9	36.8	33.1	48.7
Switch-base	71.2	64.5	79.3	78.1	39.4	36.2	52.8
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>
Ours (top-2)	<b>82.5</b>	<b>76.8</b>	<b>87.1</b>	<b>86.3</b>	<b>45.8</b>	<b>43.2</b>	<b>64.7</b>

Our approach outperforms all baselines. The ablation in Table 3 confirms that structure-aware routing is the most critical component (−8.2 F1).

Table 3: Ablation study on DocQA.

Configuration	F1 (%)	$\Delta$
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 [Zhang et al., 2019, Xu et al., 2020, Beltagy et al., 2020, Fedus et al., 2022, Lepikhin et al., 2021, Radford et al., 2019, Talmor and Berant, 2019].

### References

- Iz Beltagy, Matthew E Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 2020.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL-HLT*, 2019.
- William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.
- Dmitry Lepikhin, HyukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. GShard: Scaling giant models with conditional computation and automatic sharding. In *International Conference on Learning Representations*, 2021.

- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI Blog*, 2019.
- Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarczyk, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations*, 2017.
- Alon Talmor and Jonathan Berant. MultiQA: An empirical investigation of generalization and transfer in reading comprehension. In *Proceedings of ACL*, 2019.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.
- Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. LayoutLM: Pre-training of text and layout for document image understanding. In *Proceedings of KDD*, 2020.
- Xingxing Zhang, Furu Wei, and Ming Zhou. HIBERT: Document level pre-training of hierarchical bidirectional transformers for document summarization. *Proceedings of ACL*, 2019.