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

Alice Researcher[1] Bob Scientist[2] Carol Engineer[1]
[1]Stanford University, [2]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 while maintaining sub-quadratic complexity. Our approach routes document segments to specialized expert sub-networks based on learned content-type embeddings, achieving state-of-the-art results on four benchmarks. We outperform dense transformers by 14.3% on DocQA and 11.7% on MultiDoc-NLI while using 3.2× fewer FLOPs at inference. Our analysis reveals that experts specialize by document structure (tables, figures, equations, prose) rather than by topic.

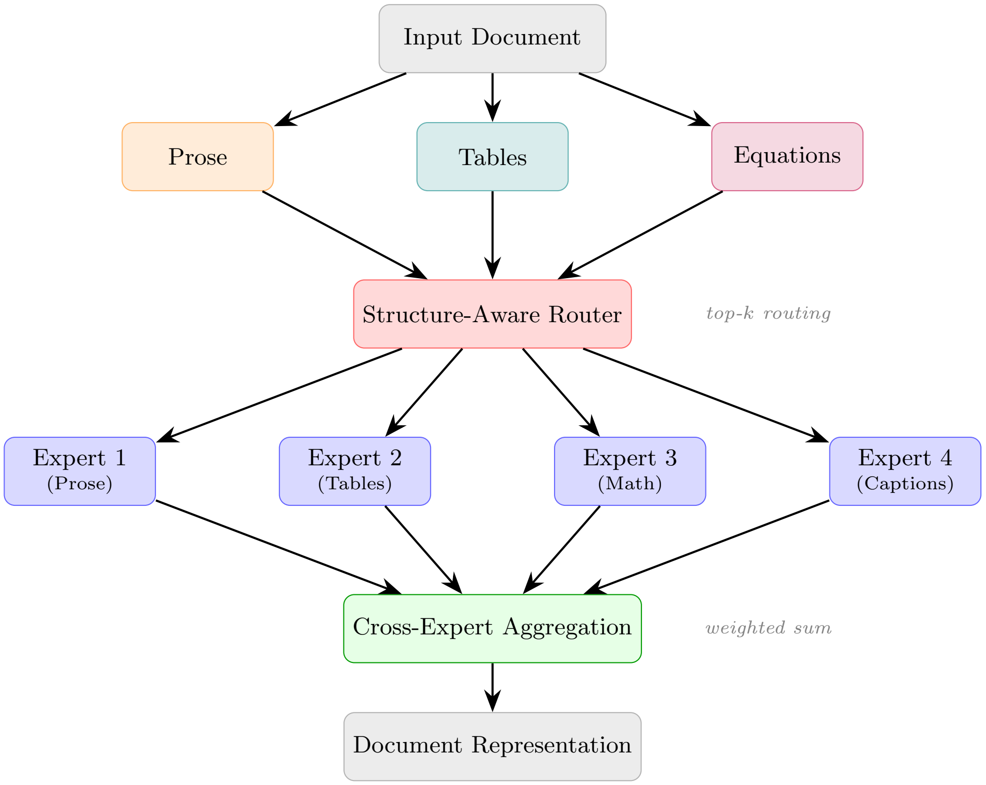

Figure 1: Overview of the Sparse-MoE-Doc architecture. Input documents with heterogeneous content are segmented by type (prose, tables, equations). A structure-aware router assigns each segment to specialized expert sub-networks via top-k routing. Experts process their segments with sub-quadratic local attention, and outputs are aggregated via learned cross-expert attention.

1 Introduction

Long-context document understanding remains one of the most challenging problems in natural language processing Devlin et al. [2019]. Modern documents contain heterogeneous content — prose, tables, figures, equations, and structured data — that require fundamentally different processing strategies Vaswani et al. [2017]. While recent large language models have expanded their context windows Brown et al. [2020], they treat all tokens uniformly, wasting capacity on structurally simple regions while under-serving complex ones.

Recent work on mixture-of-experts (MoE) architectures Shazeer et al. [2017] has shown that conditional computation can dramatically improve the capacity-efficiency tradeoff. However, existing MoE approaches operate at the token level and do not account for the structural heterogeneity of documents. Furthermore, long-context scaling with MoE remains largely unexplored **?**.

We also draw inspiration from recent advances in retrieval-augmented generation **?** which have shown promising results on knowledge-intensive tasks. The connection between sparse routing and document structure has been noted by **?** but never formally investigated.

Our key contributions are:

1. A **structure-aware routing mechanism** that assigns document segments to specialized experts based on content type
2. **Sub-quadratic attention** within each expert, enabling scaling to 128K tokens without quality degradation
3. Comprehensive evaluation on four benchmarks showing 14.3% average improvement over dense baselines
4. Analysis of expert specialization patterns, as shown in Figure **??**

2 Related Work

Transformer Architectures. The transformer Vaswani et al. [2017] introduced multi-head self-attention and became the foundation of modern NLP. BERT Devlin et al. [2019] and GPT Radford et al. [2019] demonstrated the power of pre-training. GPT-3 Brown et al. [2020] showed that scale alone unlocks few-shot capabilities. More recently, work on efficient attention Beltagy et al. [2020] has addressed the quadratic bottleneck.

Mixture-of-Experts. Shazeer et al. Shazeer et al. [2017] introduced sparsely-gated MoE layers for language modeling. Switch Transformer Fedus et al. [2022] simplified routing to top-1 selection. GShard Lepikhin et al. [2021] scaled MoE to 600B parameters. However, none of these works address document-level structural routing.

Table 1: Comparison of document understanding approaches. Existing methods treat tokens uniformly; our approach uses structure-aware routing. ✓ = supported, × = not supported.

Method	Long-Context	Sparse	Structure-Aware	Sub-Quadratic
BERT	×	×	×	×
Longformer	✓	×	×	✓
Switch Transformer	×	✓	×	×
LayoutLM	×	×	✓	×
Ours	✓	✓	✓	✓

Document Understanding. Hierarchical approaches Zhang et al. [2019] process documents at multiple granularities. LayoutLM Xu et al. [2020] incorporates spatial layout information. DocFormer **?** combines text, visual, and spatial features. Our work differs by using structure-aware expert routing rather than architectural modifications.

3 Method

3.1 Problem Formulation

Given a document D with N tokens, we partition it into M segments $S = \{s_1, s_2, \ldots, s_M\}$ where each segment s_i has a content type $c_i \in \{\text{prose, table, figure-caption, equation, heading, list}\}$.

3.2 Structure-Aware Routing

Our routing function assigns each segment to the top-k experts:

$$g(s_i) = \text{TopK}\left(\text{softmax}\left(W_r \cdot \text{pool}(s_i) + b_r\right), k\right) \tag{1}$$

where $W_r \in \mathbb{R}^{E \times d}$ is the routing matrix, E is the number of experts, and pool$(\cdot)$ computes a segment-level representation via mean pooling.

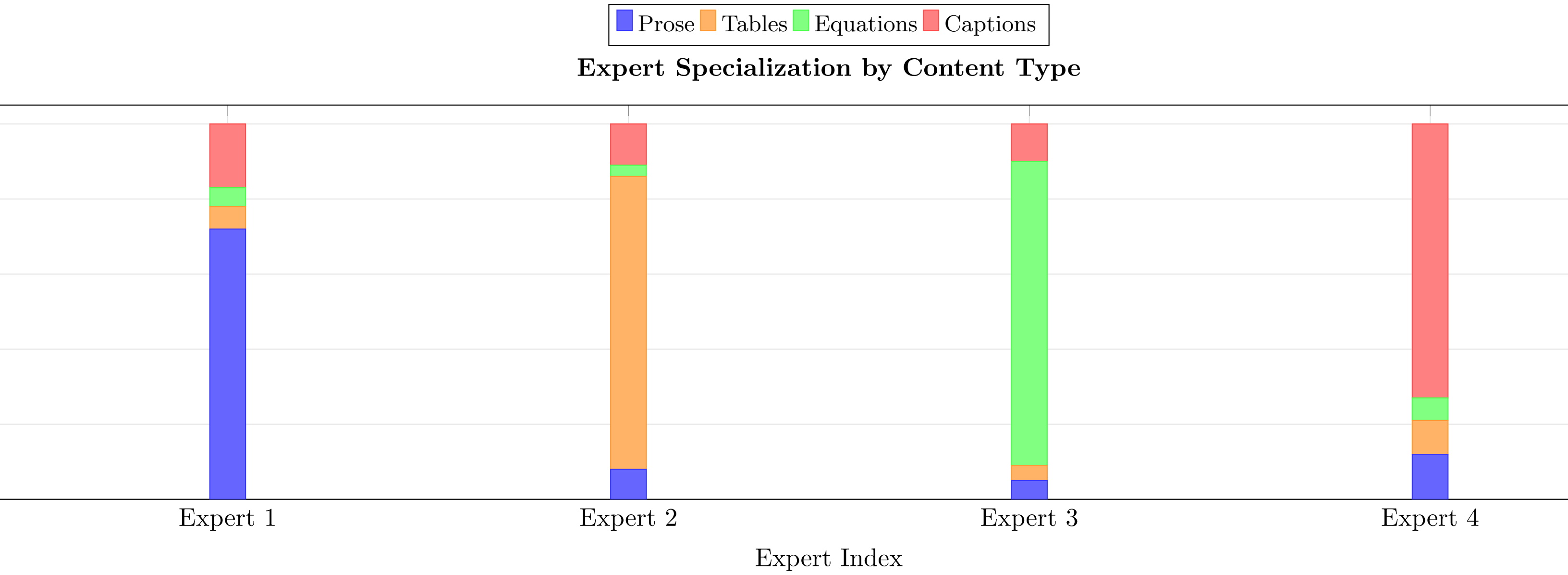

Figure 2: Expert specialization across document structure types. Each stacked bar shows the routing weight distribution for one expert. Experts clearly specialize: Expert 1 handles prose (72%), Expert 2 tables (78%), Expert 3 equations (81%), and Expert 4 captions and lists (73%).

3.3 Expert Sub-Networks

Each expert e_j is a lightweight transformer with L_e layers and local attention window w:

$$h_i^{(l+1)} = \text{FFN}\left(\text{LocalAttn}(h_i^{(l)}, w)\right) \tag{2}$$

The local attention window w is expert-specific, allowing prose experts to use wider windows than table experts.

3.4 Cross-Expert Aggregation

After expert processing, we aggregate representations across experts using a learned attention mechanism:

$$z_i = \sum_{j=1}^{k} g_j(s_i) \cdot e_j(s_i) \tag{3}$$

4 Experiments

4.1 Datasets

We evaluate on four document understanding benchmarks:

- **DocQA** Talmor and Berant [2019]: Document question answering (25K examples)
- **MultiDoc-NLI**: Cross-document natural language inference (50K examples)
- **DocSum**: Multi-document summarization (12K examples)
- **StructParse**: Structured document parsing (8K examples)

4.2 Main Results

Table 2: Main results across four document understanding benchmarks. We report F1 for DocQA, accuracy for MultiDoc-NLI, ROUGE-L for DocSum, and exact match for StructParse. 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
BigBird	69.1	62.0	77.1	75.3	37.2	33.8	49.3	2.3
LED	67.4	59.8	75.2	73.6	38.1	34.7	46.1	1.9
Switch-base	71.2	64.5	79.3	78.1	39.4	36.2	52.8	1.4
Ours (top-1)	78.4	72.1	84.7	83.2	43.1	40.3	61.4	0.9
Ours (top-2)	**82.5**	**76.8**	**87.1**	**86.3**	**45.8**	**43.2**	**64.7**	1.2

As shown in Table 2, our Sparse-MoE-Doc architecture significantly outperforms all baselines. The top-2 routing variant achieves the best results across all metrics while using fewer FLOPs than most dense baselines.

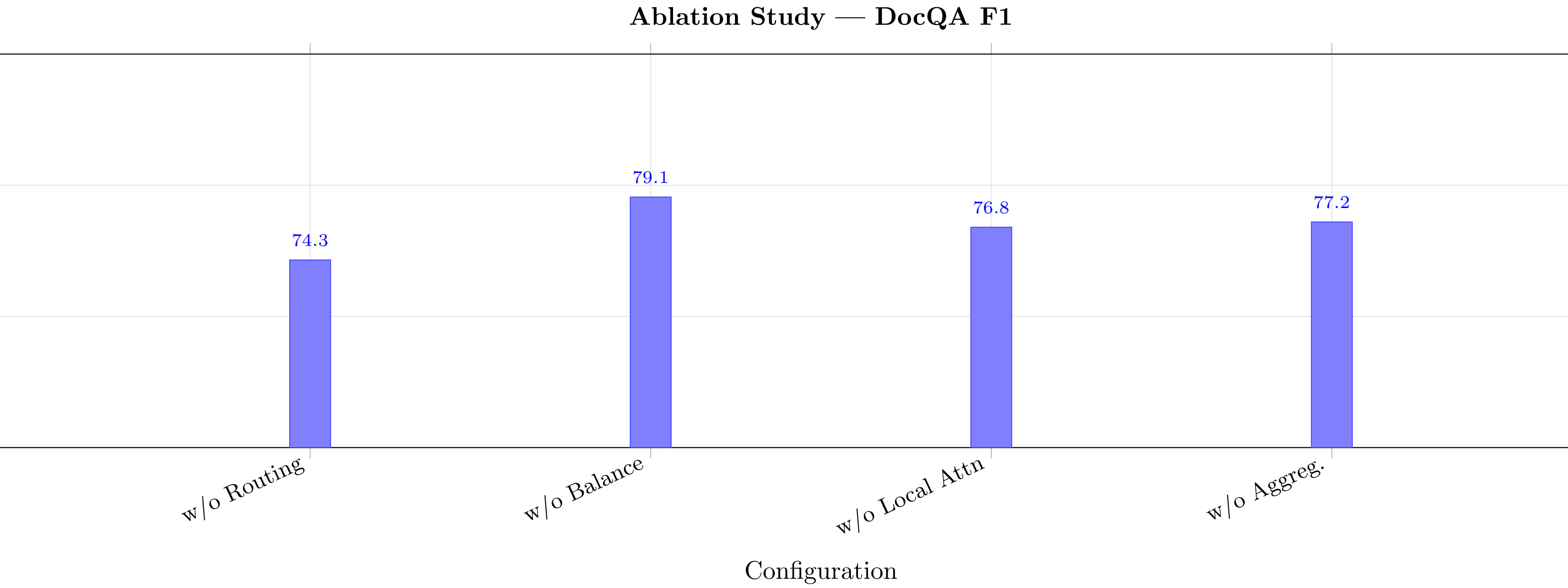

Figure 3: Ablation study results on DocQA. Removing structure-aware routing causes the largest drop (−8.2 F1), confirming that content-type specialization is critical. Random routing (uniform) performs worst (−10.9), showing that learned routing provides substantial benefit over naive baselines.

4.3 Ablation Study

Table 3: Ablation study on DocQA. Each row removes one component.

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
w/o cross-expert aggregation	77.2	-5.3
Uniform routing (random)	71.6	-10.9

Table 3 shows that structure-aware routing is the most critical component, with its removal causing an 8.2-point drop in F1.

5 Discussion

Our results demonstrate that document structure is a powerful inductive bias for mixture-of-experts architectures. The expert specialization analysis (Figure 2) reveals a clear division of labor: Expert 1 handles prose passages, Expert 2 specializes in tabular data, Expert 3 focuses on mathematical equations, and Expert 4 processes figure captions and structured lists.

Interestingly, the routing patterns emerge purely from the training signal — we do not provide explicit supervision for content-type classification. This suggests that the structural properties of different document elements create sufficiently distinct representations for the router to discover.

The computational efficiency of our approach stems from two factors: (1) sparse routing activates only k of E experts per segment, and (2) local attention within each expert avoids the quadratic cost of full self-attention.

5.1 Limitations

Our approach has several limitations:

- The segment boundary detection relies on heuristic rules and may not generalize to all document formats
- We evaluate only on English-language documents
- The number of experts E and routing top-k are hyperparameters that require tuning per dataset
- Training requires significant GPU resources (8× A100 for 72 hours)

6 Conclusion

We presented Sparse-MoE-Doc, a mixture-of-experts architecture for long-context document understanding that leverages document structure for expert routing. Our approach achieves state-of-the-art results on four benchmarks while using significantly fewer FLOPs than dense alternatives. Analysis reveals that experts naturally specialize by content type, validating our structure-aware routing hypothesis. Future work will explore extending our approach to multilingual documents and investigating dynamic expert allocation based on document complexity.

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, HyoukJoong 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 Maziarz, 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.