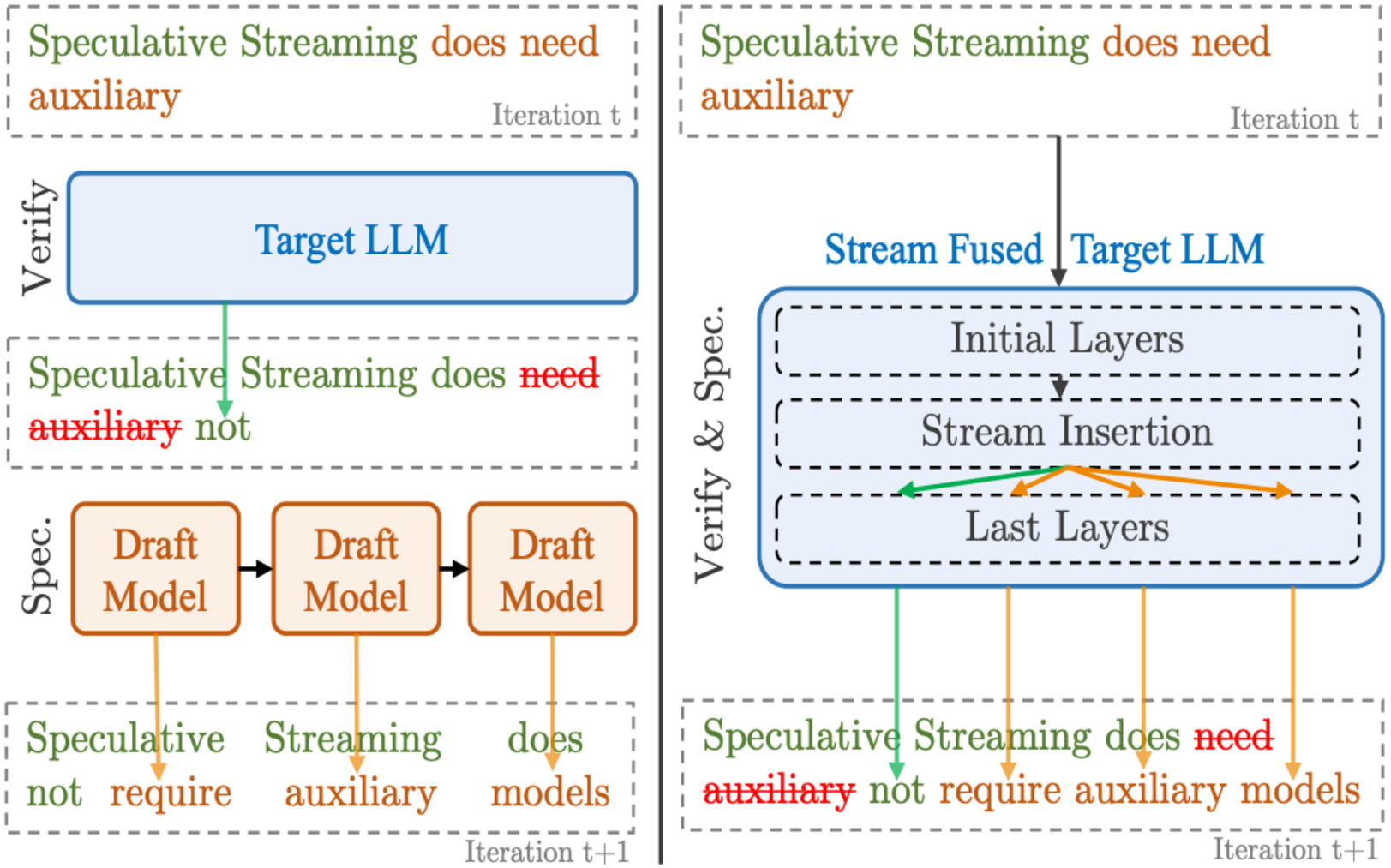
# Speculative Streaming: Fast LLM Inference without Auxiliary Models

Nikhil Bhendawade · Irina Belousova · Qichen Fu · Henry Mason · Mohammad Rastegari · Mahyar Najibi

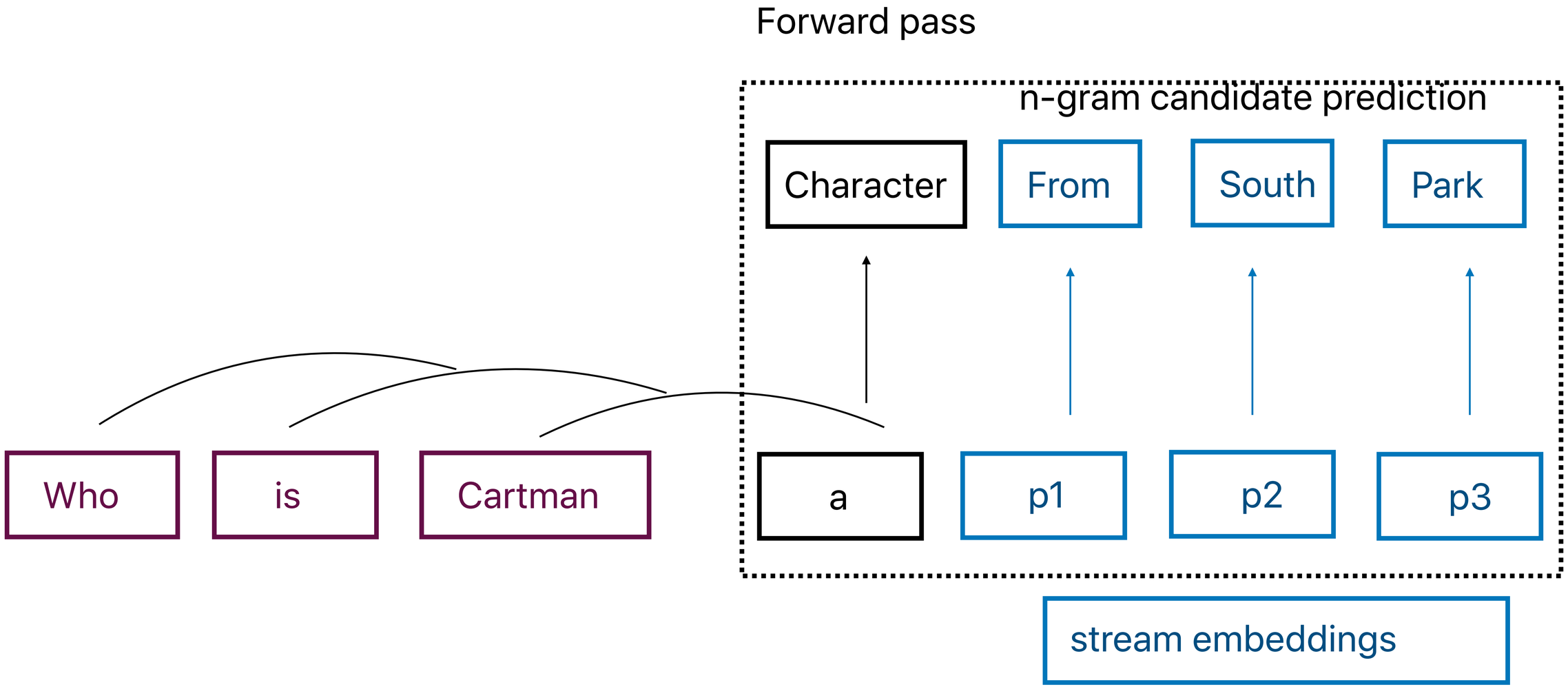
NeurIPS Workshop 2024 | Apple | 12/14/2024

# Speculative Decoding vs Speculative Streaming



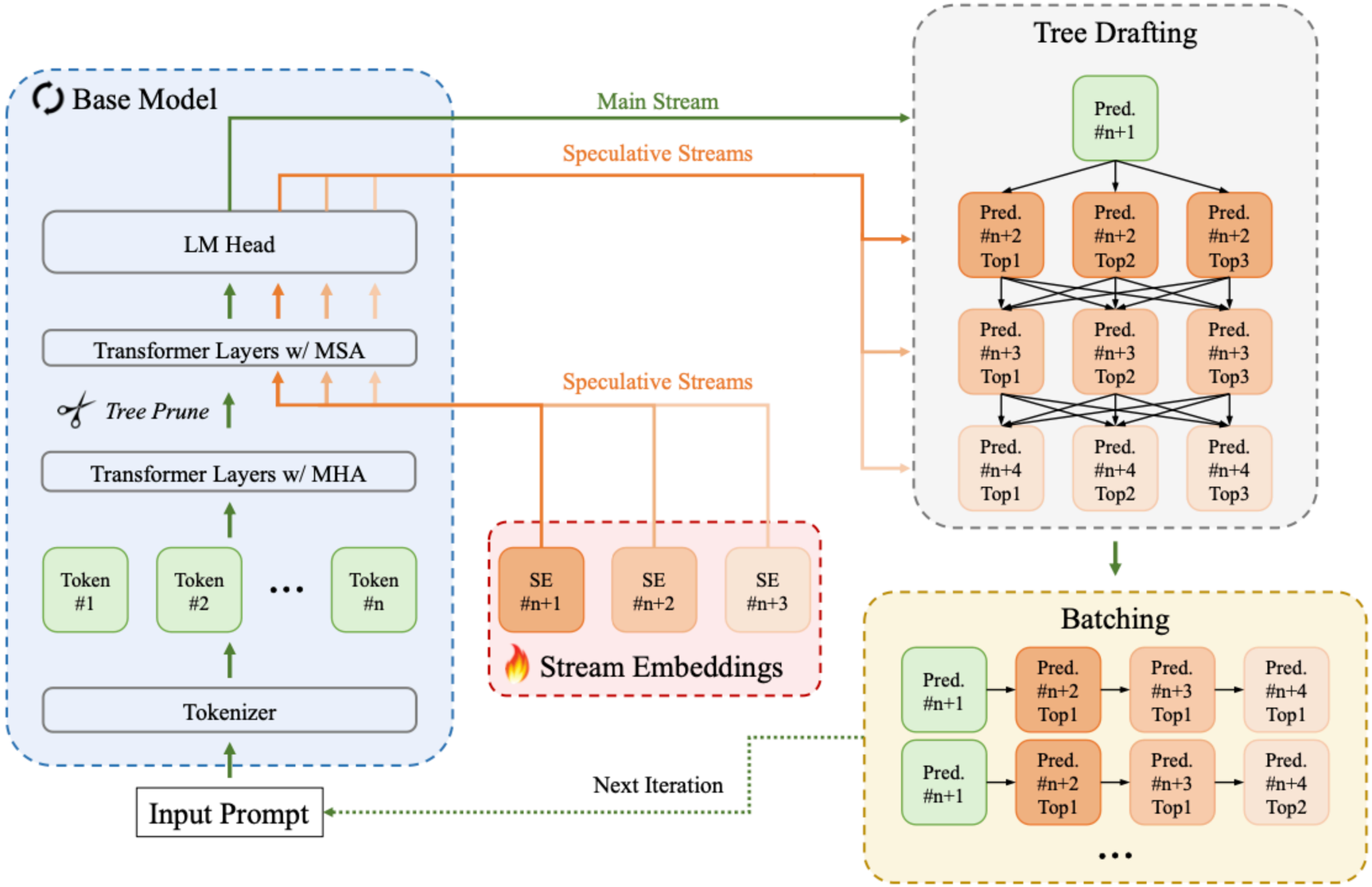

(a) Speculative Decoding (b) Speculative Streaming

# Speculative Streaming 101




*attention of prophet streams not shown

# Architecture



# Example

# Person 2 # ~~and~~

# Person 2 # thinks Lincoln ~~is a character~~

# Person 2 # thinks Lincoln was a ~~character and he~~

# Person 2 # thinks Lincoln was a man of ~~character and he~~

# Person 2 # thinks Lincoln was a man of sound character and # person

# Person 2 # thinks Lincoln was a man of sound character and # person 1 # adm ires him

# Person 2 # thinks Lincoln was a man of sound character and # person 1 # adm ires him for his courage and ~~and~~

# Person 2 # thinks Lincoln was a man of sound character and # person 1 # adm ires him for his courage and rights and humility . </s>

Figure 19: Speculative streaming on Dialog Summarization task for $\gamma = 4$ and $k = 1$, each pass verifies the previous draft and generates a maximum of 5 tokens. For instance, in pass 3, "$is$", "$a$", "$character$" are rejected and "$was$", "$a$", "$character$", "$and$", "$he$" are speculated.

# Mean walltime speedup on Vicuna and Llama-2 models

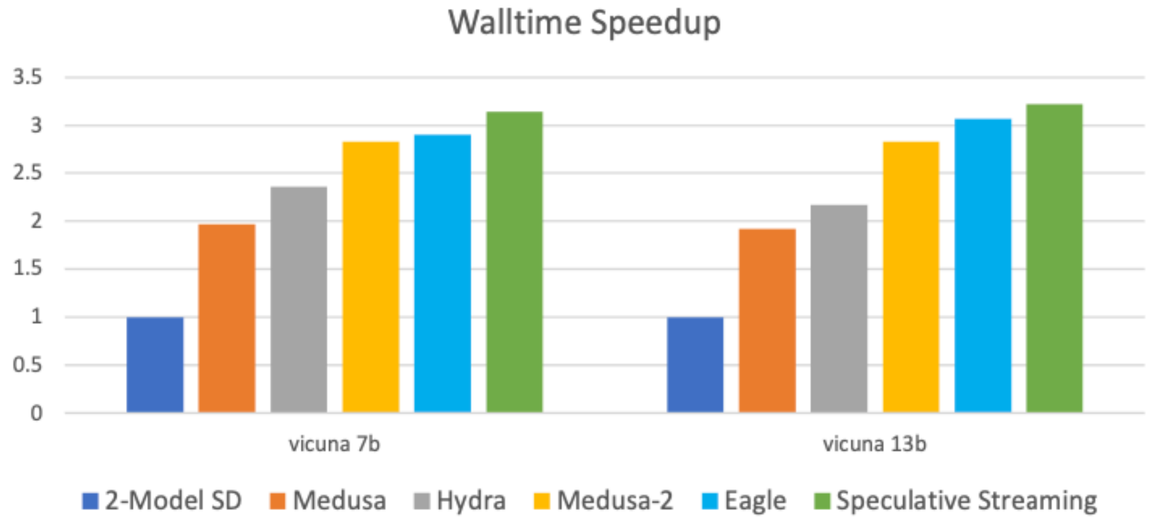



Figure 3: Mean walltime speedup on Vicuna models of various sizes to demonstrate scalability and generalizability of our approach on MT-Bench.

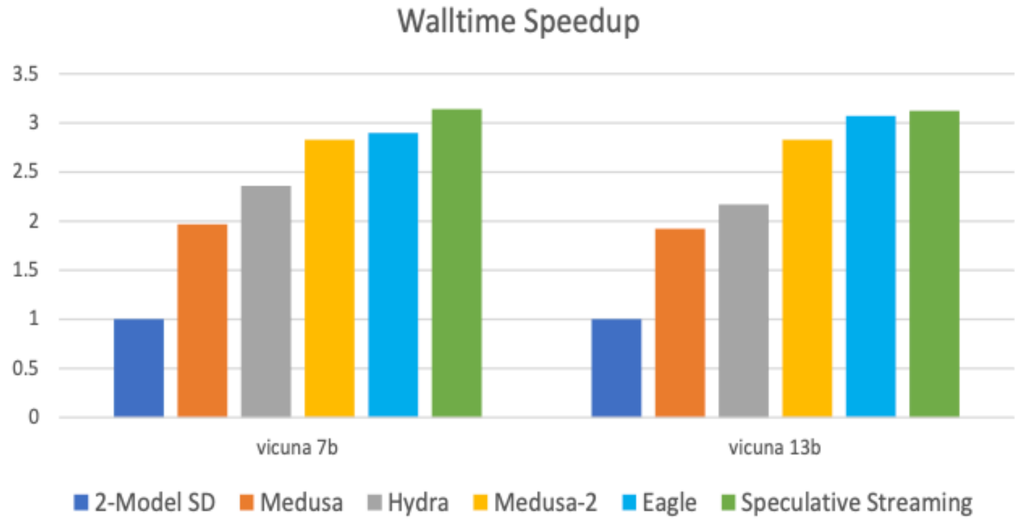



Figure 4: Mean walltime speedup on Llama-2 models of various sizes to demonstrate scalability and generalizability of our approach.

# Results

| Dataset | Model | Method | SpeedUp (↑) | CR Ratio (↑) | Metric (↑) | # Extra Parameters (↓) |
|---|---|---|---|---|---|---|
| SqlContext | Mistral-Instruct-7B | Baseline | 1.00 | 1.00 | 84.16 | – |
| | | Medusa-2 | 2.79 | 3.18 | 84.18 | $5.9E8$ |
| | | Eagle | 2.75 | 3.58 | 84.16 | $2.4E8$ |
| | | SS (ours) | **2.93** | **3.67** | **84.50** | $\underline{8.2E4}$ |
| | PHI-3-Instruct-3.8B | Baseline | 1.00 | 1.00 | 80.92 | – |
| | | Medusa-2 | 2.54 | 2.81 | 81.07 | $4.3E8$ |
| | | Eagle | 2.62 | 3.37 | 80.92 | $1.3E8$ |
| | | SS (ours) | **2.92** | **3.65** | **84.10** | $\underline{6.1E4}$ |
| | Llama2-7b | Baseline | 1.00 | 1.00 | 85.37 | – |
| | | Medusa-2 | 2.52 | 2.98 | 85.31 | $5.9E8$ |
| | | Eagle | 2.59 | 3.31 | 85.37 | $2.4E8$ |
| | | SS (ours) | **2.81** | **3.57** | **85.93** | $\underline{8.2E4}$ |
| DialogSum | Mistral-Instruct-7B | Baseline | 1.00 | 1.00 | 44.74/36.76 | – |
| | | Medusa-2 | 1.89 | 2.05 | 44.78/36.95 | $5.9E8$ |
| | | Eagle | 1.95 | 2.56 | 44.74/36.76 | $2.4E8$ |
| | | SS (ours) | **2.04** | **2.96** | **44.89/37.09** | $\underline{8.2E4}$ |
| | PHI-3-Instruct-3.8B | Baseline | 1.00 | 1.00 | 46.08/38.28 | – |
| | | Medusa-2 | 2.15 | 2.26 | 45.82/37.78 | $4.3E8$ |
| | | Eagle | 2.05 | 2.31 | 46.08/38.28 | $1.3E8$ |
| | | SS (ours) | **2.32** | **2.85** | **46.30/38.32** | $\underline{6.1E4}$ |
| | Llama2-7b | Baseline | 1.00 | 1.00 | 44.90/37.0 | – |
| | | Medusa-2 | 1.76 | 1.95 | 44.17/37.02 | $5.9E8$ |
| | | Eagle | 1.86 | 2.57 | 44.90/37.0 | $2.4E8$ |
| | | SS (ours) | **1.90** | **3.05** | **45.0/37.85** | $\underline{8.2E4}$ |
| E2E-NLG | Mistral-Instruct-7B | Baseline | 1.00 | 1.00 | 67.82/48.99 | – |
| | | Medusa-2 | 2.78 | 3.19 | 67.74/48.85 | $5.9E8$ |
| | | Eagle | 2.85 | 3.52 | 67.82/48.99 | $2.4E8$ |
| | | SS (ours) | **2.93** | **3.67** | **68.37/49.09** | $\underline{8.2E4}$ |
| | PHI-3-Instruct-3.8B | Baseline | 1.00 | 1.00 | 68.72/49.31 | – |
| | | Medusa-2 | 2.39 | 2.63 | 68.41/49.08 | $4.3E8$ |
| | | Eagle | **2.42** | **2.76** | 68.72/49.31 | $1.3E8$ |
| | | SS (ours) | 2.36 | 2.72 | **69.38/50.22** | $\underline{6.1E4}$ |
| | Llama2-7b | Baseline | 1.00 | 1.00 | 69.47/49.54 | – |
| | | Medusa-2 | 2.82 | 3.19 | 69.41/49.44 | $5.9E8$ |
| | | Eagle | 2.79 | 3.26 | 69.47/49.54 | $2.4E8$ |
| | | SS (ours) | **2.89** | **3.38** | **69.52/49.93** | $\underline{8.2E4}$ |

# Comparison with standard draft-target speculative decoding approach

| Dataset | Target | Method | Target calls | Draft Calls | Walltime Latency ($ms, \downarrow$) | Metric ($\uparrow$) |
|---|---|---|---|---|---|---|
| SqlContext | OPT-1.3b | Two-model SD | 6.59 | 22.35 | 269.24 | 84.98 |
| | | SS (ours) | 7.79 | 0 | **133.48** | **87.40** |
| | OPT-6.7b | Two-model SD | 6.60 | 22.41 | 301.10 | 89.13 |
| | | SS (ours) | 6.88 | 0 | **157.04** | **89.34** |
| Dialogsum | OPT-1.3b | Two-model SD | 11.65 | 42.59 | 493.59 | 43.40/35.60 |
| | | SS (ours) | 13.41 | 0 | **248.26** | **44.07/35.99** |
| | OPT-6.7b | Two-model SD | 12.15 | 35.76 | 555.99 | 44.40/36.60 |
| | | SS (ours) | 14.45 | 0 | **444.67** | **44.42/36.81** |
| E2E-NLG | OPT-1.3b | Two-model SD | 8.86 | 31.47 | 345.72 | **69.48**/50.17 |
| | | SS (ours) | 9.80 | 0 | **164.23** | 69.32/**50.51** |
| | OPT-6.7b | Two-model SD | 8.90 | 31.58 | 412.02 | 69.34/**49.88** |
| | | SS (ours) | 10.31 | 0 | **244.80** | **69.45**/49.78 |

# Parameter/Memory access

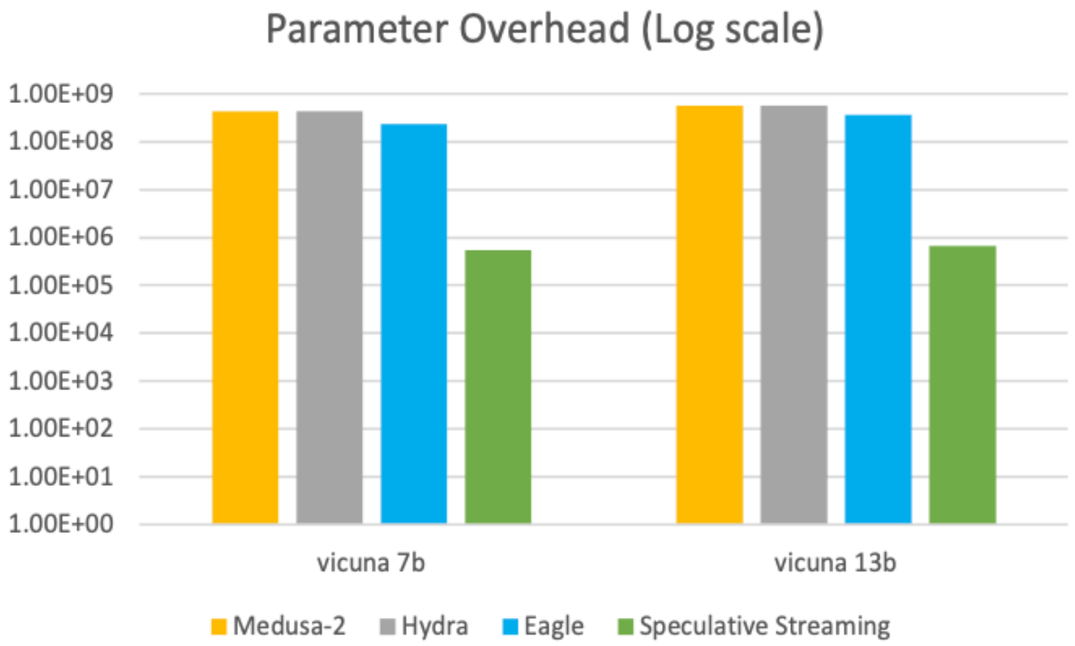



Figure 5: Parameter/Memory access overhead of different SD architectures with Vicuna models.

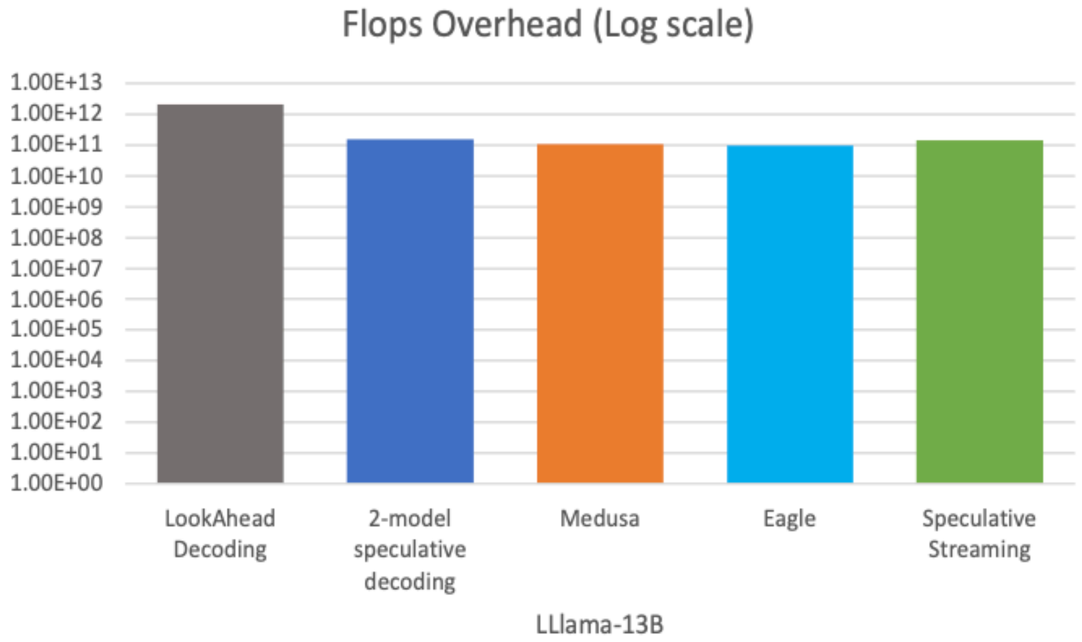



Figure 6: FLOP overhead of different SD architectures with Llama-13B.

# Conclusion

The key advantages of Speculative Streaming are as follows

- Achieves substantial 1.9 - 3X decoding speedups and improves downstream performance metrics through a single, streamlined fine-tuning process leveraging multi-stream attention
- Demonstrates resource efficiency with ~10000X fewer additional parameters compared to Medusa, Hydra and Eagle, while still surpassing them in speedup gains
- Simplifies deployment by removing the complexity of managing, aligning, and switching between multiple models during inference