Bringing LLMs to the Edge: Scalable and Fault-Tolerant AI

Miquel Sirera Perelló, Joshua Groen, Wan Liu, Stratis Ioannidis, and Kaushik Chowdhury

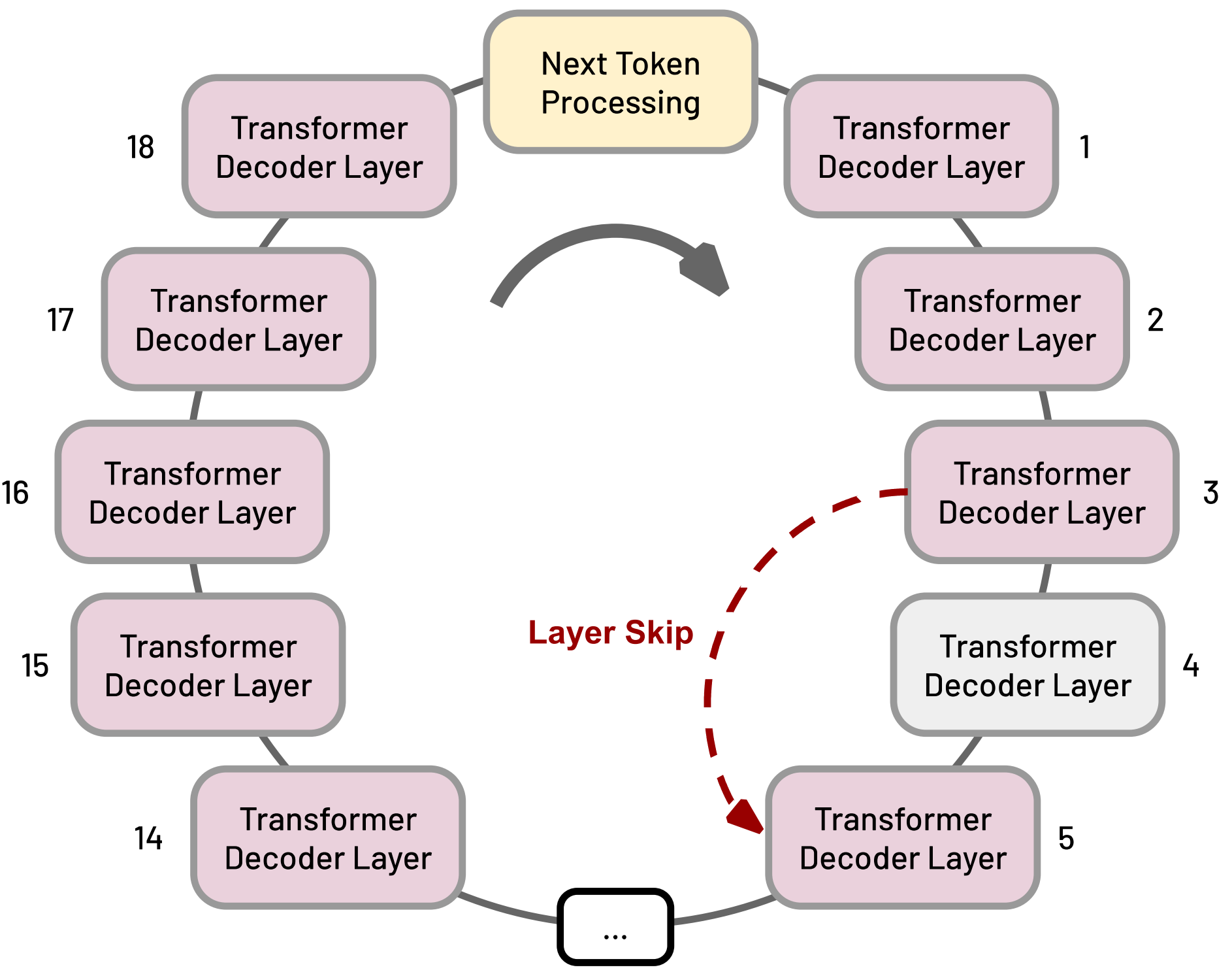
Need for Distributed LLMs

- Real-time decision support
- Avoid relying on remote infrastructure
- Intelligence analysis
- Communication assistance

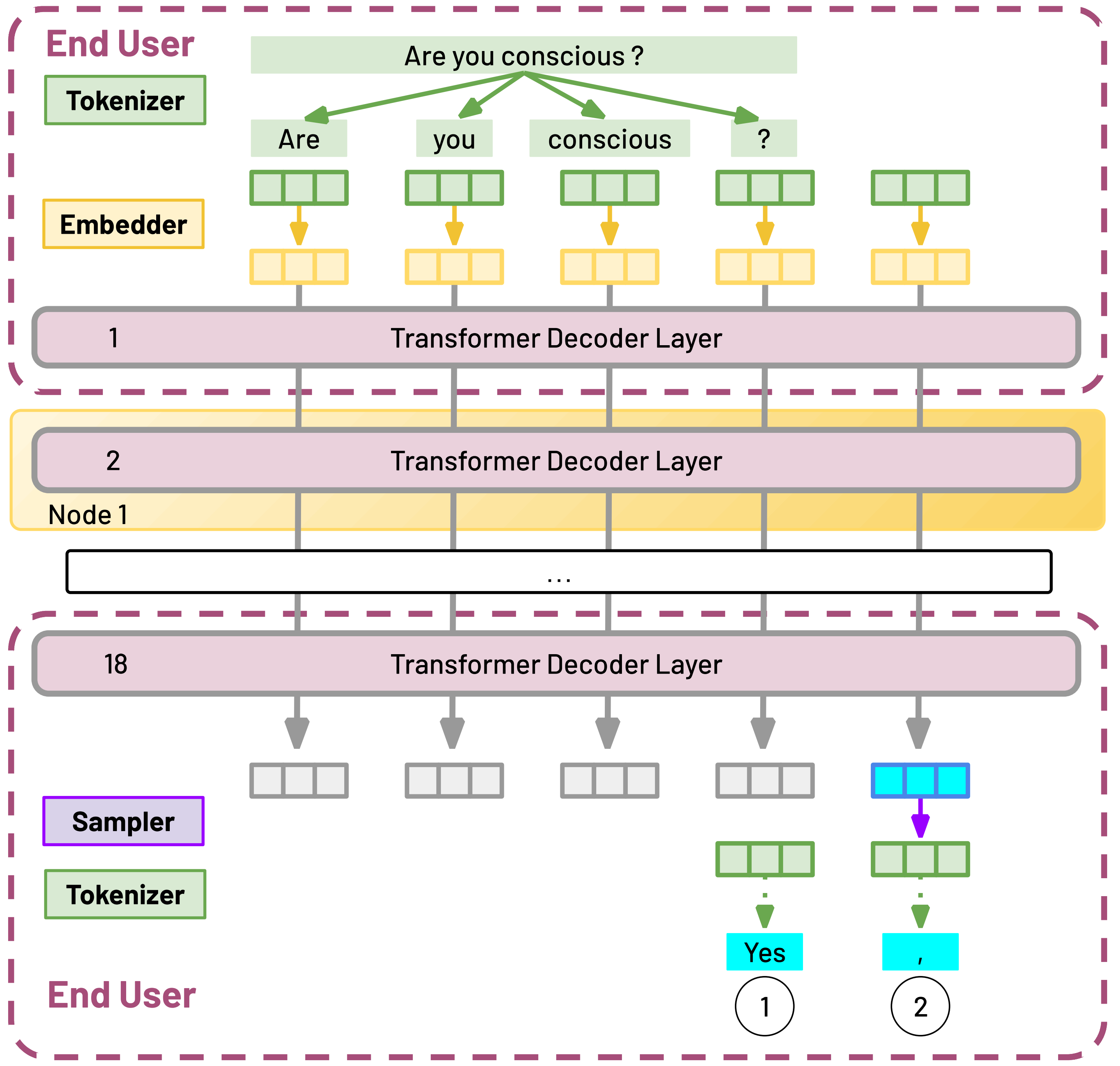
JARVIS: A distributed LLM orchestration framework.
- **Resilience**: Layer skipping enhances node failure resistance.
- **Recovery**: Supports peer communication and layer redundancy.
- **Efficiency**: Enables local LLM execution without centralized cloud reliance.

System Design

- Capitalizes on LLMs' resilience to partial loss of weights: When a node fails, its layer gets skipped.
- LLMs can handle informational gaps without significantly impacting performance.
- For example, if node 4 fails, **JARVIS** will bypass this node, redirecting the workflow directly from layer 3 to layer 5.


Next Token Processing
Transformer Decoder Layer 1
Transformer Decoder Layer 2
Transformer Decoder Layer 3
Transformer Decoder Layer 4
Transformer Decoder Layer 5
...
Transformer Decoder Layer 14
Transformer Decoder Layer 15
Transformer Decoder Layer 16
Transformer Decoder Layer 17
Transformer Decoder Layer 18
Layer Skip

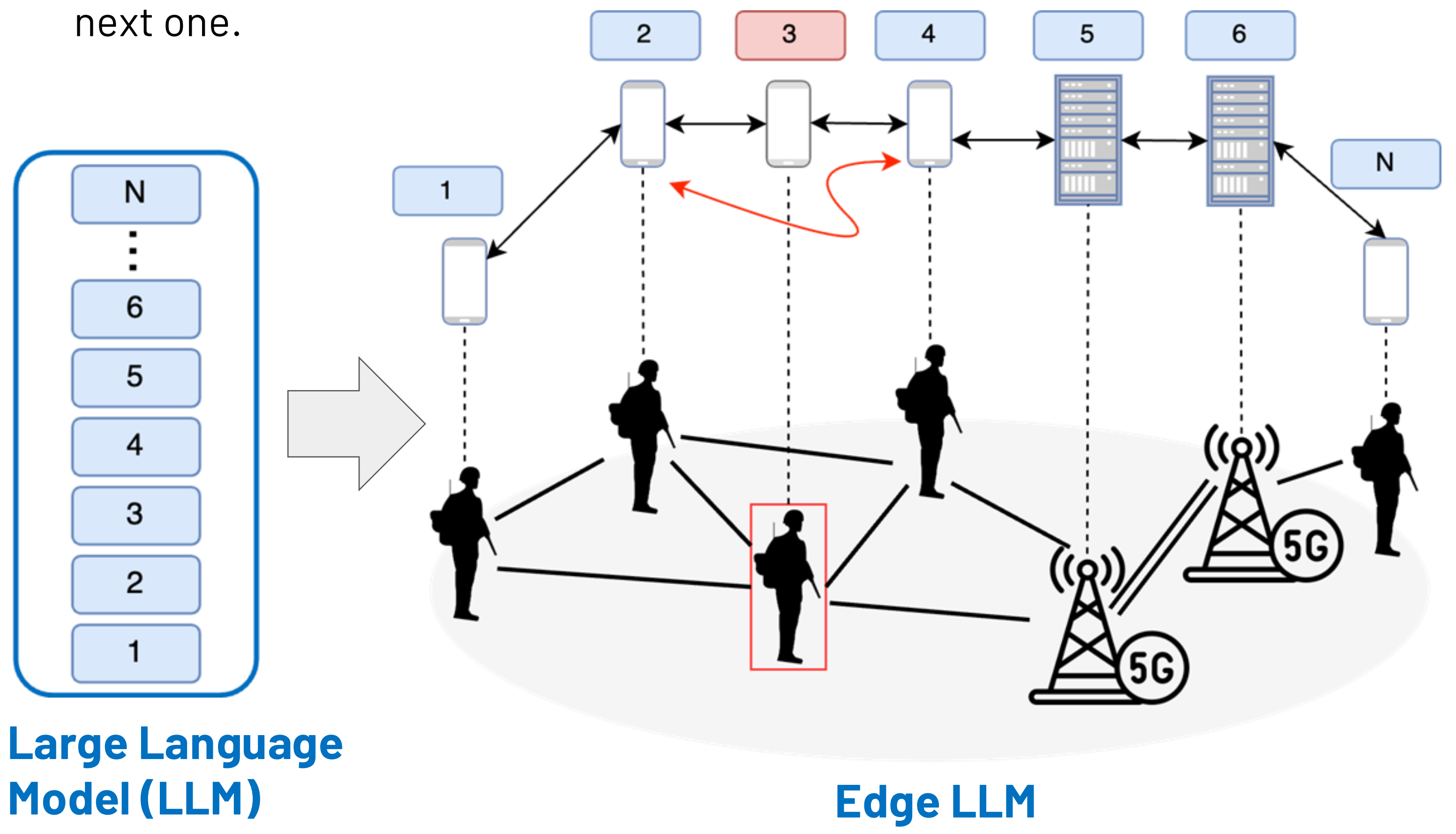
What runs where?



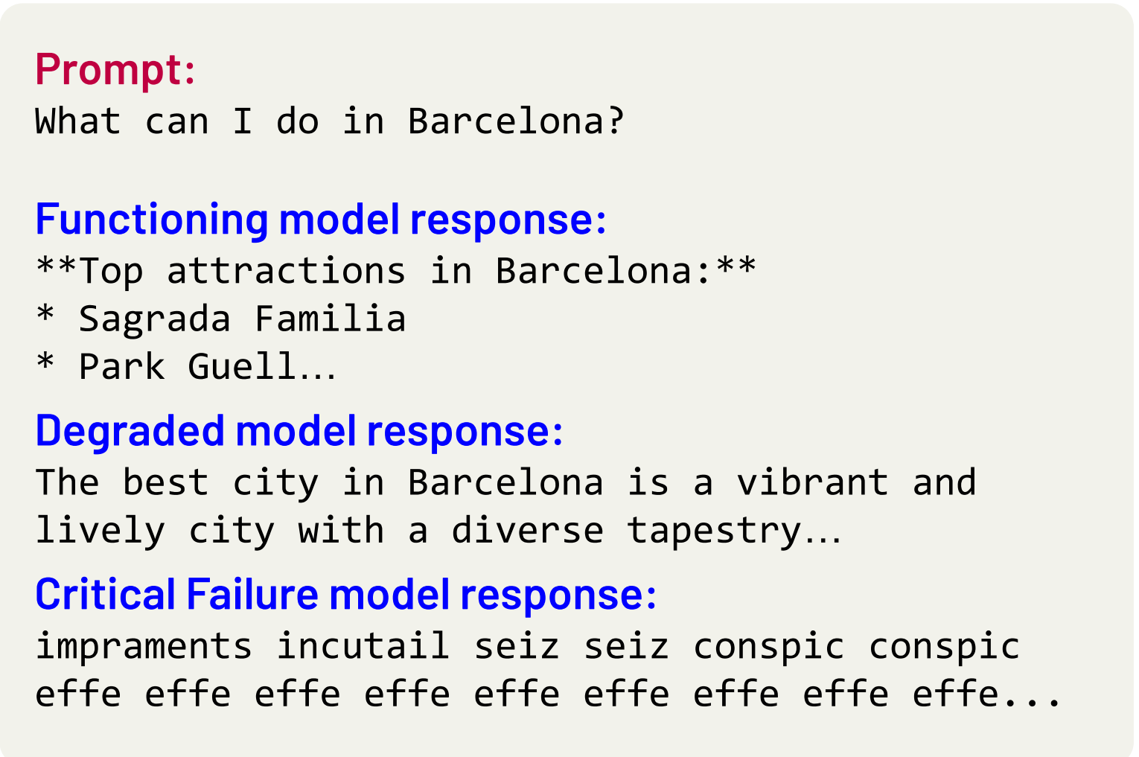
➔ Remember, text generation is a **causal** process, which means that we will generate our *words* one by one.

Resilience at the Center

➔ In **JARVIS**, resilience is achieved by avoiding central points of failure. If a node fails, it's skipped, and the previous node sends data directly to the next one.



What happens when we skip a layer?



Prompt:
```
What can I do in Barcelona?
```

Functioning model response:
```
**Top attractions in Barcelona:**
* Sagrada Familia
* Park Guell...
```

Degraded model response:
```
The best city in Barcelona is a vibrant and
lively city with a diverse tapestry...
```

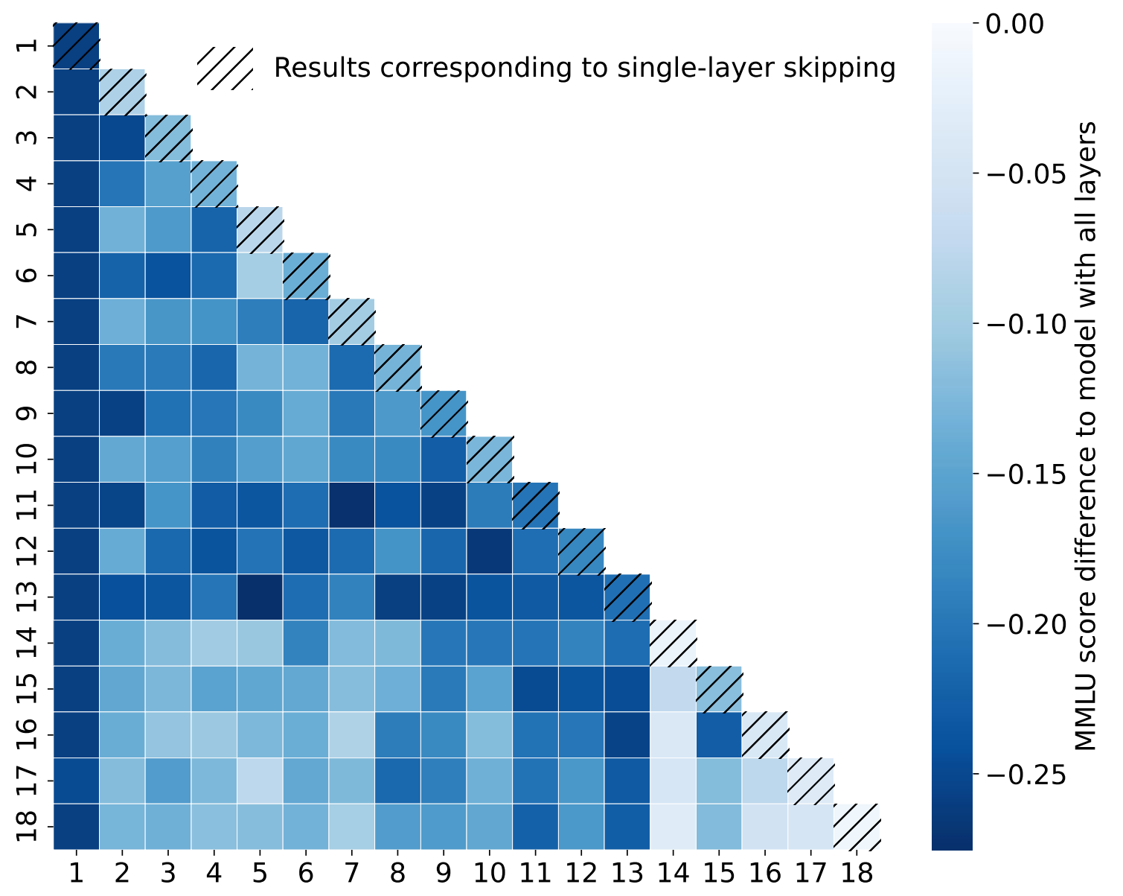
Critical Failure model response:
```
impraments incutail seiz seiz conspic conspic
effe effe effe effe effe effe effe effe effe...
```

The effect of skipping layers on the generation capacity of the model will depend on:

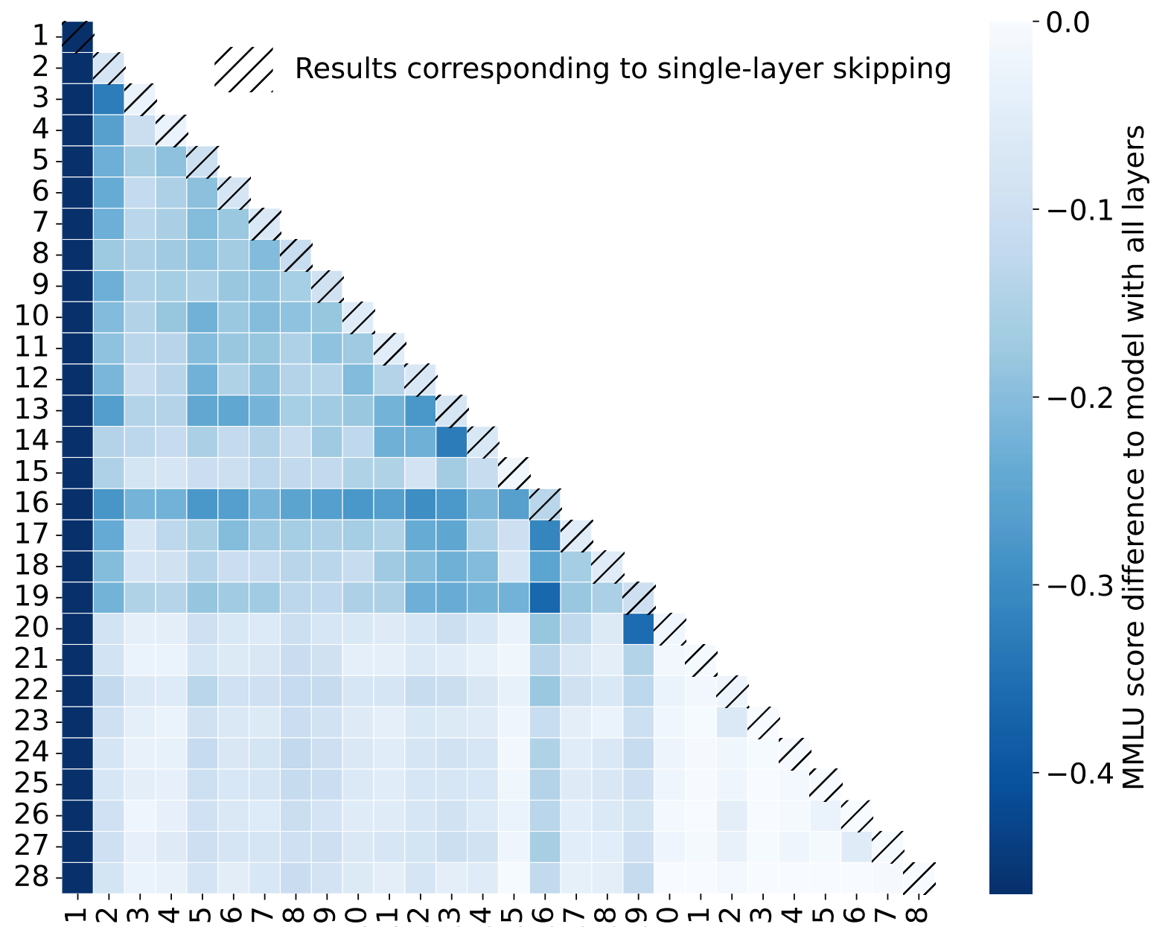
- The number of layers skipped.
  - For Gemma 2B, even skipping 5 layers at a time would not result in a critical failure.
- The specific layers involved.
  - First and last layers are critical.

★ We analyze the impact of skipping layers on the Massive Multitask Language Understanding (MMLU) benchmark. We compare full model performance to single and double layer skips in these two models:

Gemma 2B



Gemma 7B



➔ The omission of most of individual layers does not significantly impact performance.
➔ The 7B model shows less relative degradation than 2B, overparameterization helps.

Acknowledgment

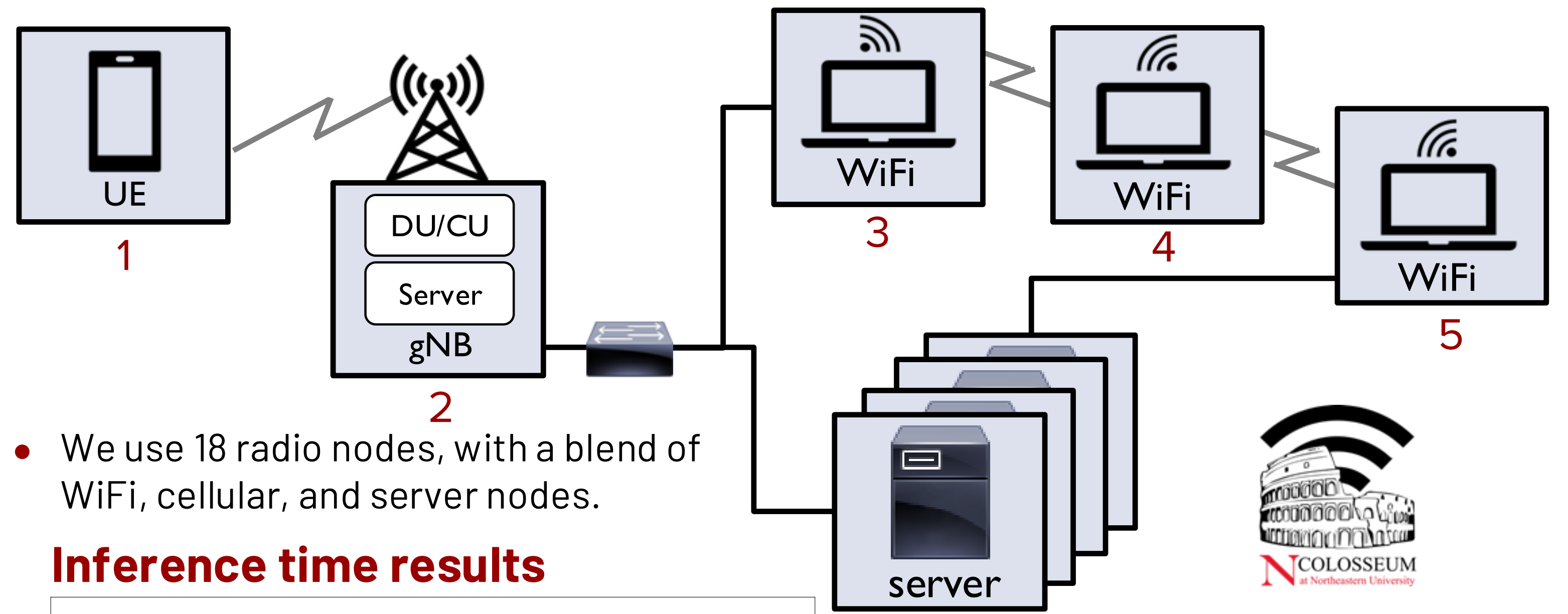
This work was supported in part by the U.S. National Science Foundation under grants CNS-1925601 and CNS-2112471.

Evaluation

NETWORK TOPOLOGY

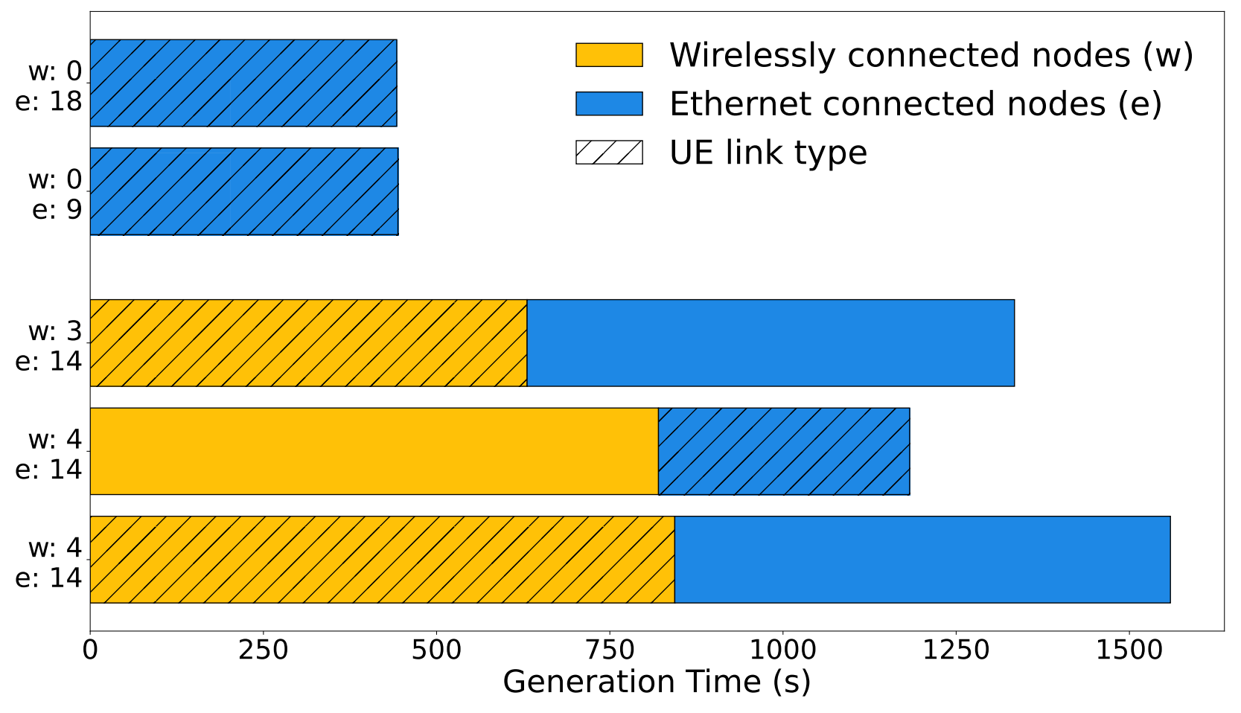
- **Model**: Gemma 2B, with 18 layers.
- **End User Role**: Acts as the prompt entry point and output display.
- **Node Role**: Acts as server for preceding node, client for subsequent one.
- **Node Failure Resilience**: Bypasses and reroutes around affected nodes.
- **Testbeds**: Evaluated on Colosseum and on a Raspberry Pis testbed.

Colosseum, the World's Largest Wireless Network Emulator

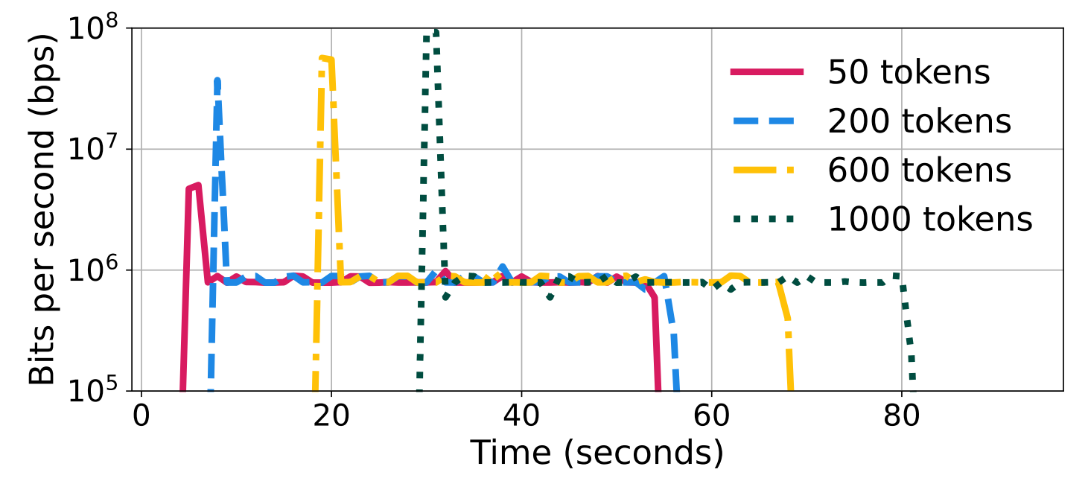


- We use 18 radio nodes, with a blend of WiFi, cellular, and server nodes.

Inference time results

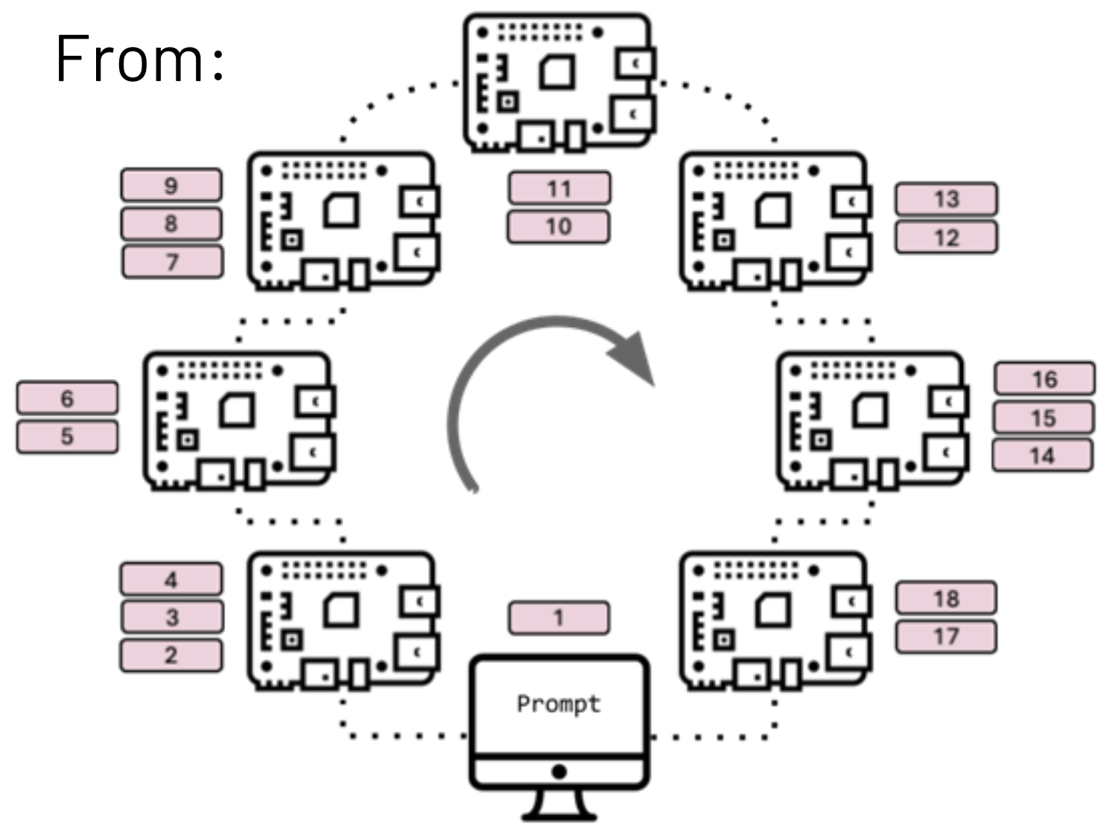


Network behaviour

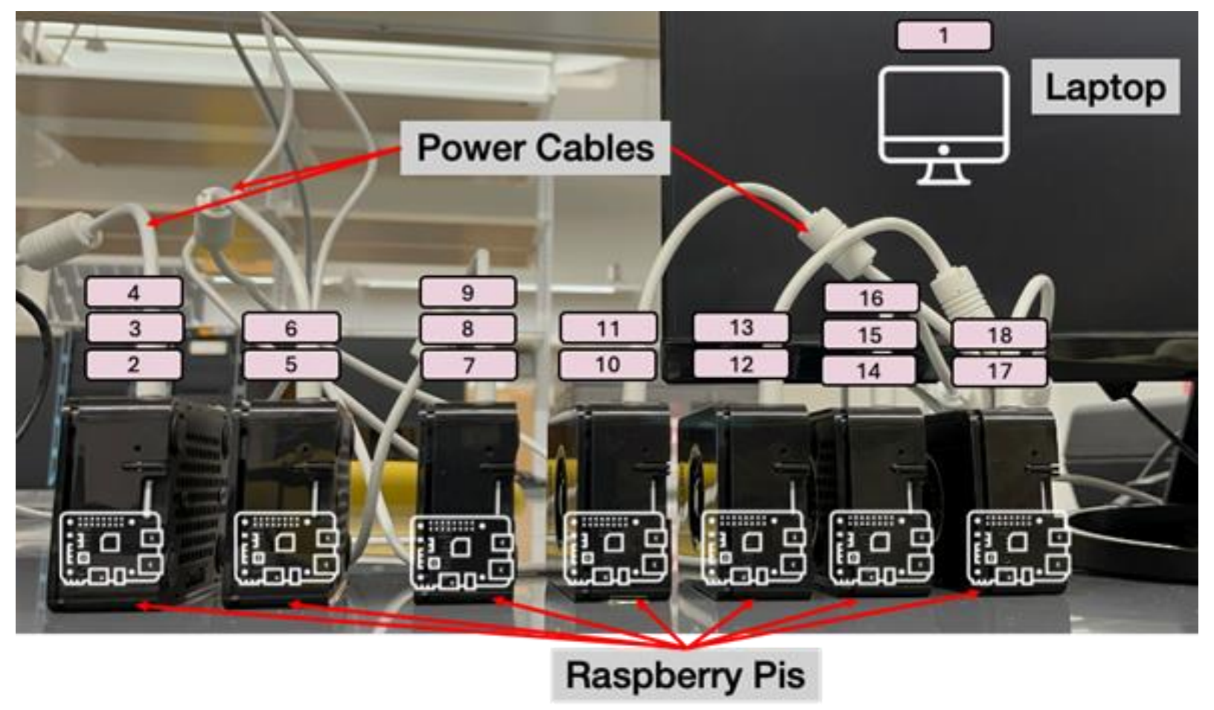


Our own Edge Implementation, a Raspberry Pis testbed

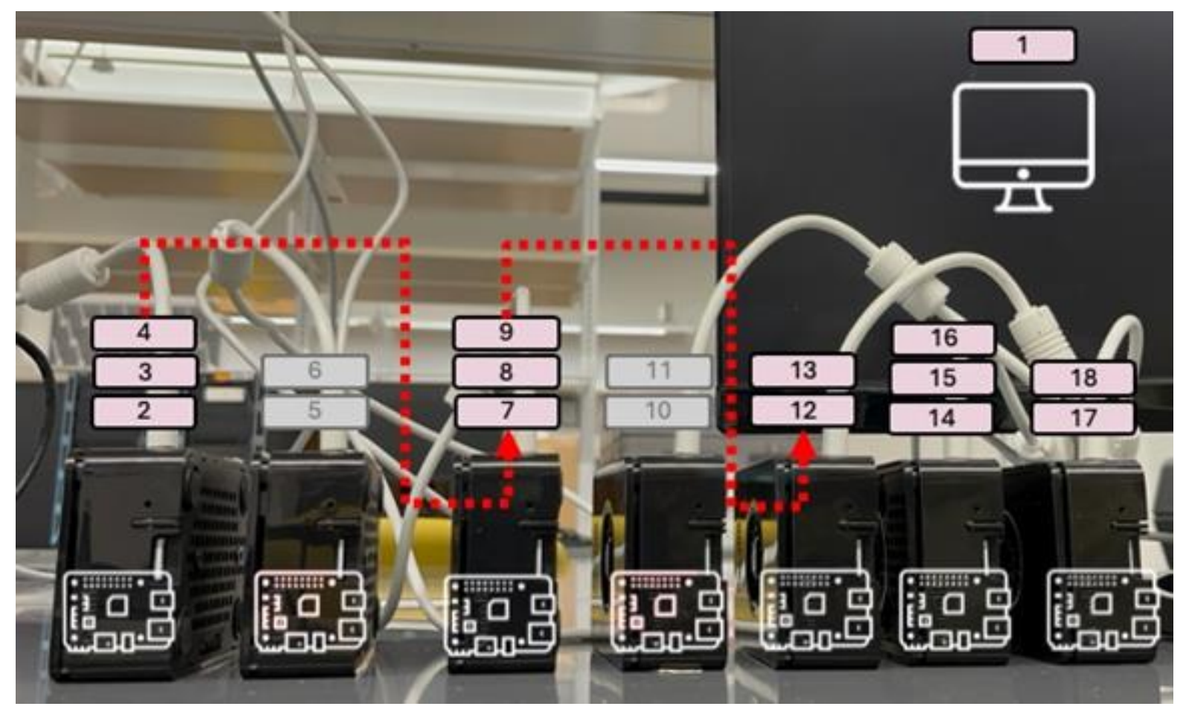
From:



To:

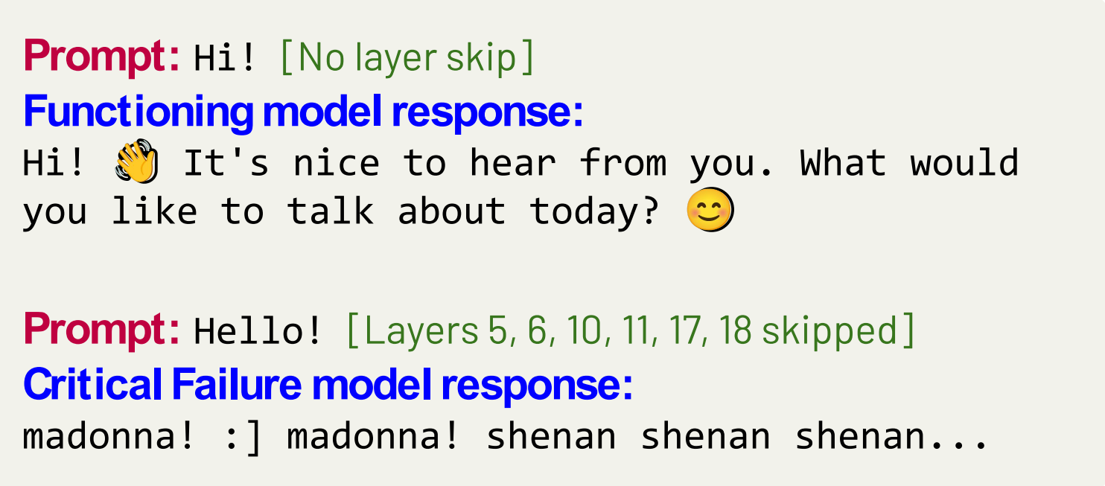


Layer skip:



- We use 7 Raspberry Pis as distributed nodes, allocating 2 to 3 layers at each device.
- The nodes are connected wirelessly through Wi-Fi.

➔ Example responses:



Prompt: `Hi!` [No layer skip]
Functioning model response:
```
Hi! 👋 It's nice to hear from you. What would
you like to talk about today? 😊
```

Prompt: `Hello!` [Layers 5, 6, 10, 11, 17, 18 skipped]
Critical Failure model response:
```
madonna! :] madonna! shenan shenan shenan...
```

Watch the full video at:

M. Sirera Perelló, J. Groen, W. Liu, S. Ioannidis and K. Chowdhury, "JARVIS: Disjoint Large Language Models on Radio VLANs for Intelligent Services," 2024 IEEE Military Communications Conference (MILCOM), Washington, DC, USA, Oct. 2024