



- Llama3-8B (20%)
- Llama3-70B (9%)
- Mistral-7B (33%)
- Janus-7B (34%)
- Mixtral-8x7B (26%)
- GPT-4o-mini (17%)
- DeepSeek-R1-70B (27%)
- Gemma-2-9B (12%)
- Gemma-2-27B (9%)
- Qwen3-8B (20%)
- Qwen3-8B-Thinking (23%)
- Qwen3-32B (21%)