**Syntax vs NLU Preference Accuracy**

Diverging comparison across models (T=0.0, Few-Shot)

■ Syntax Accuracy   ■ NLU Preference

| Model | NLU Preference | Syntax Accuracy |
|---|---|---|
| gemini-2.5-flash | 51.2% | 100.0% |
| glm-4.6 | 51.2% | 100.0% |
| gpt-oss-20b | 51.2% | 100.0% |
| gemini-2.5-pro | 51.2% | 100.0% |
| deepseek-v3.1 | 54.4% | 96.9% |
| kimi-k2-instruct | 55.0% | 95.0% |
| gemini-2.5-flash-lite | 55.0% | 81.2% |
| llama-3.3-70b-instruct | 51.2% | 75.0% |
| qwen3-next-80b-a3b-instruct | 38.1% | 68.1% |
| llama-3.2-3b-instruct | 65.0% | 67.5% |
| gemma-3-27b-it | 39.4% | 63.1% |
| llama-3.1-8b-instruct | 41.2% | 60.0% |
| qwen3-next-80b-a3b-thinking | 76.2% | 52.5% |
| mixtral-8x22b-instruct | 76.2% | 52.5% |
| llama-3.2-1b-instruct | 24.4% | 46.9% |

← NLU Preference (%)  |  Syntax Accuracy (%) →