Hallucination Detection Method Comparison

Detection Rate by Method

- BERTScore: 73.0%
- NLI: 7.0%

Score Distribution
- BERTScore
- NLI

Hallucination by Category

- Scientific: 44.0%
- TruthfulQA: 42.0%
- Medical: 38.0%
- Factual_Knowledge: 36.0%

Model Performance

- qwen2.5:3b
- deepseek-llm
- mistral:7b
- llama3:8b
- gemma:7b