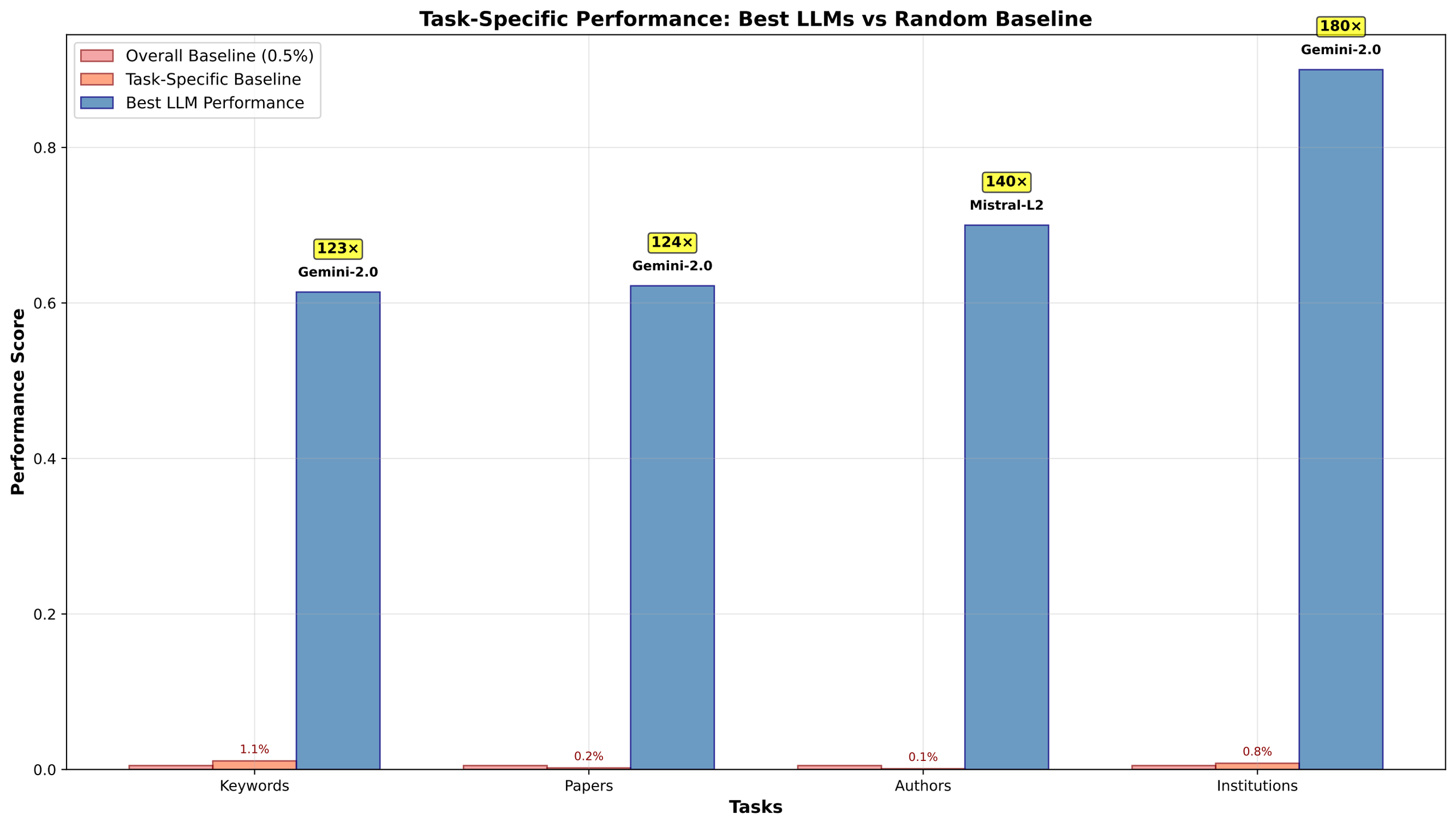
Task-Specific Performance: Best LLMs vs Random Baseline
Overall Baseline (0.5%)
Task-Specific Baseline
Best LLM Performance
Performance Score
0.0
0.2
0.4
0.6
0.8
123×
Gemini-2.0
124×
Gemini-2.0
140×
Mistral-L2
180×
Gemini-2.0
1.1%
0.2%
0.1%
0.8%
Keywords
Papers
Authors
Institutions
Tasks