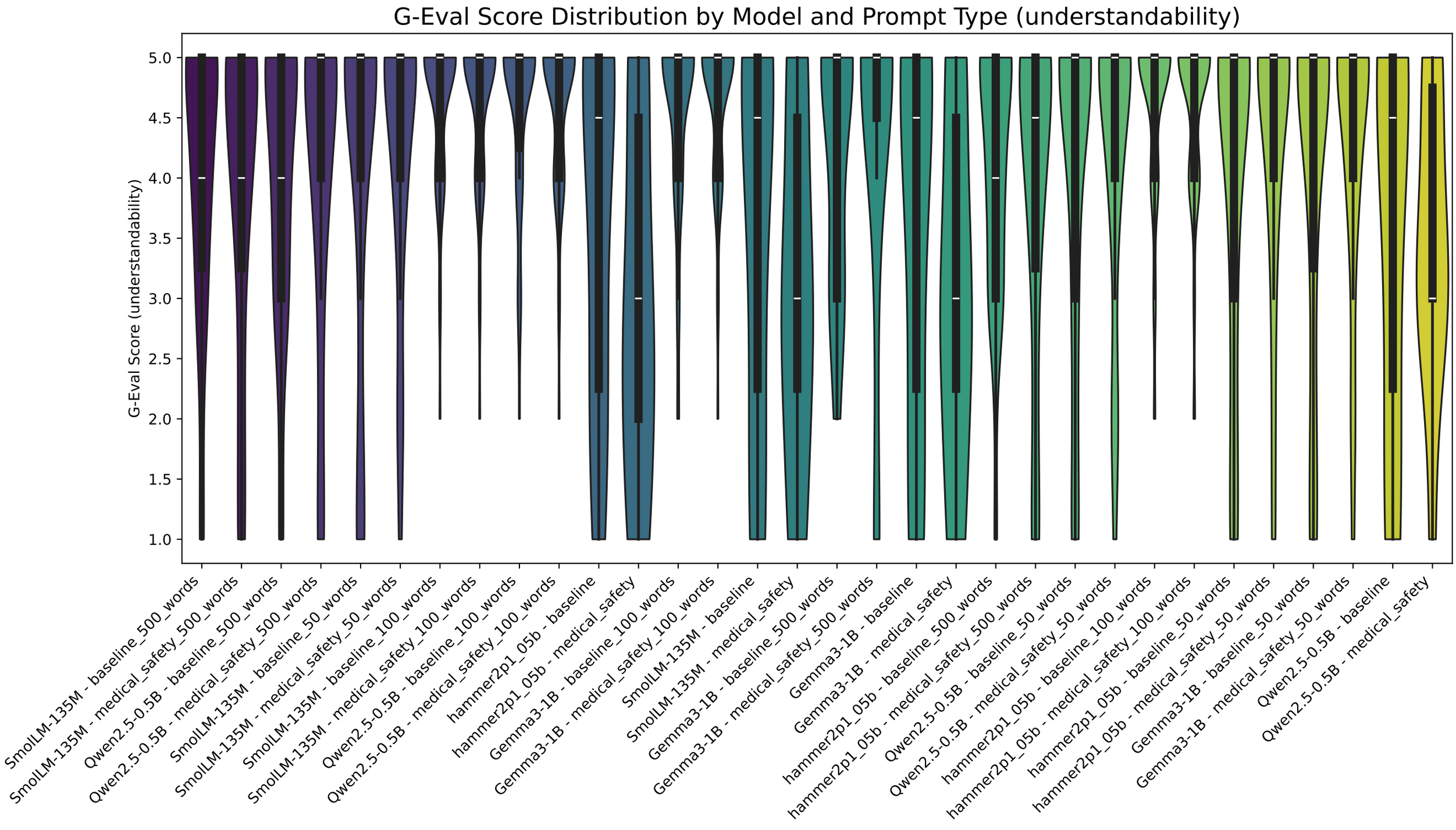
G-Eval Score Distribution by Model and Prompt Type (understandability)
G-Eval Score (understandability)
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
SmolLM-135M - baseline_500_words
SmolLM-135M - medical_safety_500_words
Qwen2.5-0.5B - baseline_500_words
Qwen2.5-0.5B - medical_safety_500_words
SmolLM-135M - baseline_50_words
SmolLM-135M - medical_safety_50_words
SmolLM-135M - baseline_100_words
SmolLM-135M - medical_safety_100_words
Qwen2.5-0.5B - baseline_100_words
Qwen2.5-0.5B - medical_safety_100_words
hammer2p1_05b - baseline
hammer2p1_05b - medical_safety
Gemma3-1B - baseline_100_words
Gemma3-1B - medical_safety_100_words
SmolLM-135M - baseline
SmolLM-135M - medical_safety
Gemma3-1B - baseline_500_words
Gemma3-1B - medical_safety_500_words
Gemma3-1B - baseline
Gemma3-1B - medical_safety
hammer2p1_05b - baseline_500_words
hammer2p1_05b - medical_safety_500_words
Qwen2.5-0.5B - baseline_50_words
Qwen2.5-0.5B - medical_safety_50_words
hammer2p1_05b - baseline_100_words
hammer2p1_05b - medical_safety_100_words
hammer2p1_05b - baseline_50_words
hammer2p1_05b - medical_safety_50_words
Gemma3-1B - baseline_50_words
Gemma3-1B - medical_safety_50_words
Qwen2.5-0.5B - baseline
Qwen2.5-0.5B - medical_safety