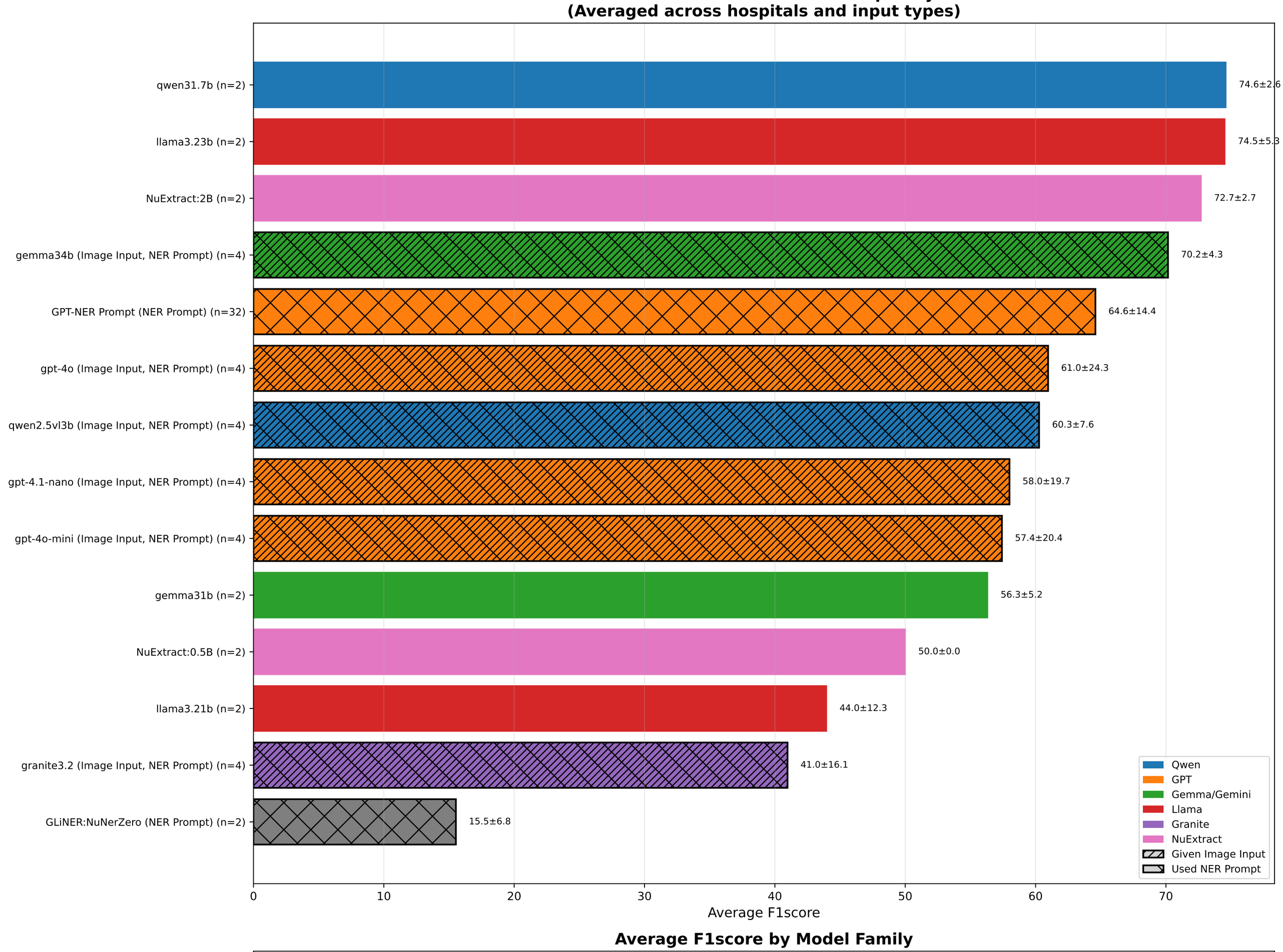
COMPREHENSIVE LLM PERFORMANCE ANALYSIS
==========
Report Generated: 2025-06-29 13:59:44

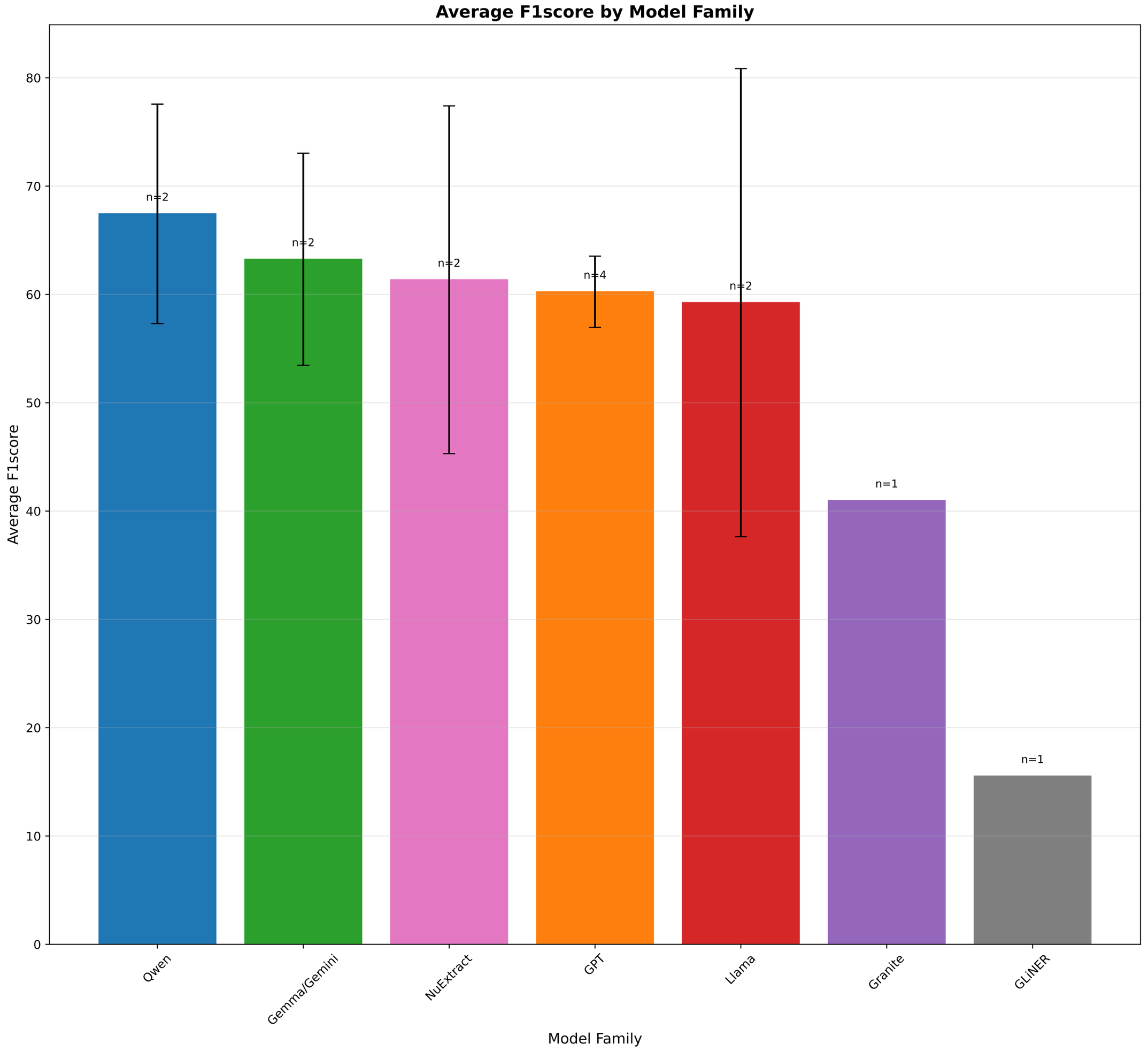
OVERALL PERFORMANCE METRICS:
- Average F1 Score: 60.476 (Std: 17.169)
- Average Accuracy: 45.408 (Std: 16.497)
- Average Precision: 64.073 (Std: 20.554)
- Average Recall: 60.364 (Std: 17.576)

GROUPED MODEL F1 SCORE STATISTICS:
- Unique Base Models: 14
- Total Test Instances: 70
- Best Performing Model: qwen31.7b (F1: 74.60)
- Worst Performing Model: GLiNER:NuNerZero (F1: 15.53)
- Overall Average F1: 57.14
- Models with Vision: 6

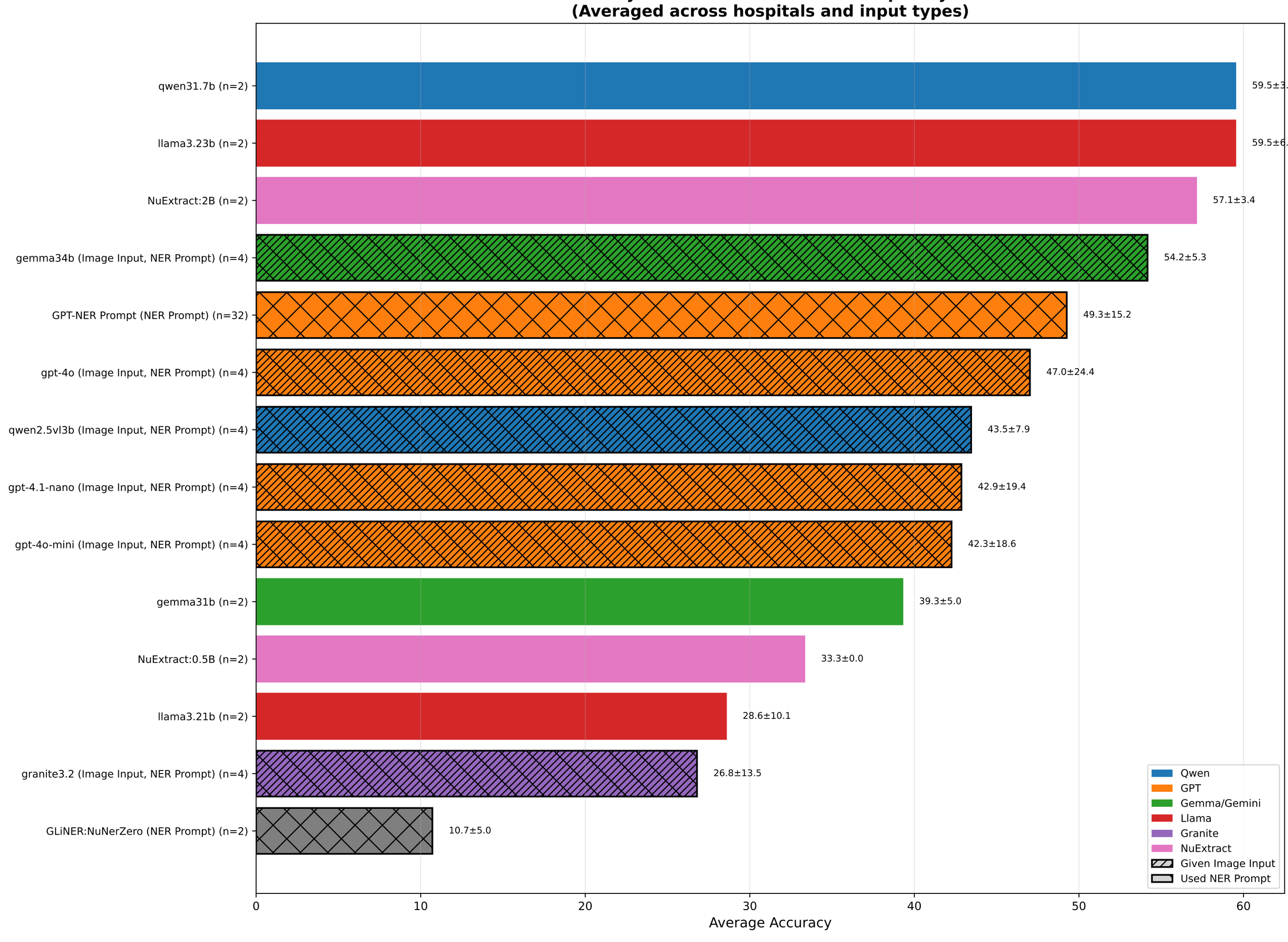
Overall F1score Performance - Models Grouped by Base Name
(Averaged across hospitals and input types)



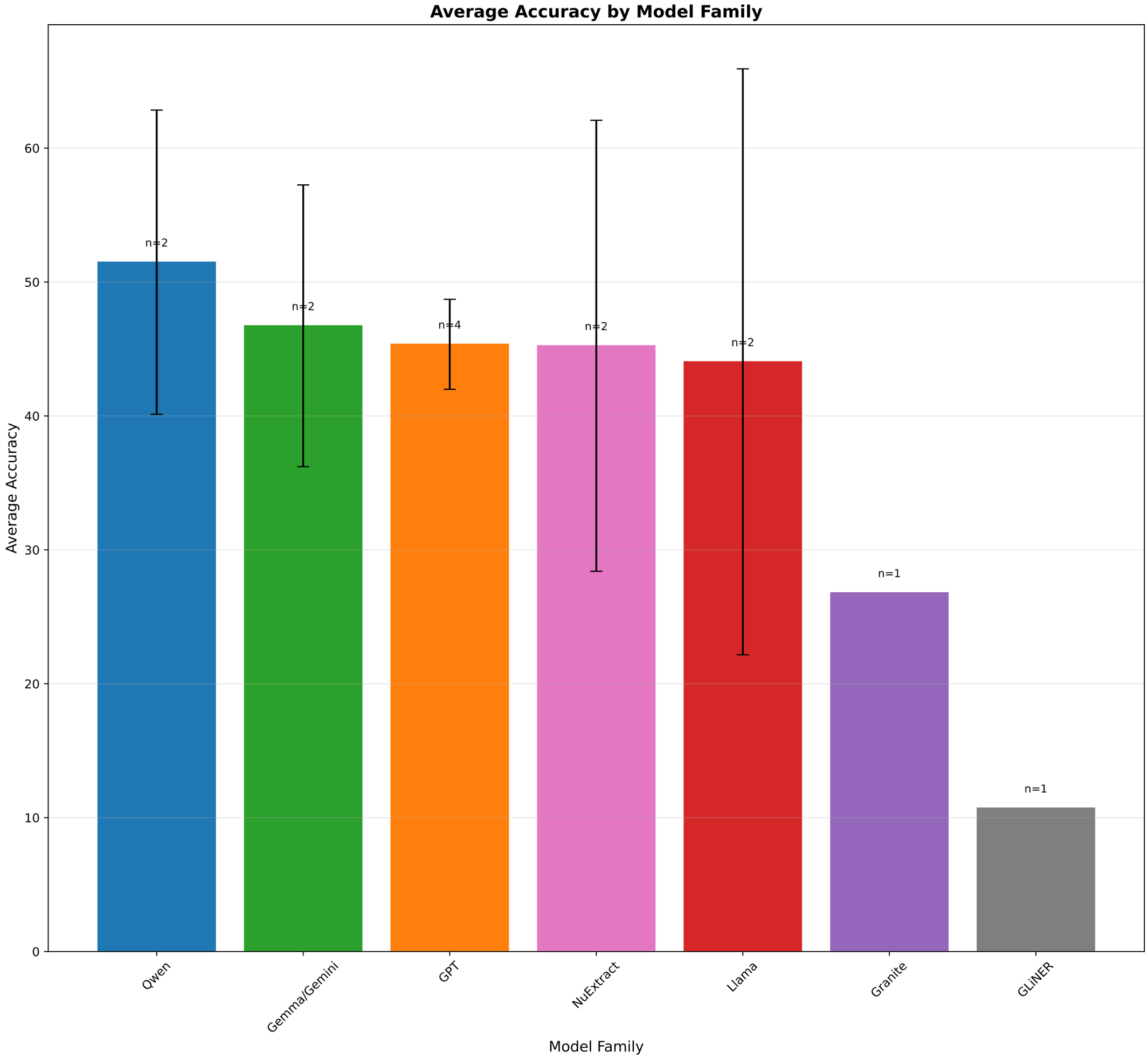
Average F1score by Model Family



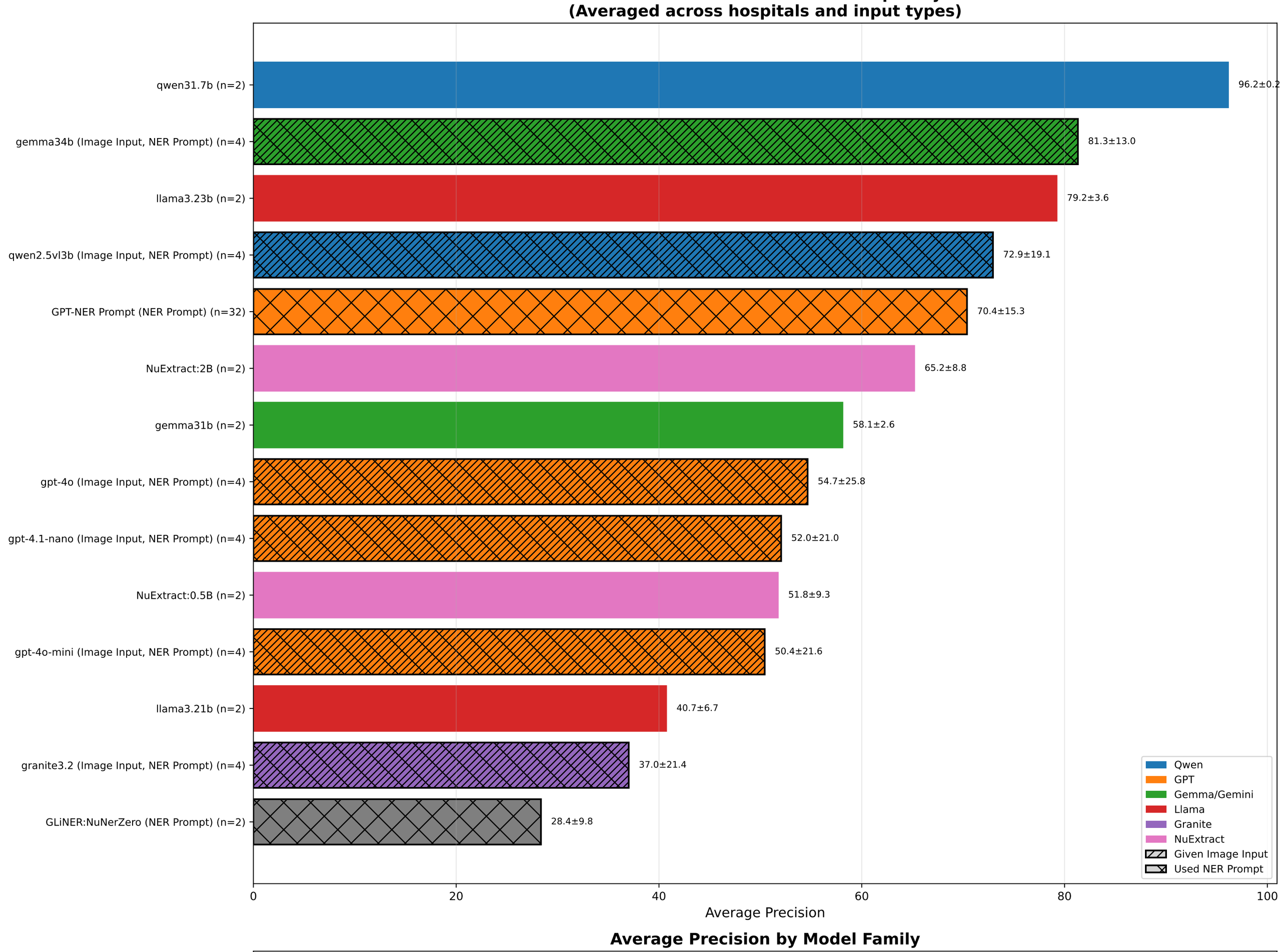
Overall Accuracy Performance - Models Grouped by Base Name
(Averaged across hospitals and input types)



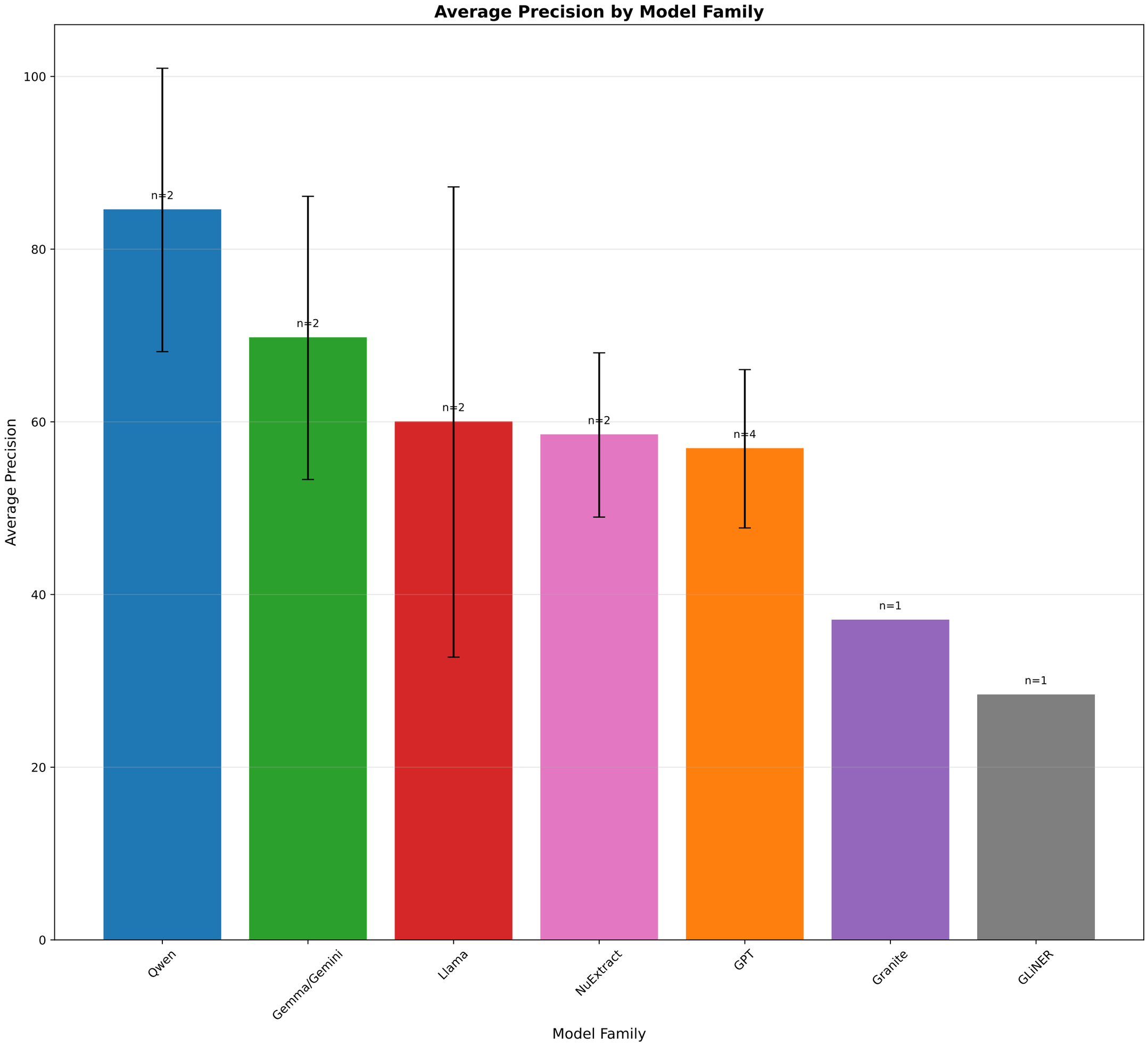
Average Accuracy by Model Family



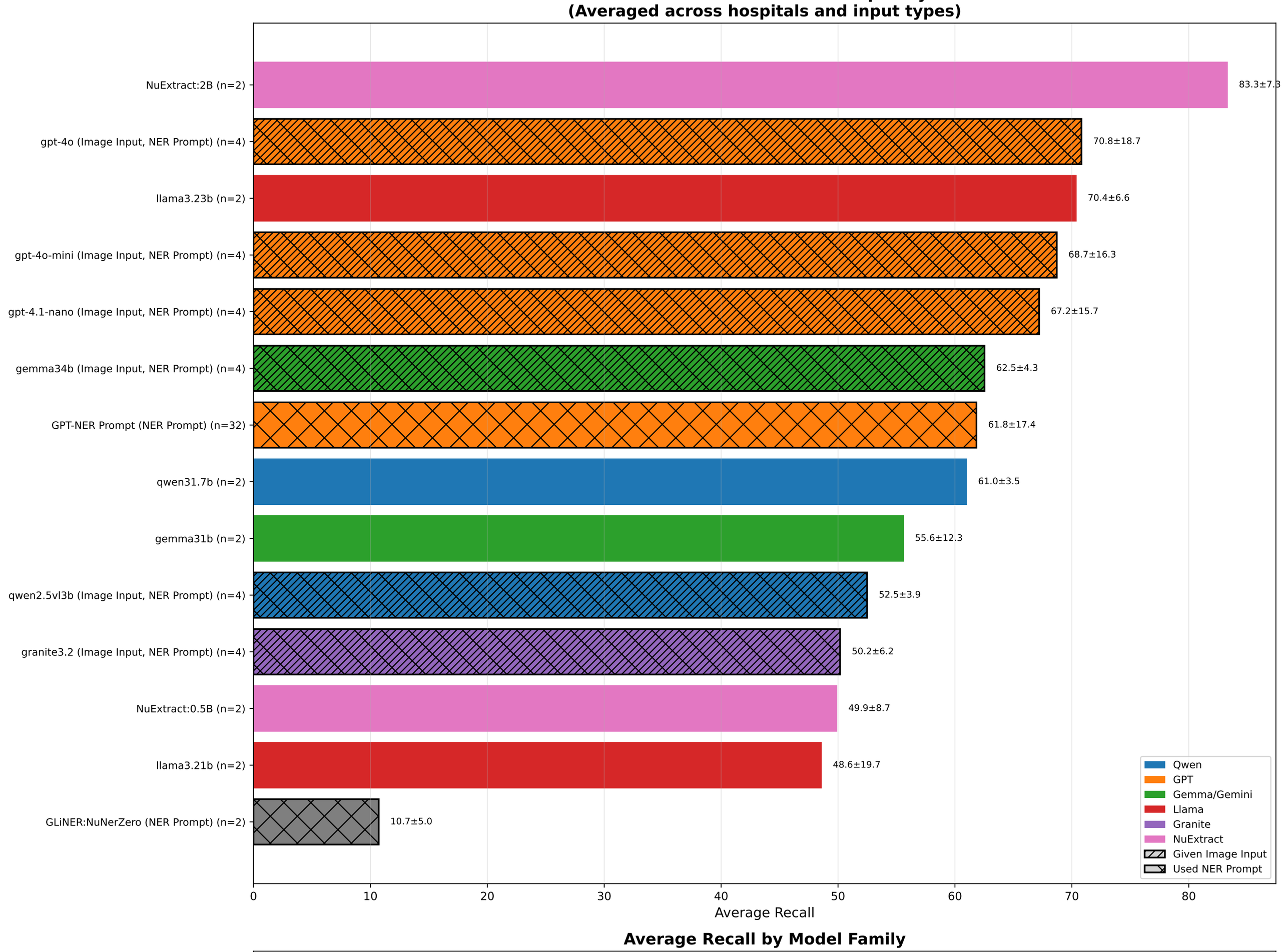
Overall Precision Performance - Models Grouped by Base Name
(Averaged across hospitals and input types)



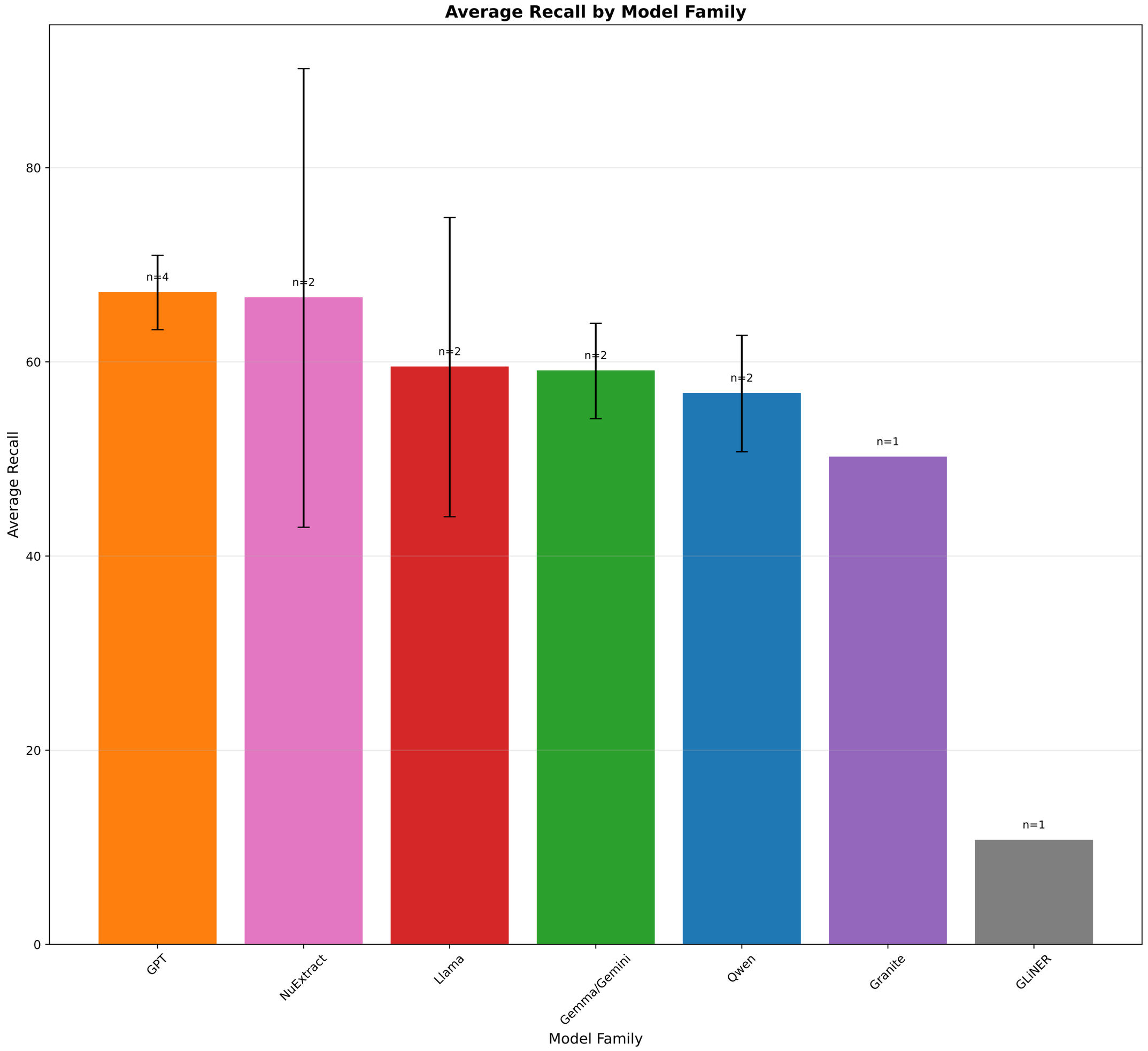
Average Precision by Model Family

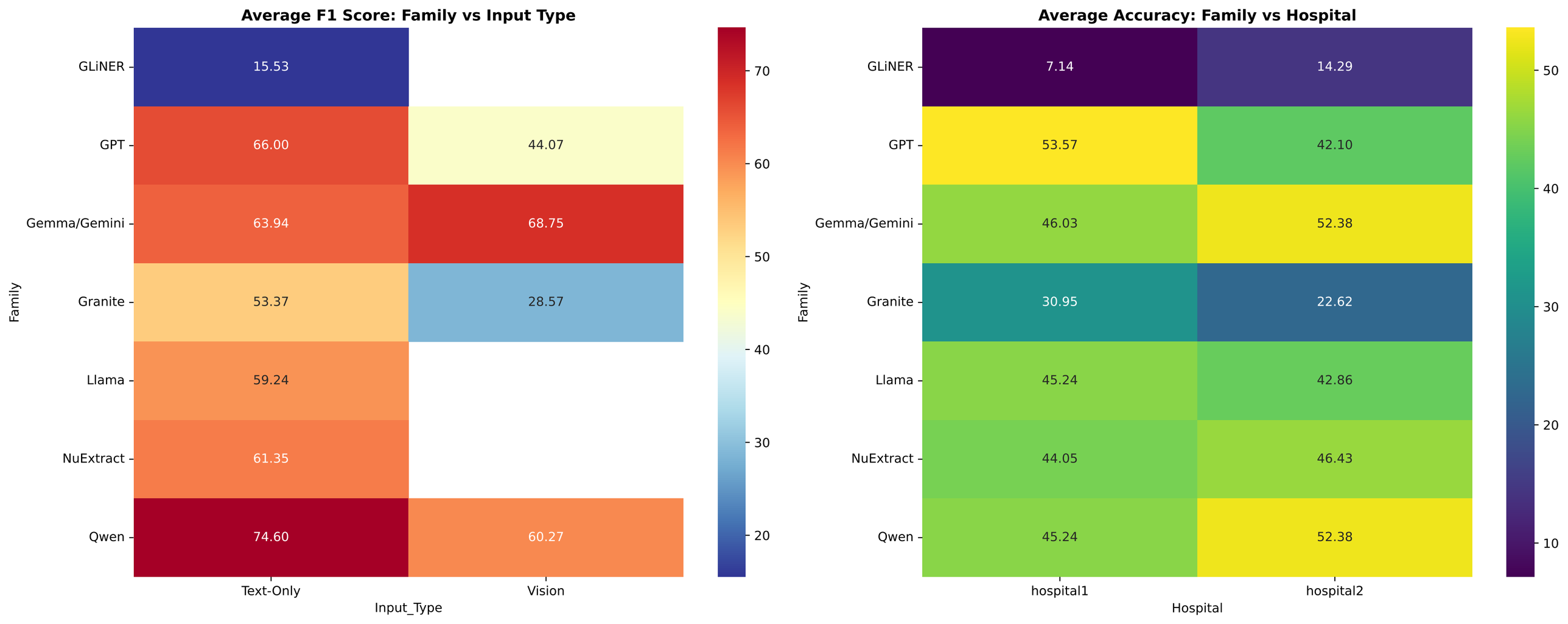


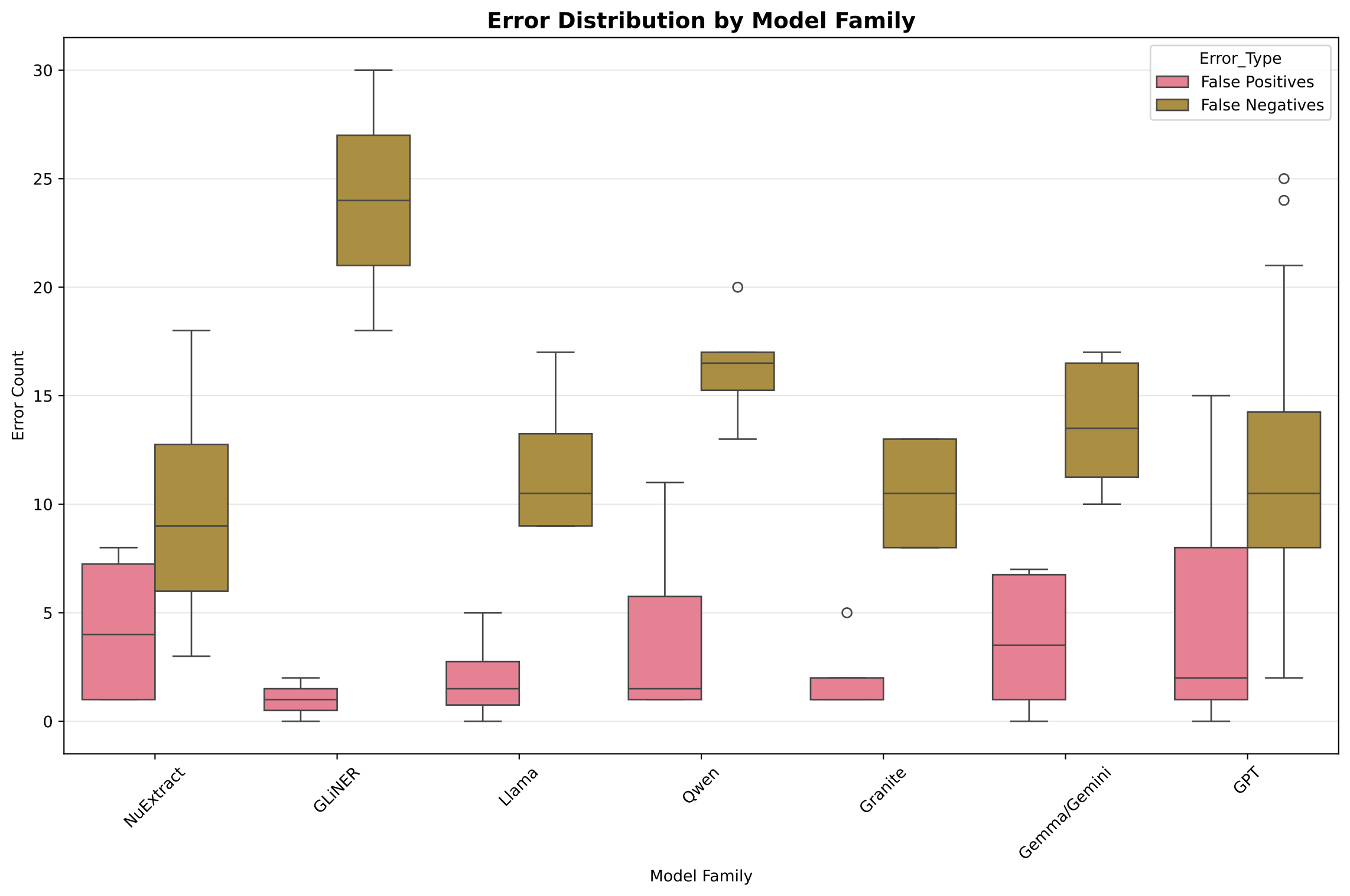
Overall Recall Performance - Models Grouped by Base Name
(Averaged across hospitals and input types)

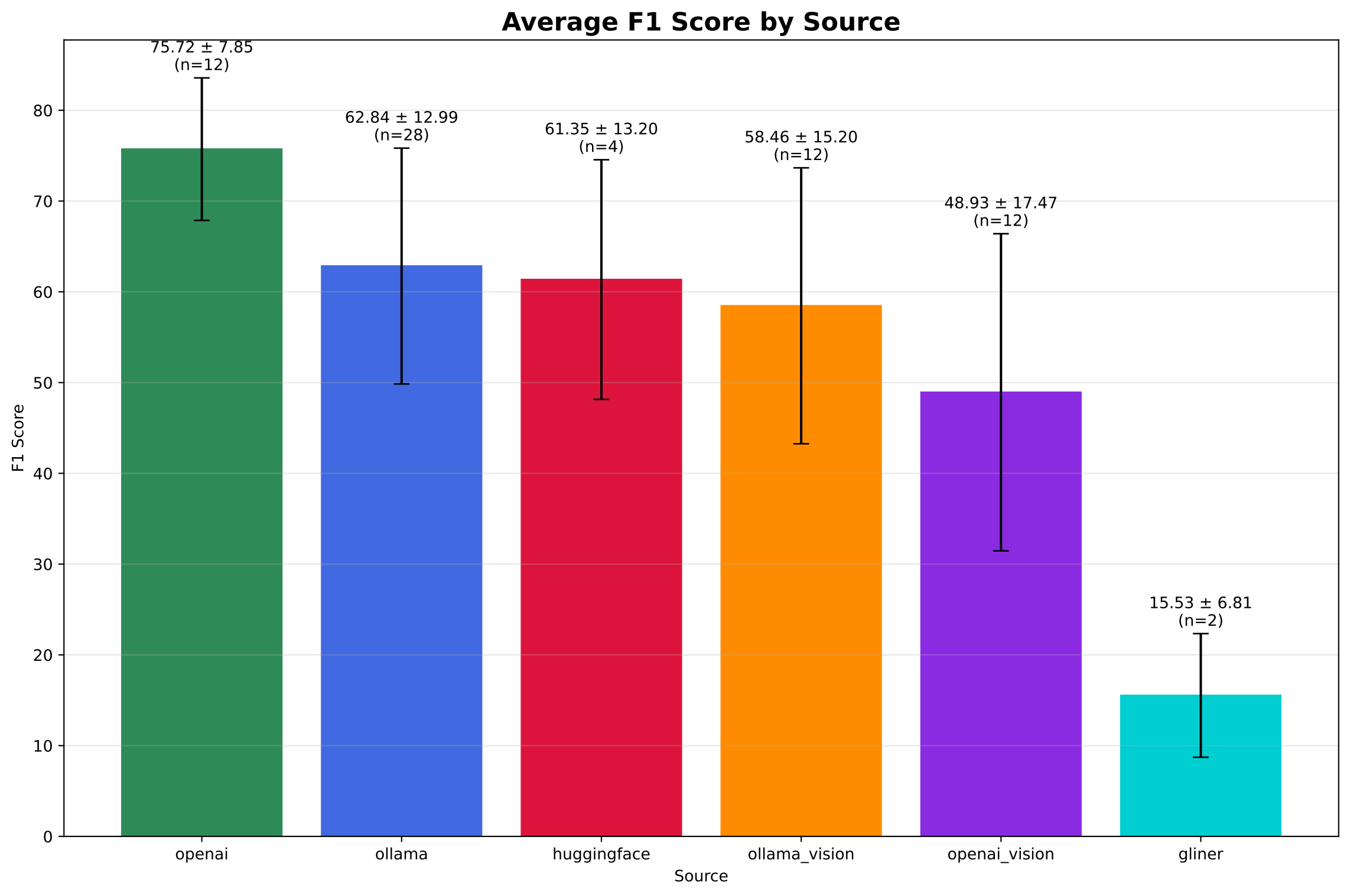


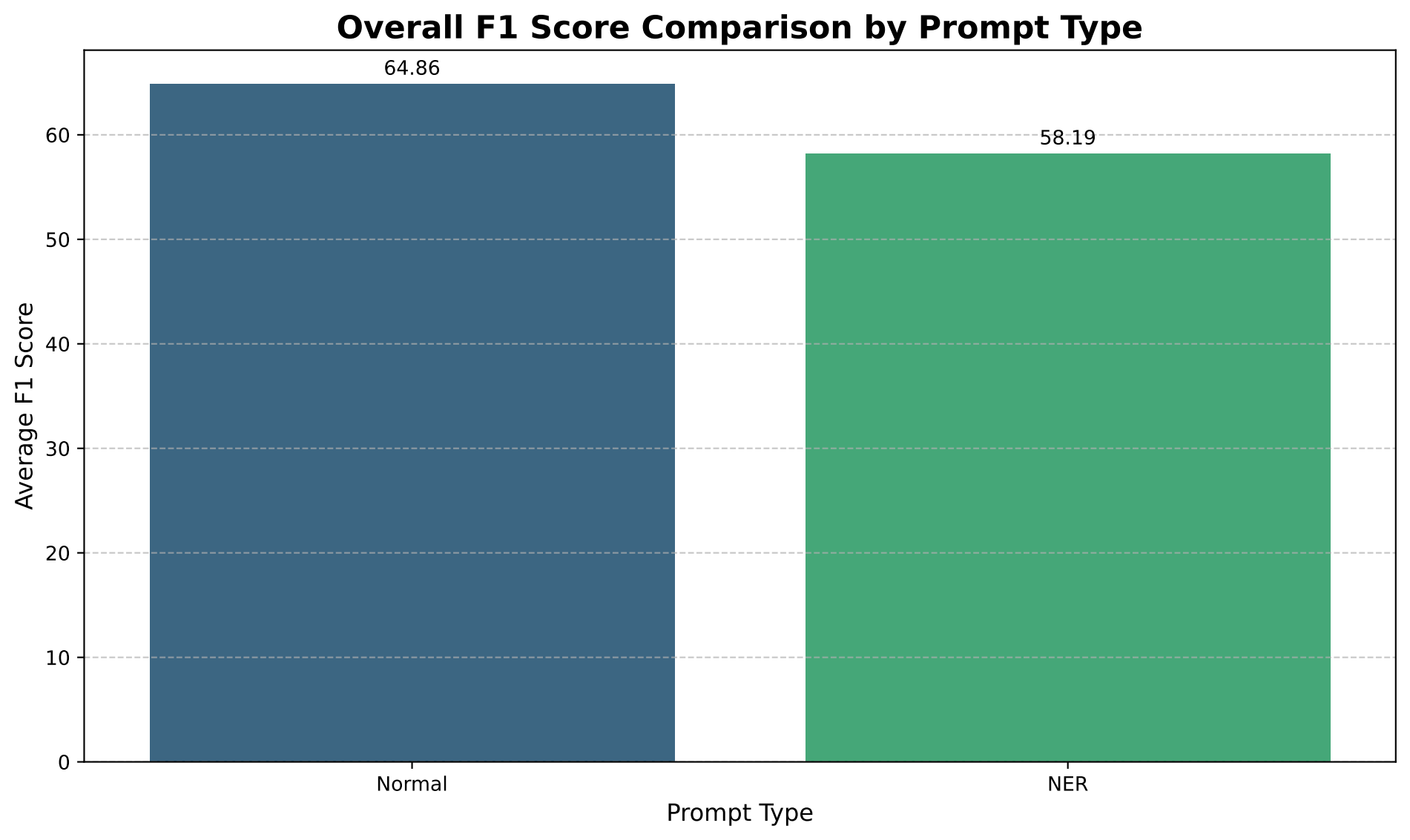
Average Recall by Model Family



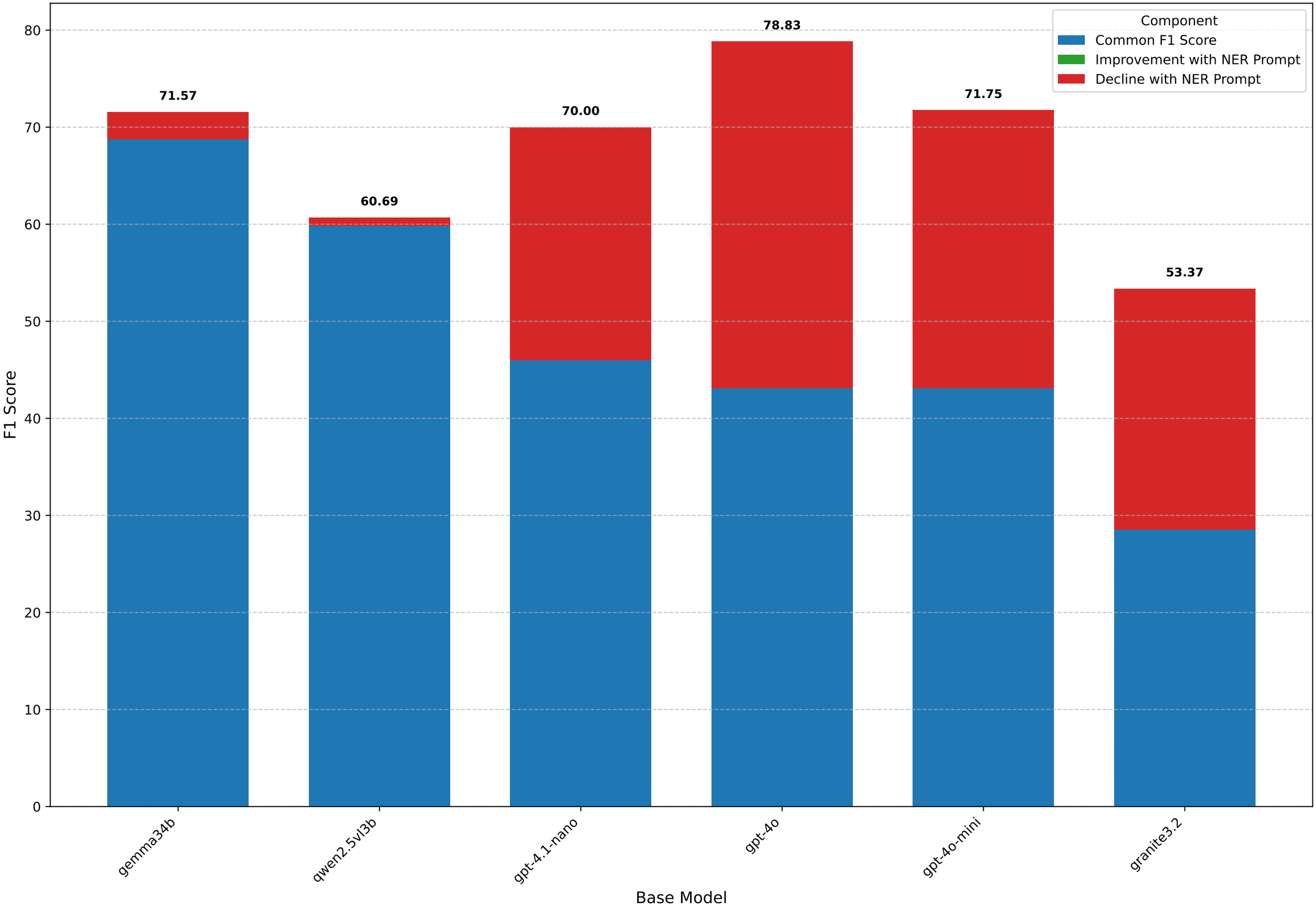

Average F1 Score: Family vs Input Type
GLiNER
GPT
Gemma/Gemini
Granite
Llama
NuExtract
Qwen
Family
15.53
66.00
44.07
63.94
68.75
53.37
28.57
59.24
61.35
74.60
60.27
Text-Only
Vision
Input_Type
70
60
50
40
30
20
Average Accuracy: Family vs Hospital
GLiNER
GPT
Gemma/Gemini
Granite
Llama
NuExtract
Qwen
Family
7.14
14.29
53.57
42.10
46.03
52.38
30.95
22.62
45.24
42.86
44.05
46.43
45.24
52.38
hospital1
hospital2
Hospital
50
40
30
20
10

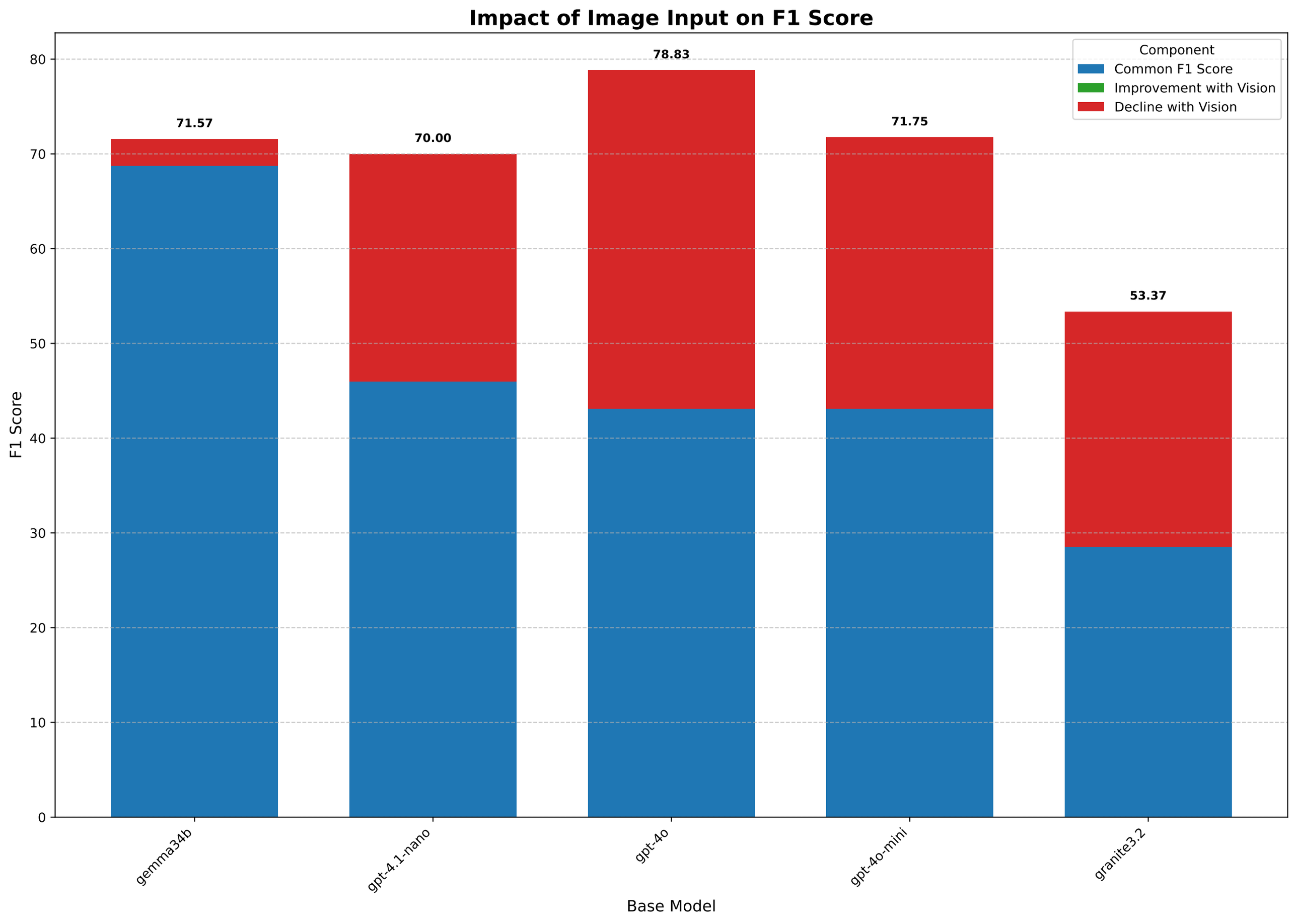

Error Distribution by Model Family
Error_Type
False Positives
False Negatives
Error Count
0
5
10
15
20
25
30
NuExtract
GLiNER
Llama
Qwen
Granite
Gemma/Gemini
GPT
Model Family


Average F1 Score by Source
75.72 ± 7.85
(n=12)
62.84 ± 12.99
(n=28)
61.35 ± 13.20
(n=4)
58.46 ± 15.20
(n=12)
48.93 ± 17.47
(n=12)
15.53 ± 6.81
(n=2)
F1 Score
0
10
20
30
40
50
60
70
80
openai
ollama
huggingface
ollama_vision
openai_vision
gliner
Source


Overall F1 Score Comparison by Prompt Type
64.86
58.19
Average F1 Score
0
10
20
30
40
50
60
Normal
NER
Prompt Type

Impact of NER Prompt on F1 Score




Impact of Image Input on F1 Score
Component
Common F1 Score
Improvement with Vision
Decline with Vision
F1 Score
0
10
20
30
40
50
60
70
80
71.57
70.00
78.83
71.75
53.37
gemma34b
gpt-4.1-nano
gpt-4o
gpt-4o-mini
granite3.2
Base Model