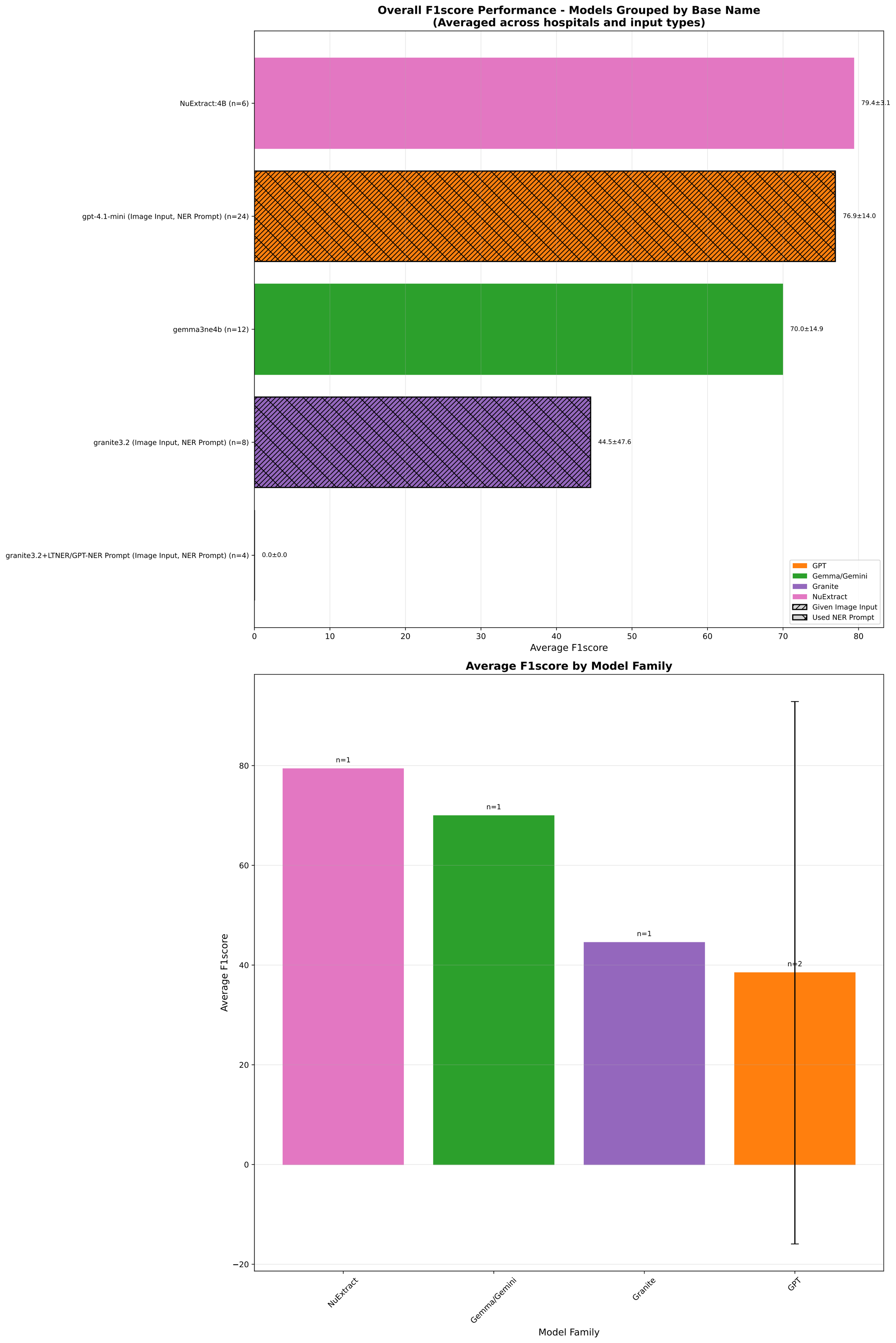
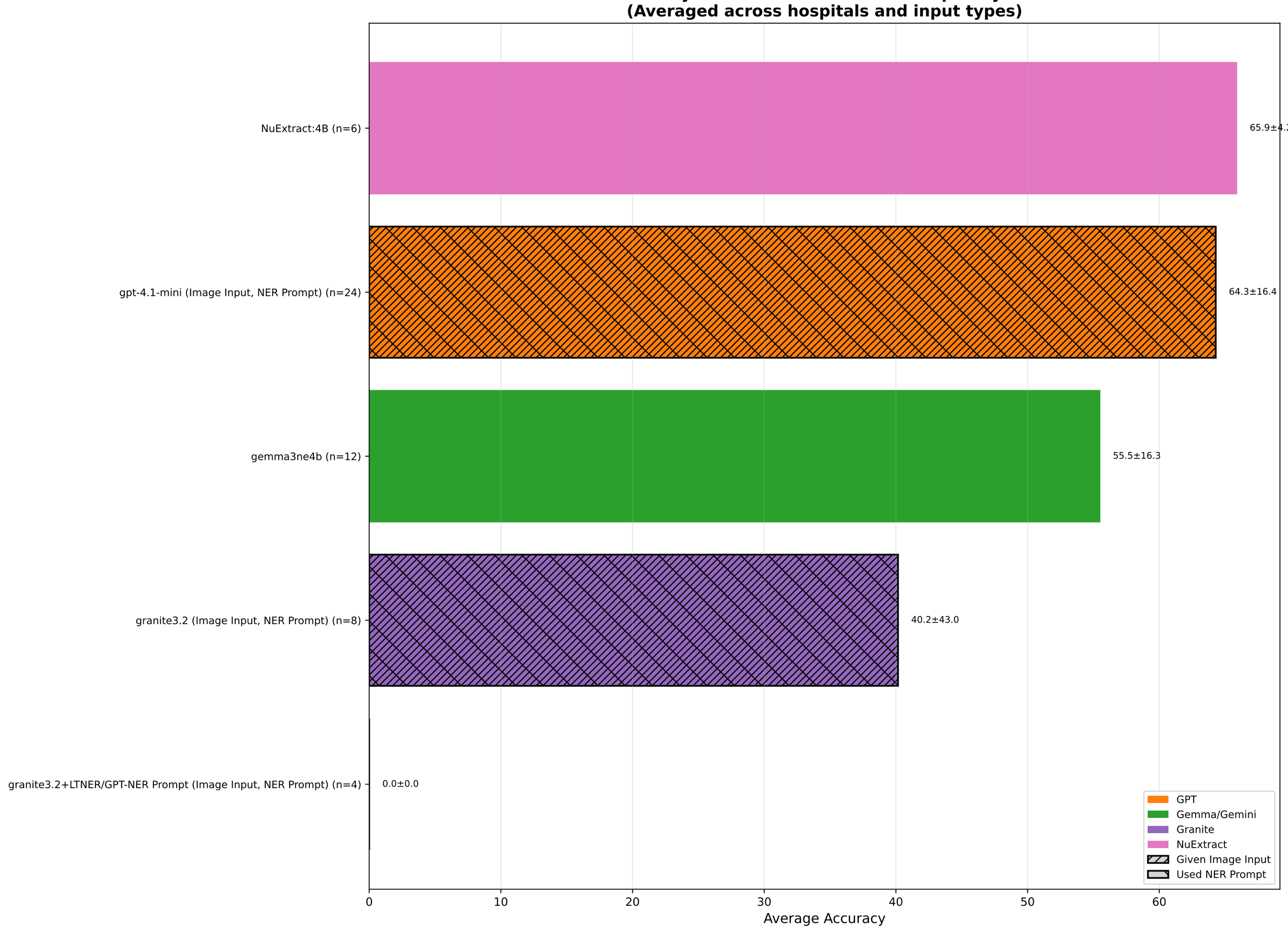
COMPREHENSIVE LLM PERFORMANCE ANALYSIS
==========
Report Generated: 2025-07-06 23:52:13

OVERALL PERFORMANCE METRICS:
- Average F1 Score: 65.147 (Std: 30.121)
- Average Accuracy: 54.169 (Std: 27.037)
- Average Precision: 68.464 (Std: 33.019)
- Average Recall: 63.863 (Std: 30.084)

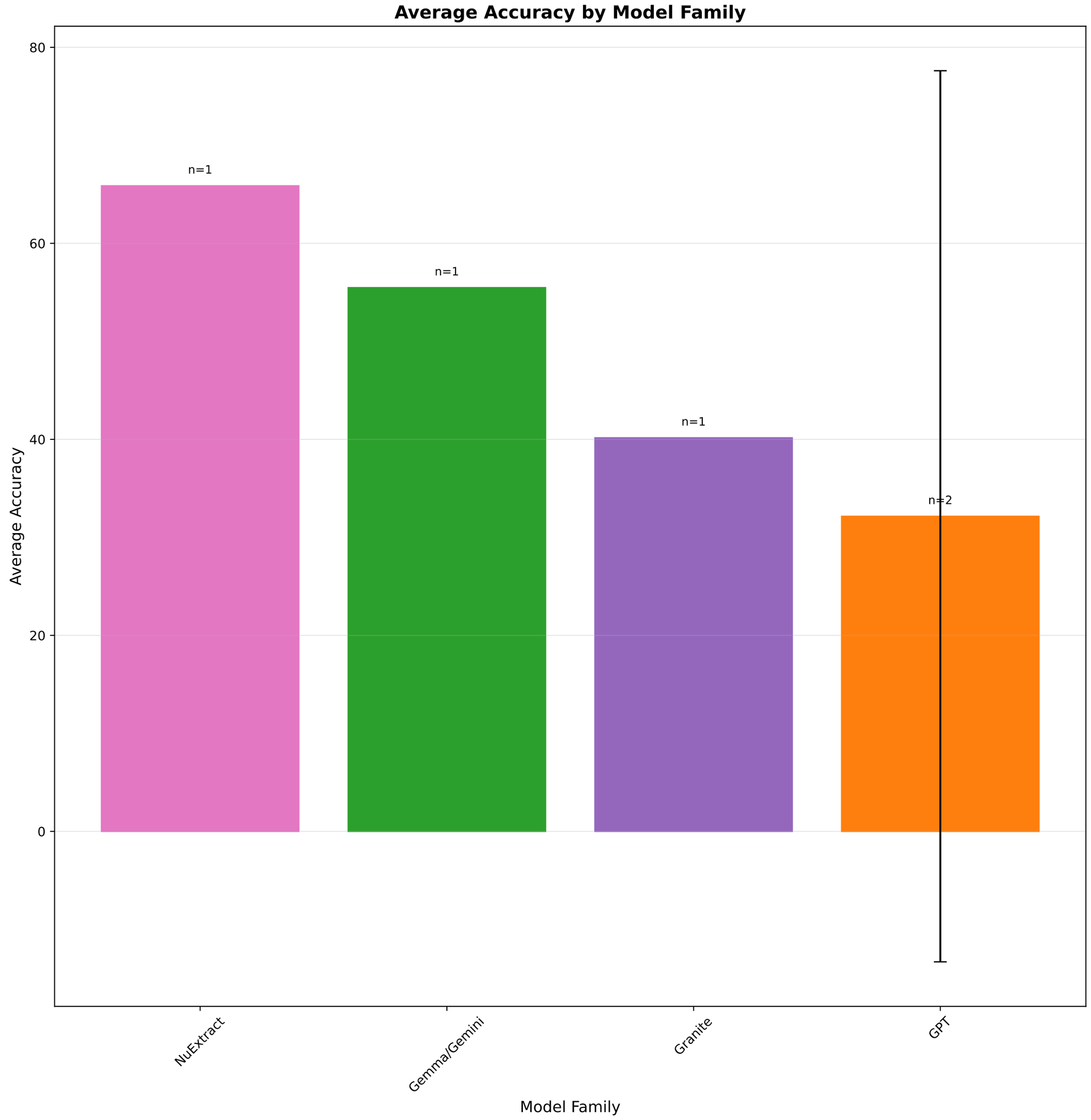
GROUPED MODEL F1 SCORE STATISTICS:
- Unique Base Models: 5
- Total Test Instances: 54
- Best Performing Model: NuExtract:4B (F1: 79.36)
- Worst Performing Model: granite3.2+LTNER/GPT-NER Prompt (F1: 0.00)
- Overall Average F1: 54.15
- Models with Vision: 3


Overall F1score Performance - Models Grouped by Base Name
(Averaged across hospitals and input types)
NuExtract:4B (n=6)
79.4±3.1
gpt-4.1-mini (Image Input, NER Prompt) (n=24)
76.9±14.0
gemma3ne4b (n=12)
70.0±14.9
granite3.2 (Image Input, NER Prompt) (n=8)
44.5±47.6
granite3.2+LTNER/GPT-NER Prompt (Image Input, NER Prompt) (n=4)
0.0±0.0
GPT
Gemma/Gemini
Granite
NuExtract
Given Image Input
Used NER Prompt
0
10
20
30
40
50
60
70
80
Average F1score
Average F1score by Model Family
n=1
n=1
n=1
n=2
80
60
40
20
0
−20
Average F1score
NuExtract
Gemma/Gemini
Granite
GPT
Model Family

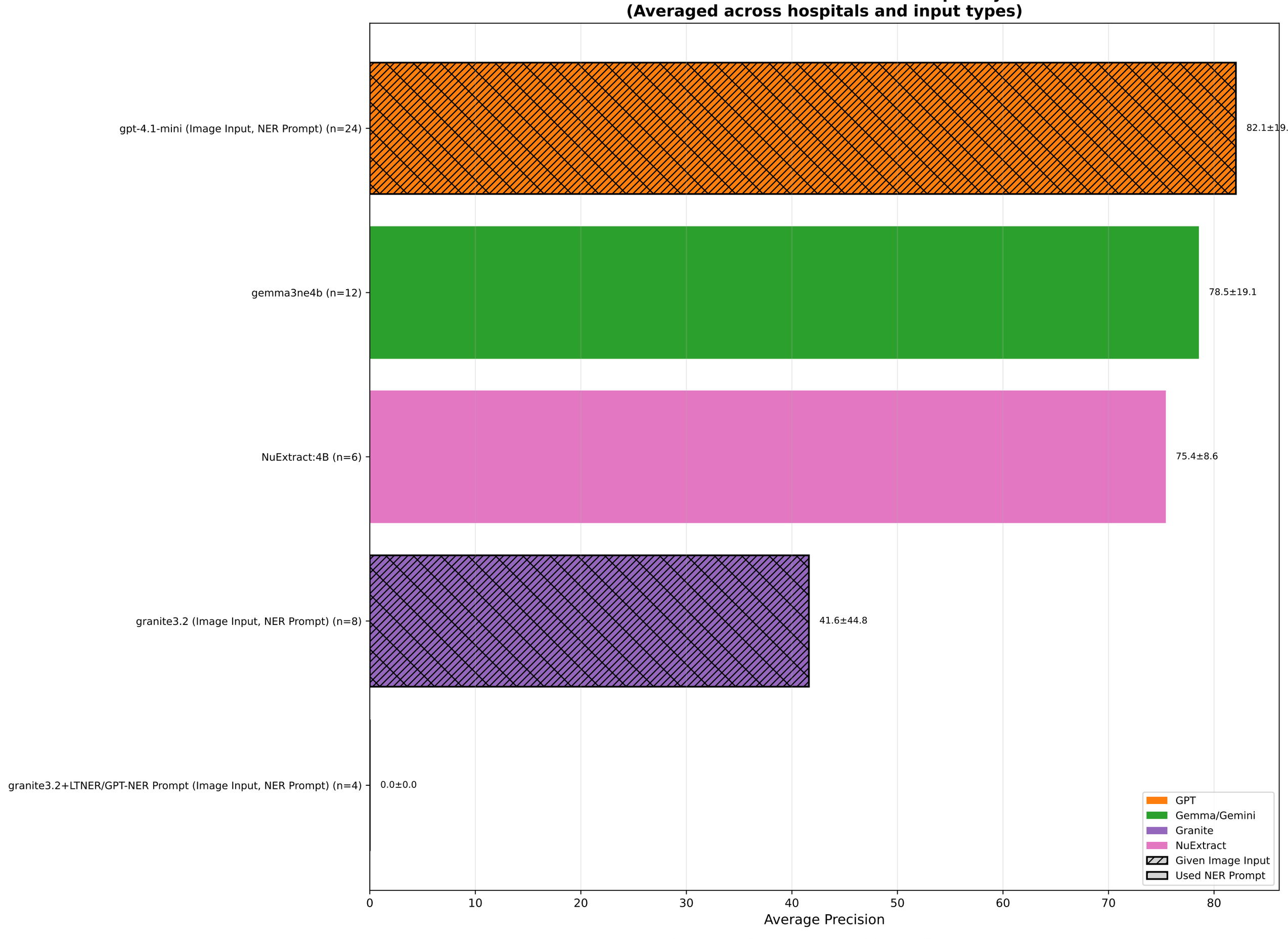
Overall Accuracy Performance - Models Grouped by Base Name
(Averaged across hospitals and input types)



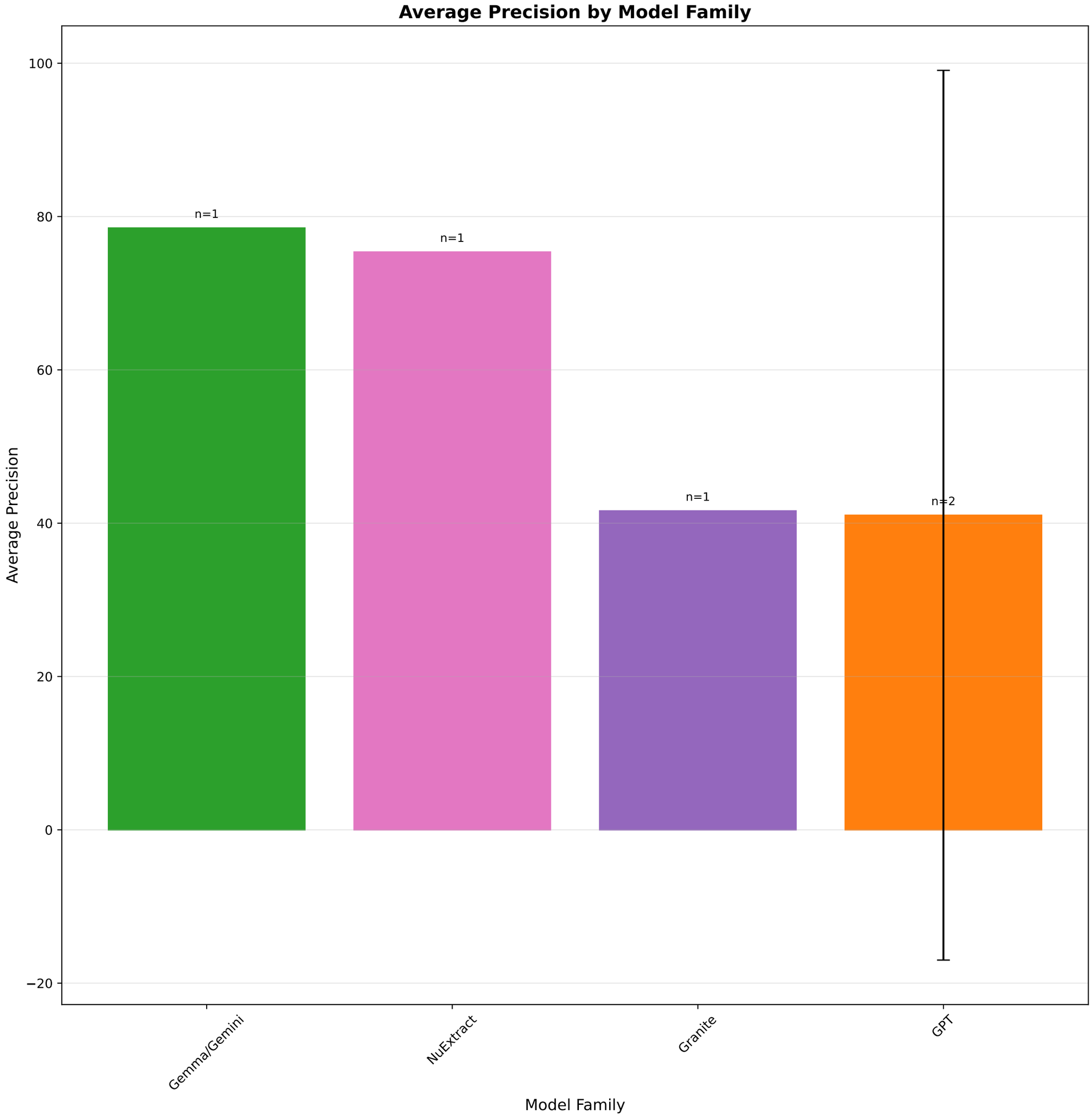
Average Accuracy by Model Family



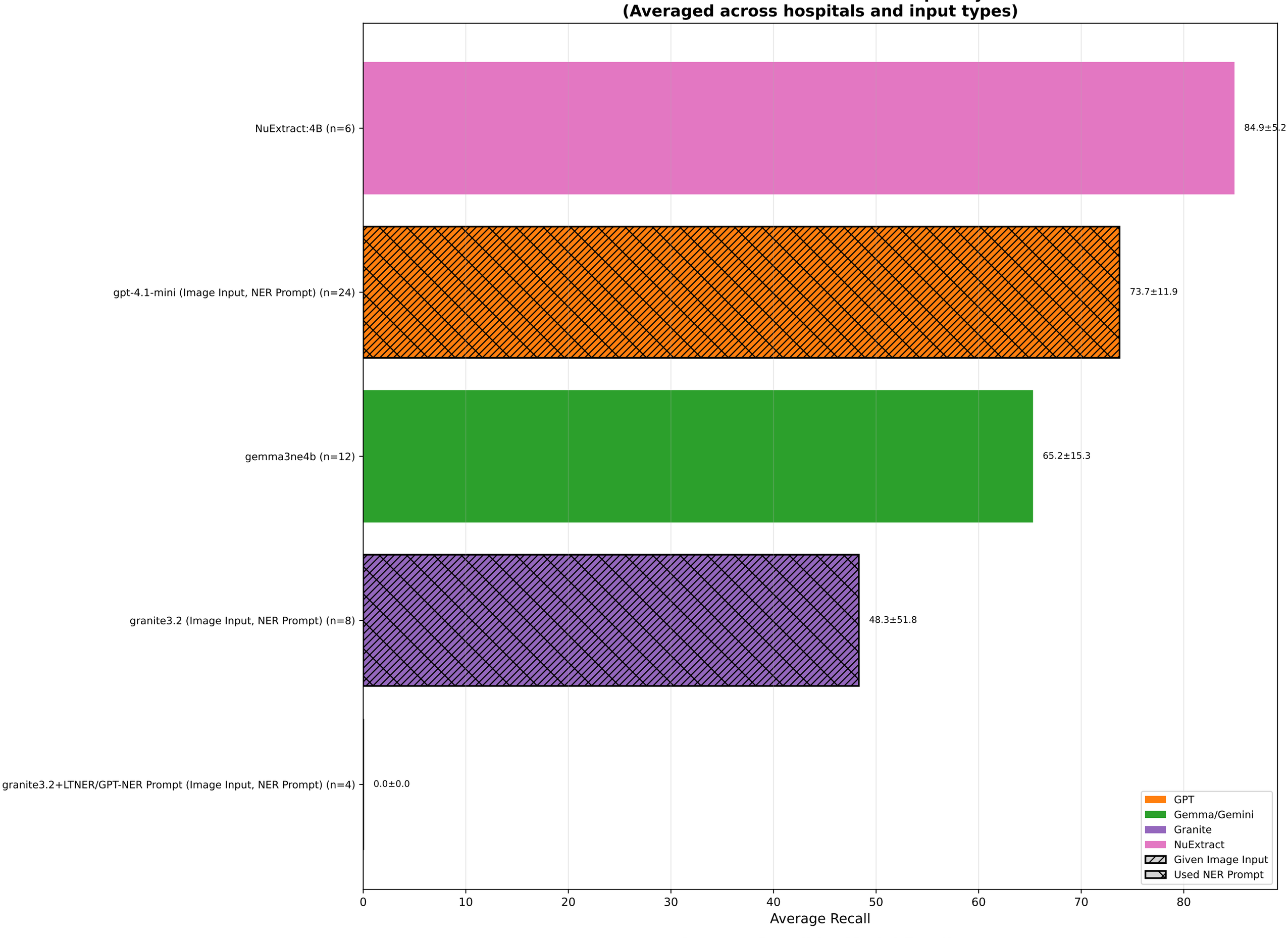
Overall Precision Performance - Models Grouped by Base Name
(Averaged across hospitals and input types)



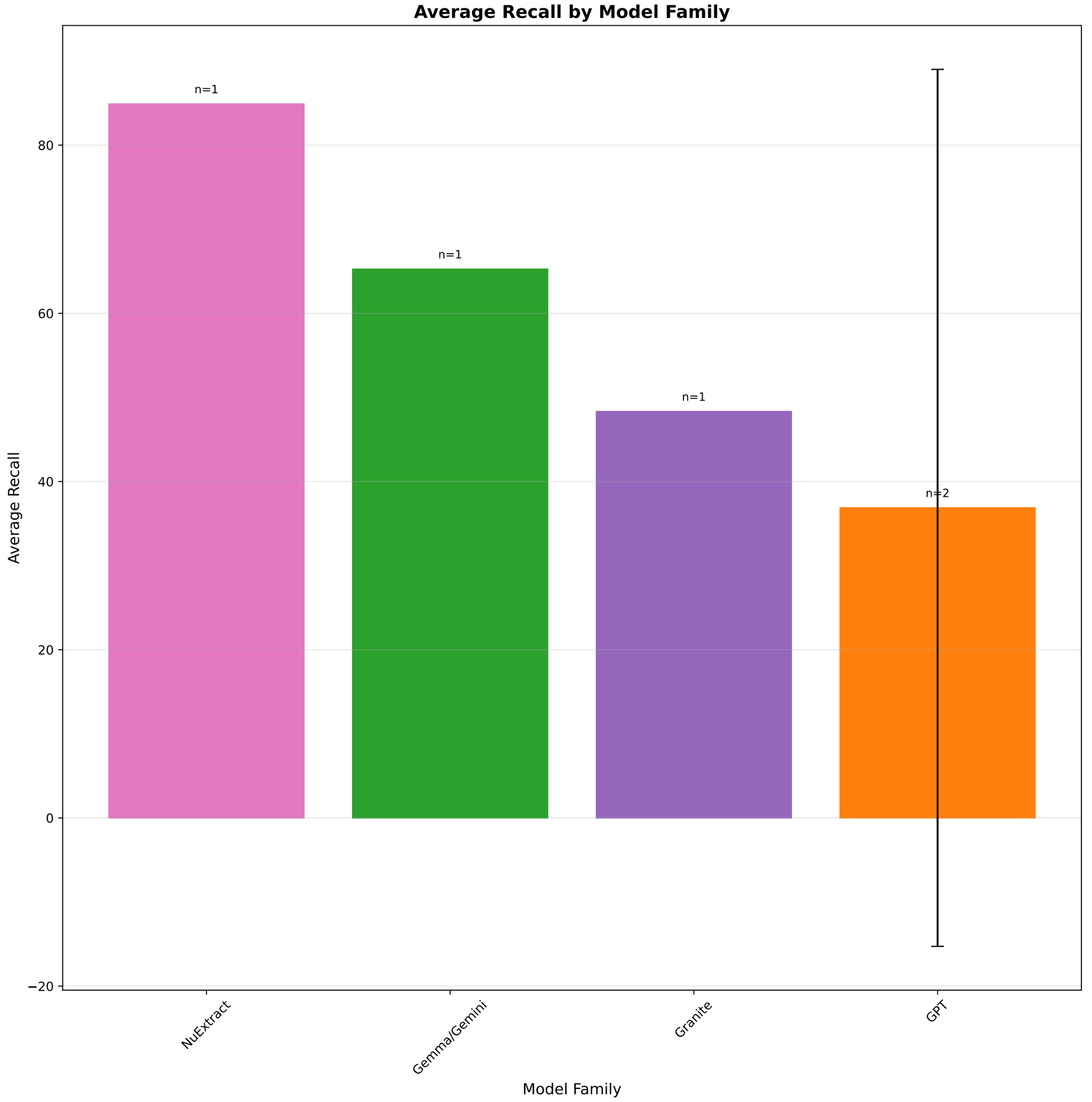
Average Precision by Model Family

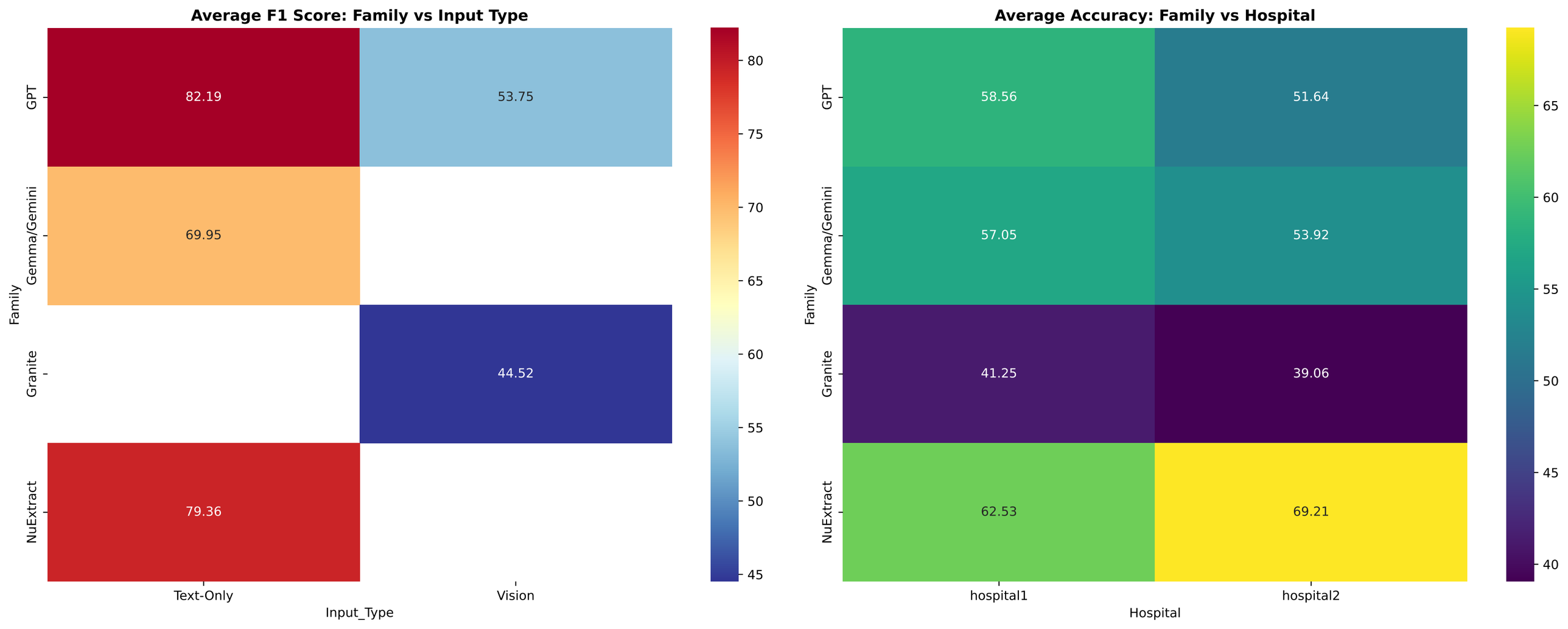


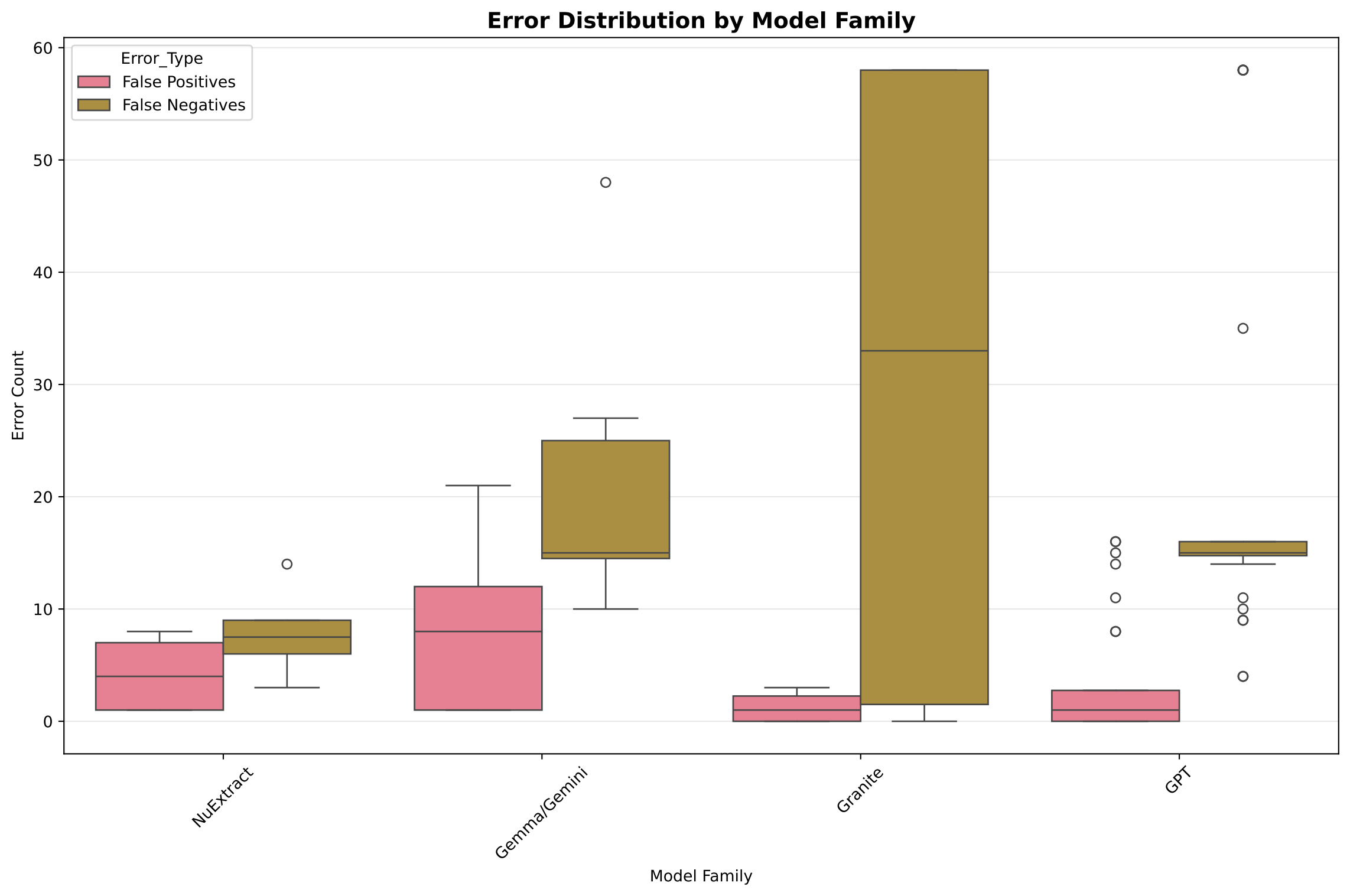
Overall Recall Performance - Models Grouped by Base Name
(Averaged across hospitals and input types)

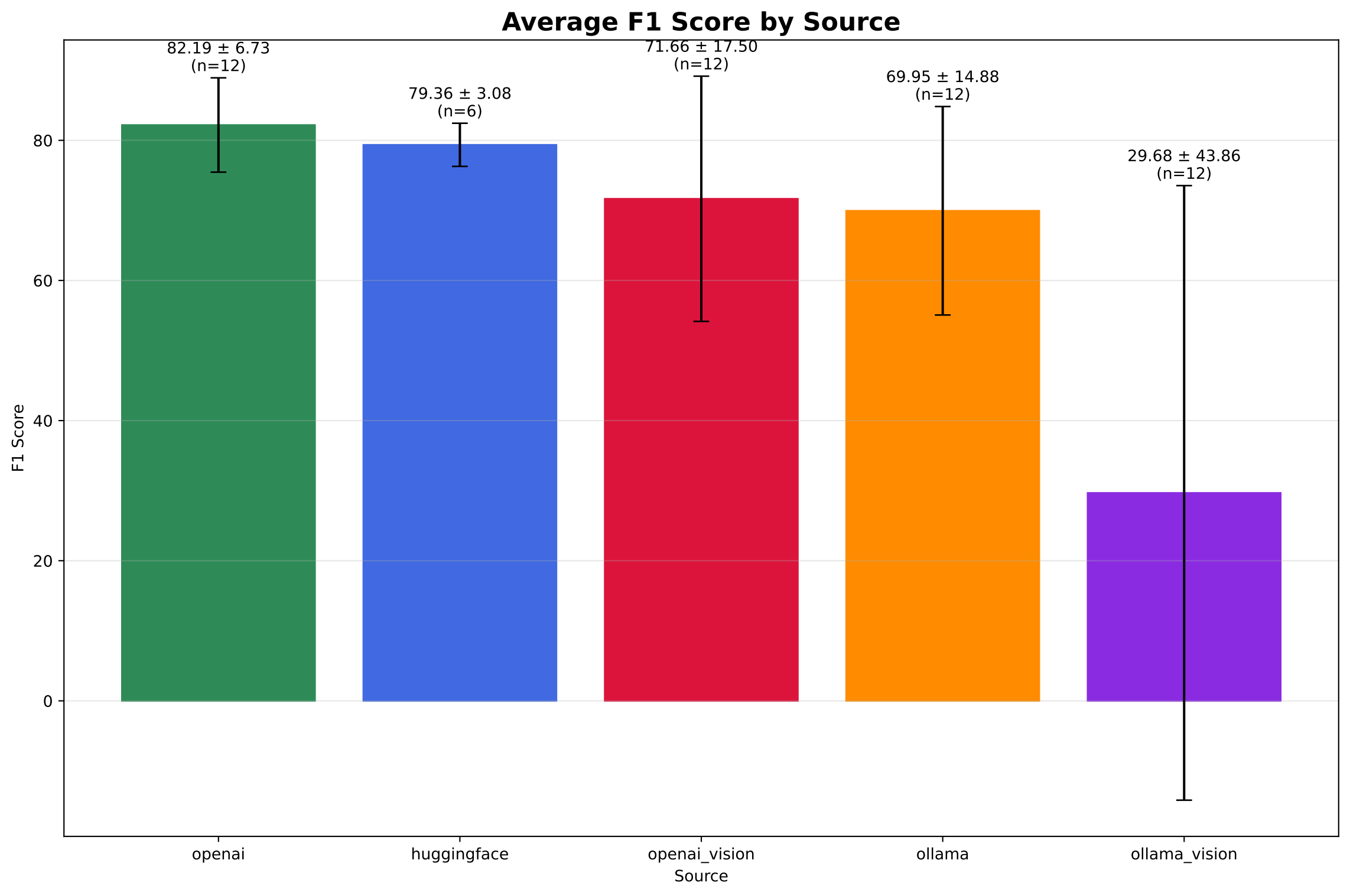


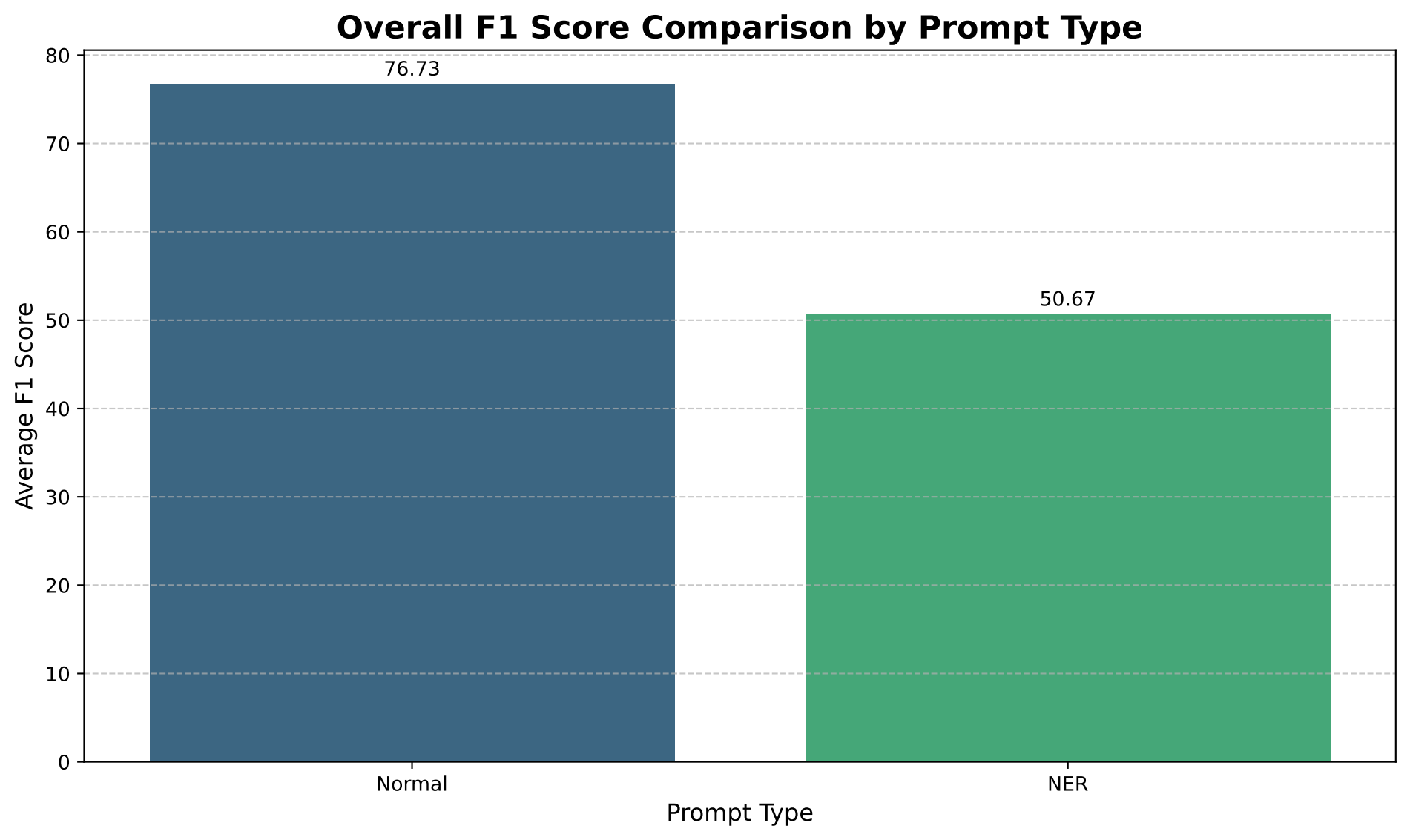
Average Recall by Model Family



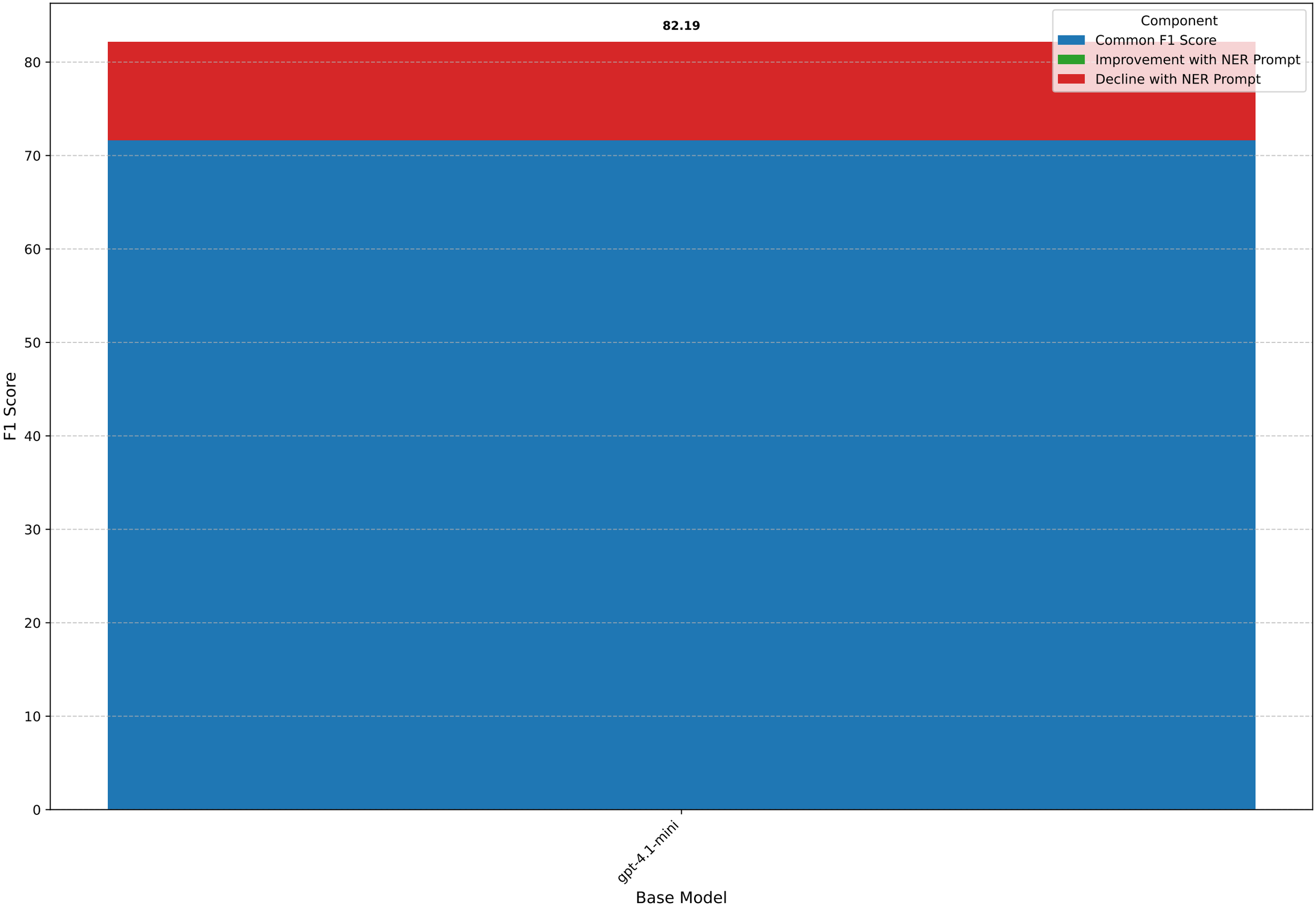

Average F1 Score: Family vs Input Type
Family
GPT
Gemma/Gemini
Granite
NuExtract
82.19
53.75
69.95
44.52
79.36
Text-Only
Vision
Input_Type
80
75
70
65
60
55
50
45
Average Accuracy: Family vs Hospital
Family
GPT
Gemma/Gemini
Granite
NuExtract
58.56
51.64
57.05
53.92
41.25
39.06
62.53
69.21
hospital1
hospital2
Hospital
65
60
55
50
45
40


Error Distribution by Model Family
Error_Type
False Positives
False Negatives
Error Count
0
10
20
30
40
50
60
NuExtract
Gemma/Gemini
Granite
GPT
Model Family


Average F1 Score by Source
82.19 ± 6.73
(n=12)
79.36 ± 3.08
(n=6)
71.66 ± 17.50
(n=12)
69.95 ± 14.88
(n=12)
29.68 ± 43.86
(n=12)
F1 Score
80
60
40
20
0
openai
huggingface
openai_vision
ollama
ollama_vision
Source


Overall F1 Score Comparison by Prompt Type
Average F1 Score
Prompt Type
Normal
NER
76.73
50.67
0
10
20
30
40
50
60
70
80

Impact of NER Prompt on F1 Score



Impact of Image Input on F1 Score