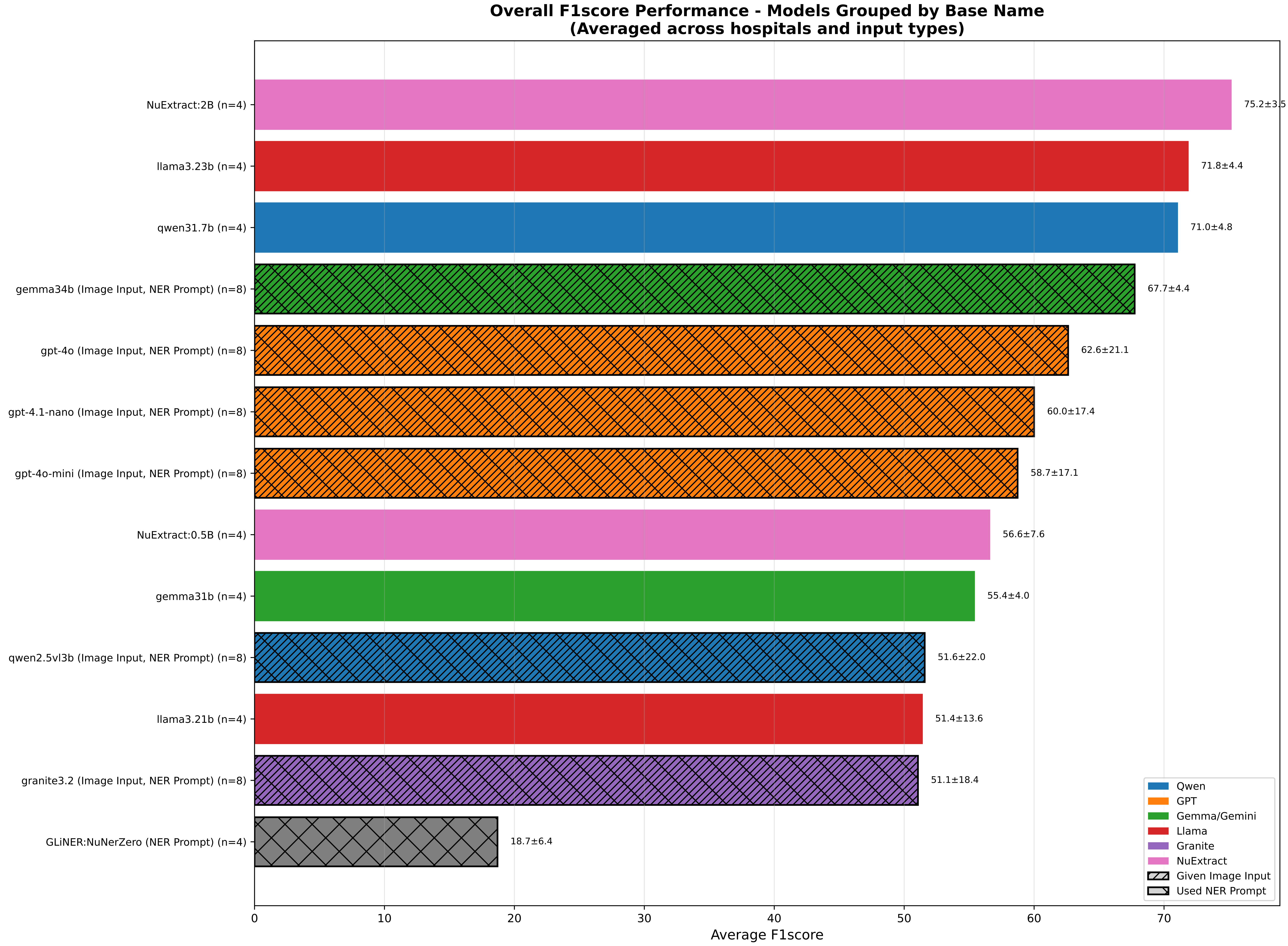
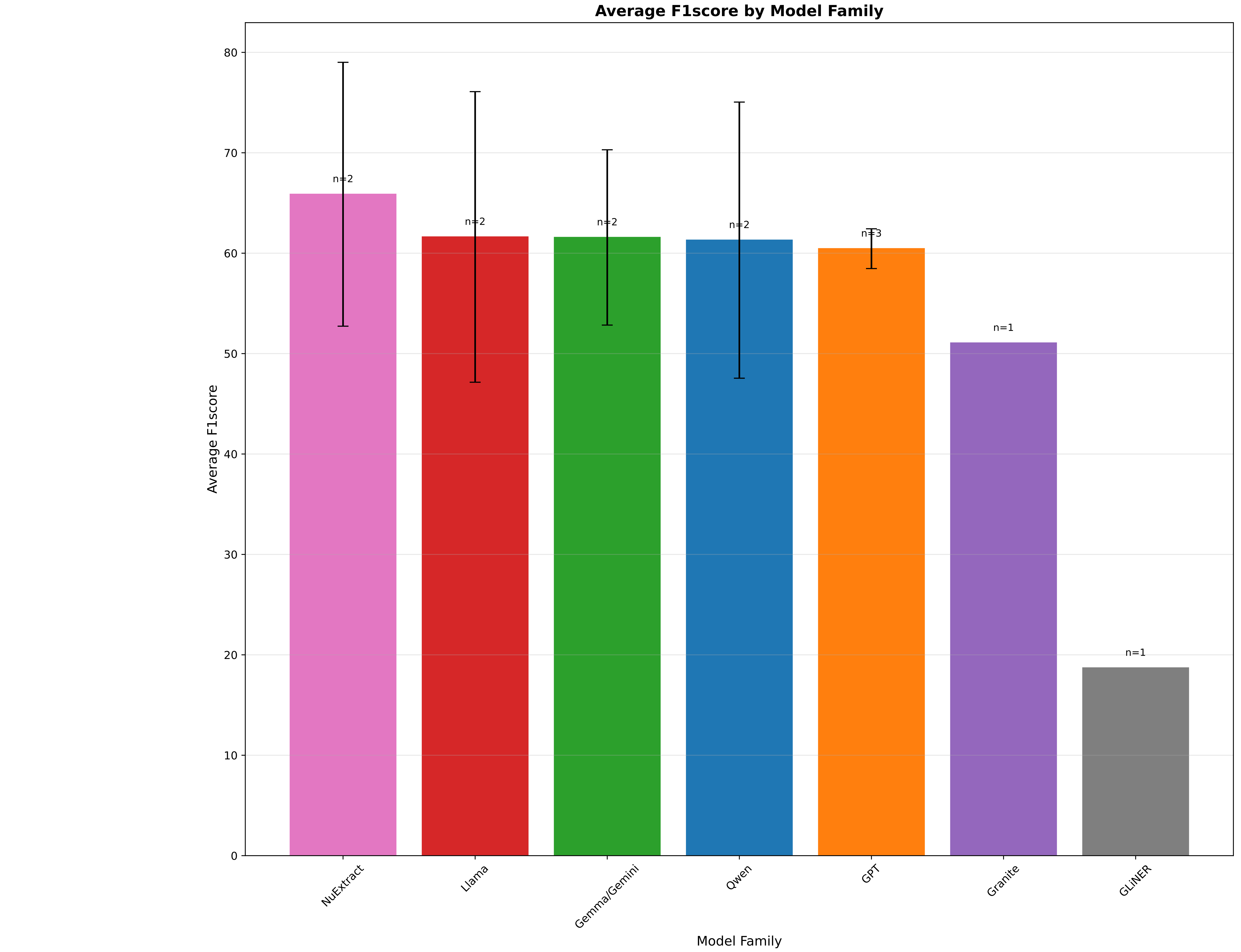
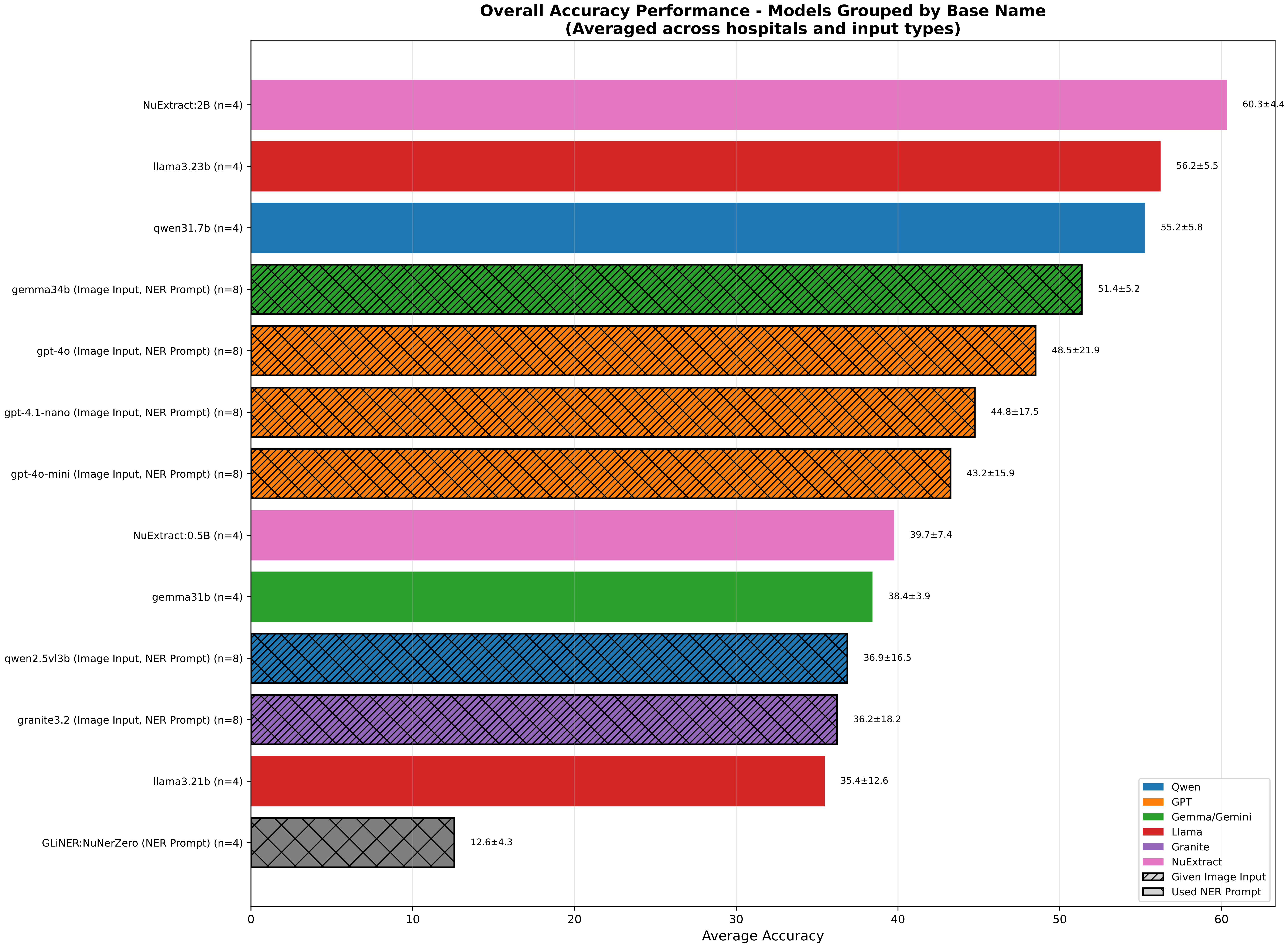
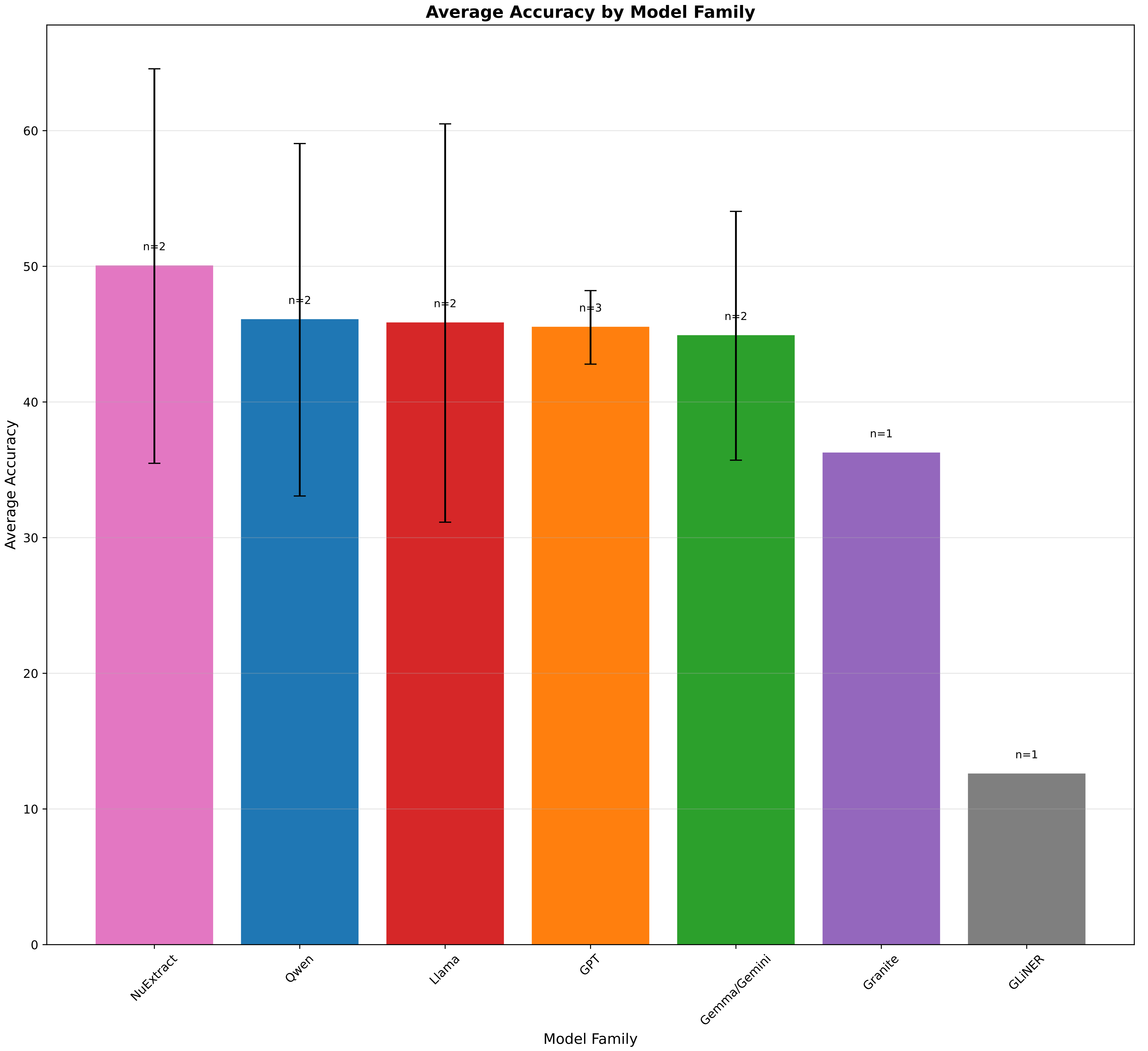
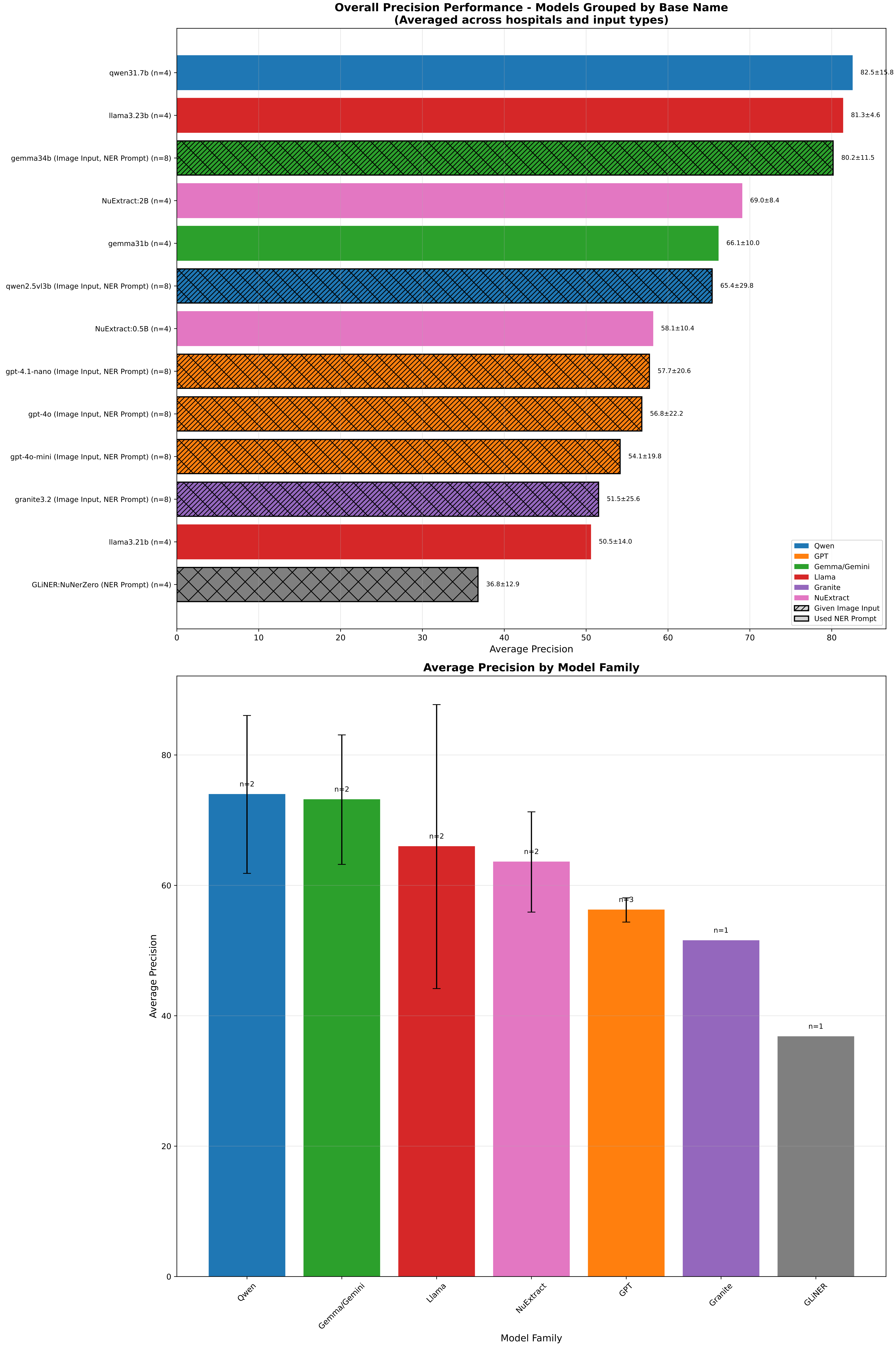
```
COMPREHENSIVE LLM PERFORMANCE ANALYSIS
========================================
Report Generated: 2025-06-26 15:32:30

OVERALL PERFORMANCE METRICS:
   • Average F1 Score: 58.082 (Std: 18.227)
   • Average Accuracy: 43.149 (Std: 16.620)
   • Average Precision: 61.895 (Std: 21.505)
   • Average Recall: 57.942 (Std: 18.549)

GROUPED MODEL F1 SCORE STATISTICS:
   • Unique Base Models: 13
   • Total Test Instances: 76
   • Best Performing Model: NuExtract:2B (F1: 75.16)
   • Worst Performing Model: GLiNER:NuNerZero (F1: 18.70)
   • Overall Average F1: 57.83
   • Models with Vision: 6
```

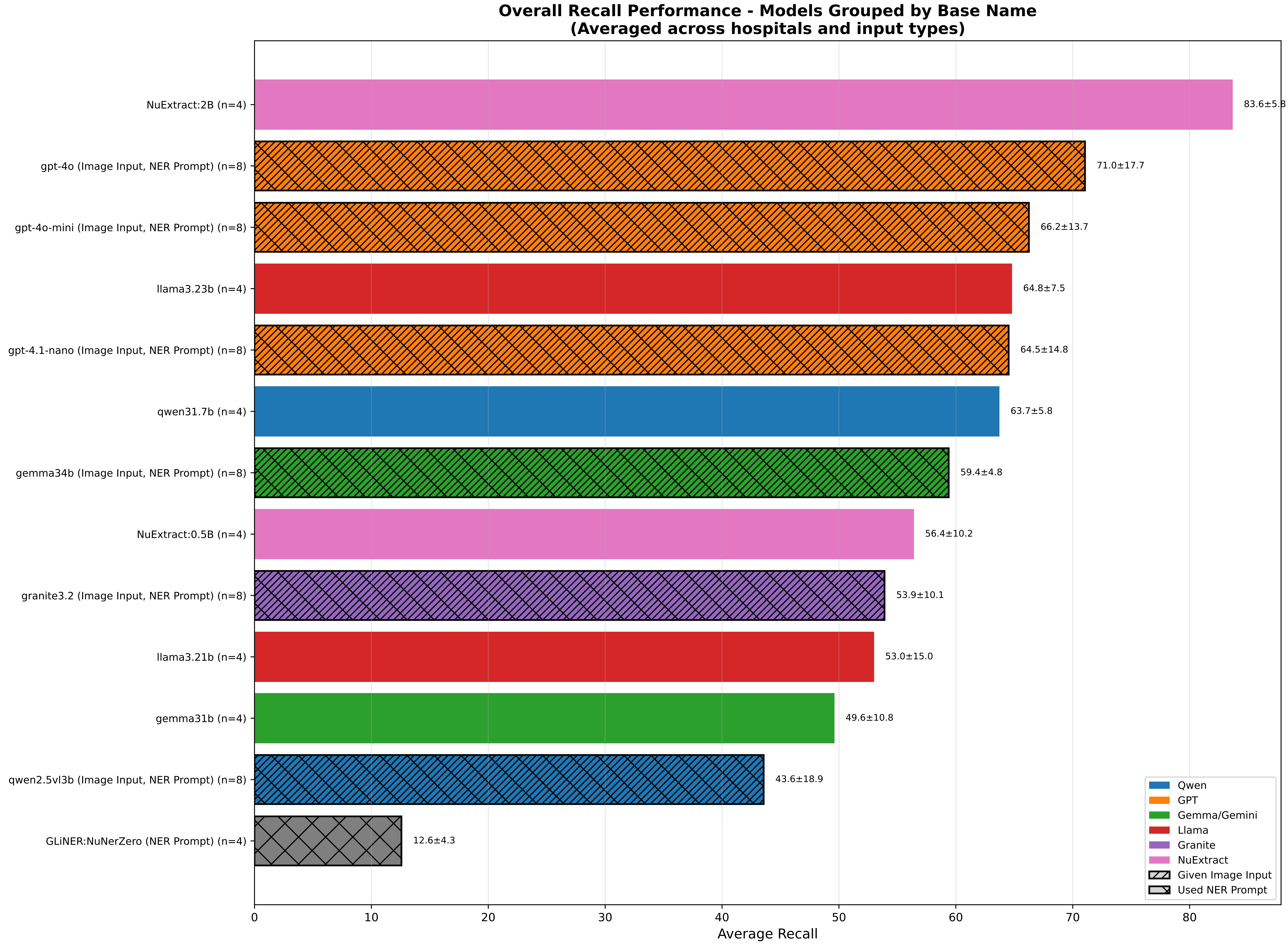

Overall F1score Performance - Models Grouped by Base Name
(Averaged across hospitals and input types)
NuExtract:2B (n=4)
75.2±3.5
llama3.23b (n=4)
71.8±4.4
qwen31.7b (n=4)
71.0±4.8
gemma34b (Image Input, NER Prompt) (n=8)
67.7±4.4
gpt-4o (Image Input, NER Prompt) (n=8)
62.6±21.1
gpt-4.1-nano (Image Input, NER Prompt) (n=8)
60.0±17.4
gpt-4o-mini (Image Input, NER Prompt) (n=8)
58.7±17.1
NuExtract:0.5B (n=4)
56.6±7.6
gemma31b (n=4)
55.4±4.0
qwen2.5vl3b (Image Input, NER Prompt) (n=8)
51.6±22.0
llama3.21b (n=4)
51.4±13.6
granite3.2 (Image Input, NER Prompt) (n=8)
51.1±18.4
GLiNER:NuNerZero (NER Prompt) (n=4)
18.7±6.4
Qwen
GPT
Gemma/Gemini
Llama
Granite
NuExtract
Given Image Input
Used NER Prompt
0
10
20
30
40
50
60
70
Average F1score

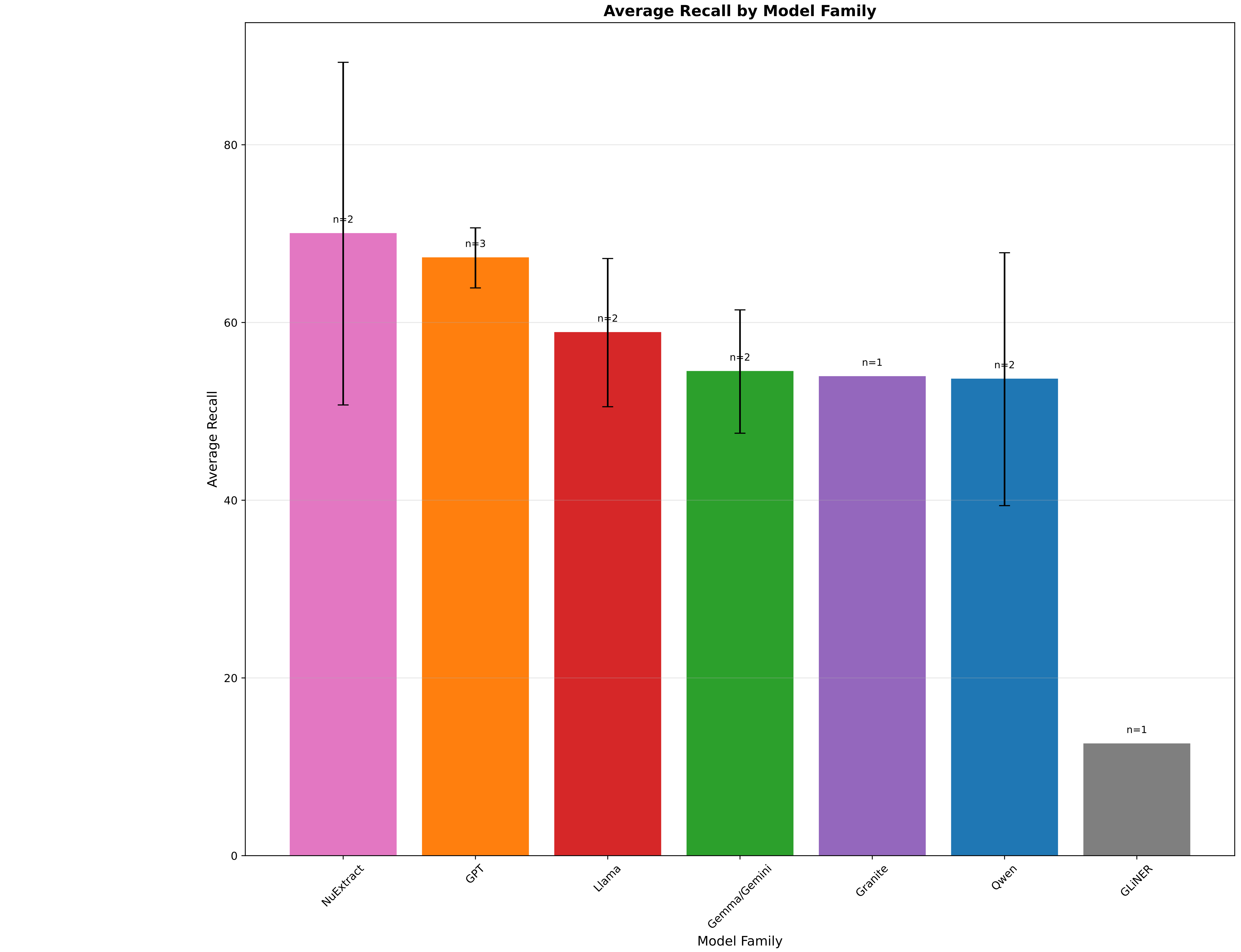

Average F1score by Model Family
Average F1score
0
10
20
30
40
50
60
70
80
n=2
n=2
n=2
n=2
n=3
n=1
n=1
NuExtract
Llama
Gemma/Gemini
Qwen
GPT
Granite
GLiNER
Model Family

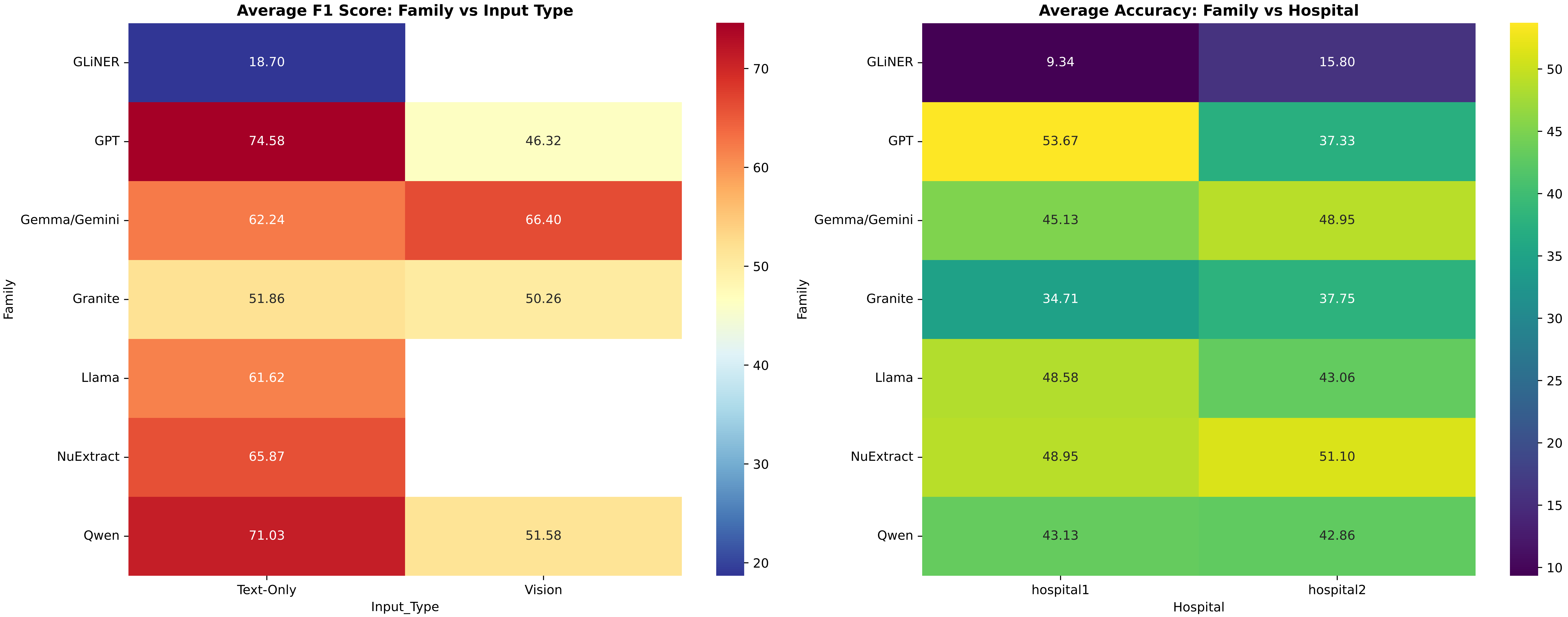

Overall Accuracy Performance - Models Grouped by Base Name
(Averaged across hospitals and input types)


Average Accuracy by Model Family

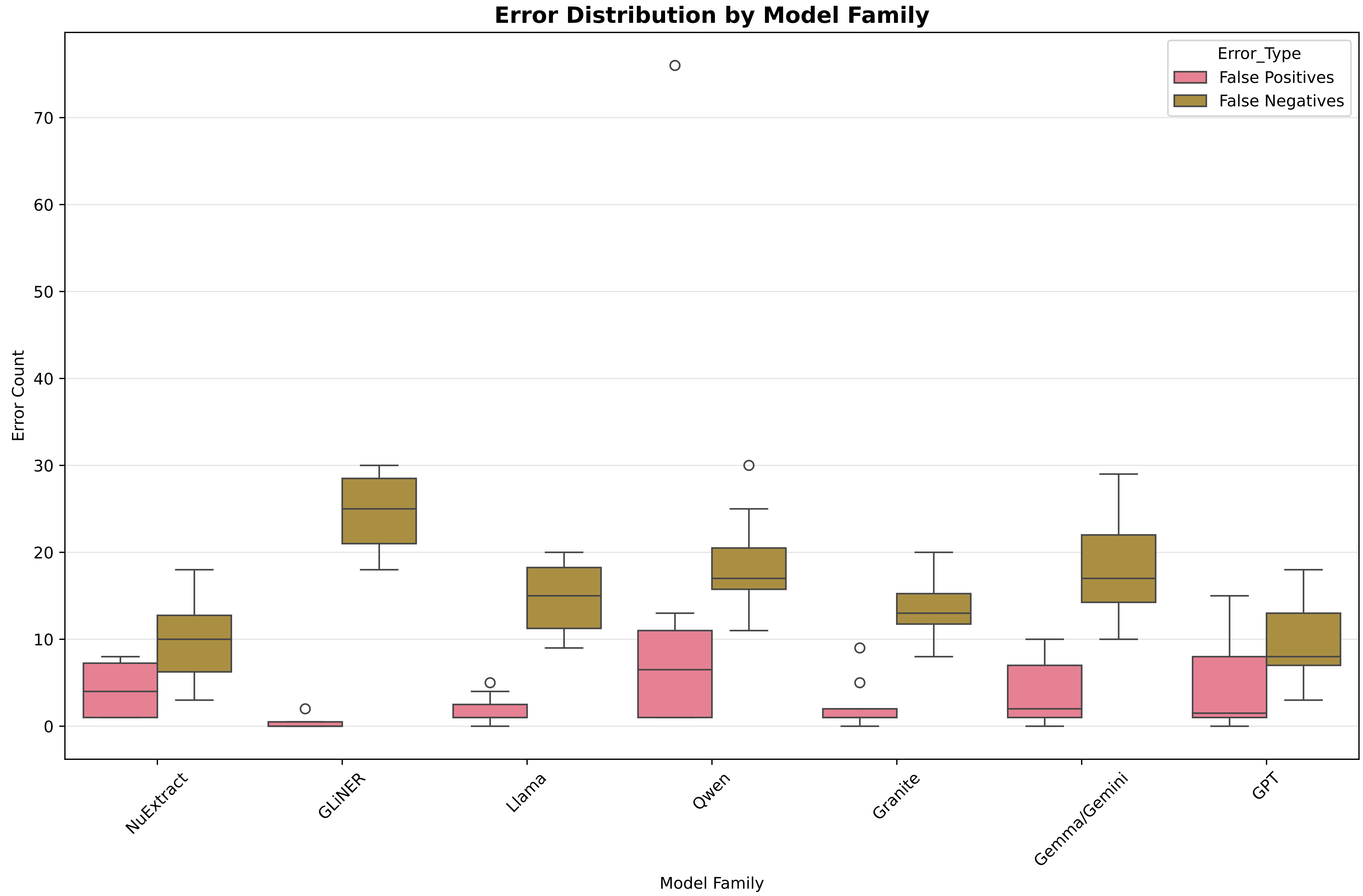

Overall Precision Performance - Models Grouped by Base Name
(Averaged across hospitals and input types)
qwen31.7b (n=4)
82.5±15.8
llama3.23b (n=4)
81.3±4.6
gemma34b (Image Input, NER Prompt) (n=8)
80.2±11.5
NuExtract:2B (n=4)
69.0±8.4
gemma31b (n=4)
66.1±10.0
qwen2.5vl3b (Image Input, NER Prompt) (n=8)
65.4±29.8
NuExtract:0.5B (n=4)
58.1±10.4
gpt-4.1-nano (Image Input, NER Prompt) (n=8)
57.7±20.6
gpt-4o (Image Input, NER Prompt) (n=8)
56.8±22.2
gpt-4o-mini (Image Input, NER Prompt) (n=8)
54.1±19.8
granite3.2 (Image Input, NER Prompt) (n=8)
51.5±25.6
llama3.21b (n=4)
50.5±14.0
GLiNER:NuNerZero (NER Prompt) (n=4)
36.8±12.9
Qwen
GPT
Gemma/Gemini
Llama
Granite
NuExtract
Given Image Input
Used NER Prompt
0
10
20
30
40
50
60
70
80
Average Precision
Average Precision by Model Family
0
20
40
60
80
Average Precision
n=2
n=2
n=2
n=2
n=3
n=1
n=1
Qwen
Gemma/Gemini
Llama
NuExtract
GPT
Granite
GLiNER
Model Family

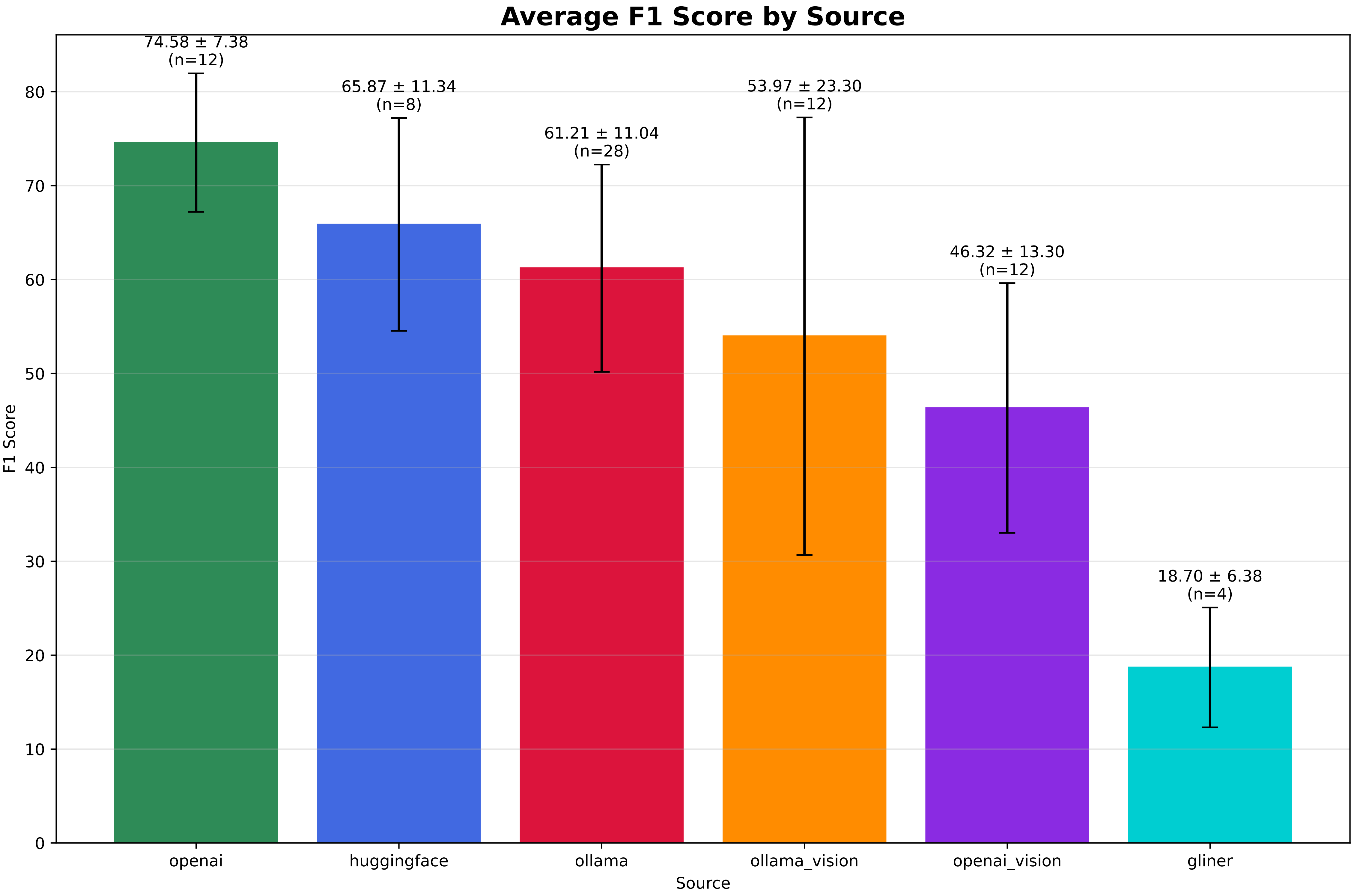

Overall Recall Performance - Models Grouped by Base Name
(Averaged across hospitals and input types)
NuExtract:2B (n=4)
83.6±5.8
gpt-4o (Image Input, NER Prompt) (n=8)
71.0±17.7
gpt-4o-mini (Image Input, NER Prompt) (n=8)
66.2±13.7
llama3.23b (n=4)
64.8±7.5
gpt-4.1-nano (Image Input, NER Prompt) (n=8)
64.5±14.8
qwen31.7b (n=4)
63.7±5.8
gemma34b (Image Input, NER Prompt) (n=8)
59.4±4.8
NuExtract:0.5B (n=4)
56.4±10.2
granite3.2 (Image Input, NER Prompt) (n=8)
53.9±10.1
llama3.21b (n=4)
53.0±15.0
gemma31b (n=4)
49.6±10.8
qwen2.5vl3b (Image Input, NER Prompt) (n=8)
43.6±18.9
GLiNER:NuNerZero (NER Prompt) (n=4)
12.6±4.3
Qwen
GPT
Gemma/Gemini
Llama
Granite
NuExtract
Given Image Input
Used NER Prompt
0
10
20
30
40
50
60
70
80
Average Recall

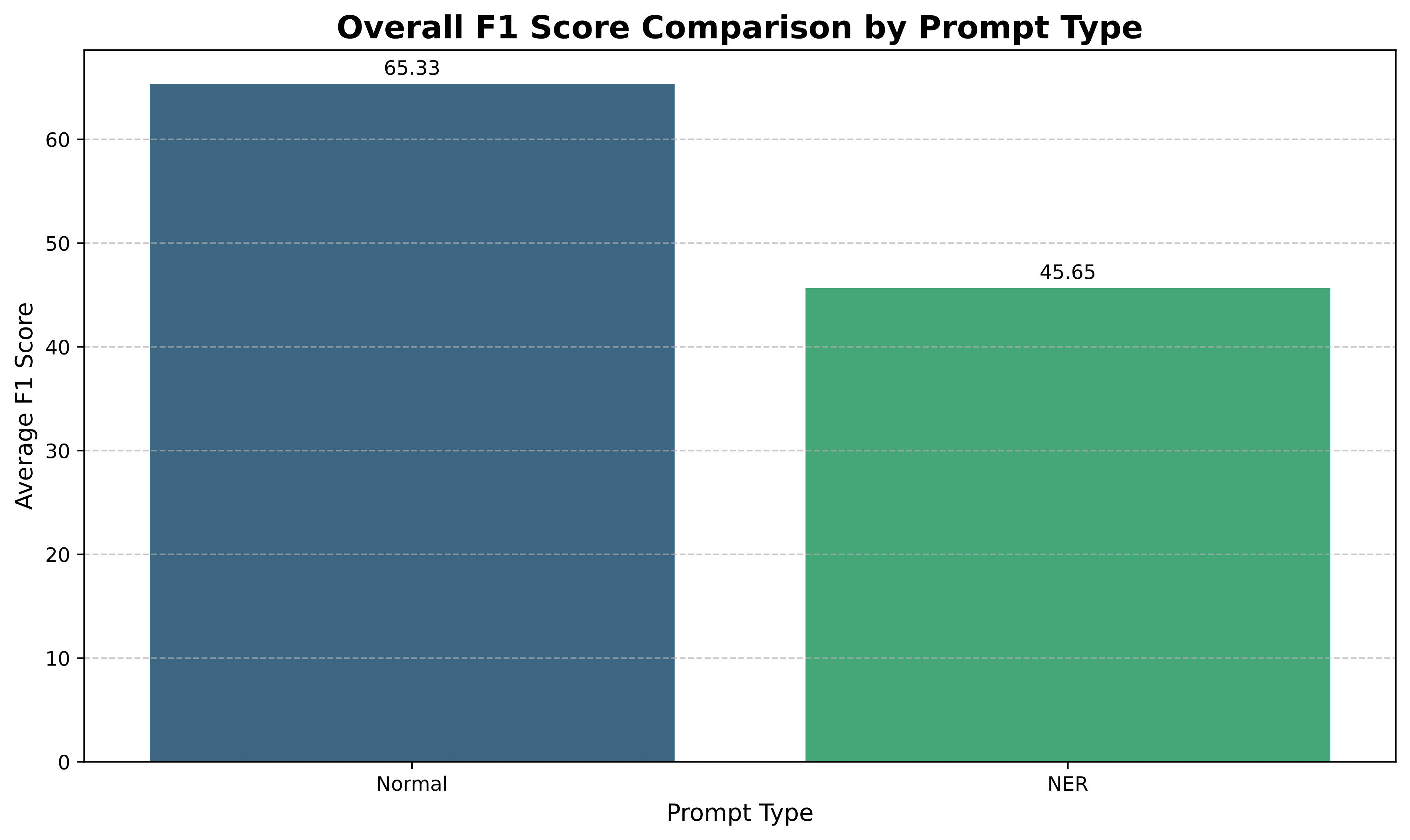

Average Recall by Model Family
Average Recall
0
20
40
60
80
n=2
n=3
n=2
n=2
n=1
n=2
n=1
NuExtract
GPT
Llama
Gemma/Gemini
Granite
Qwen
GLiNER
Model Family

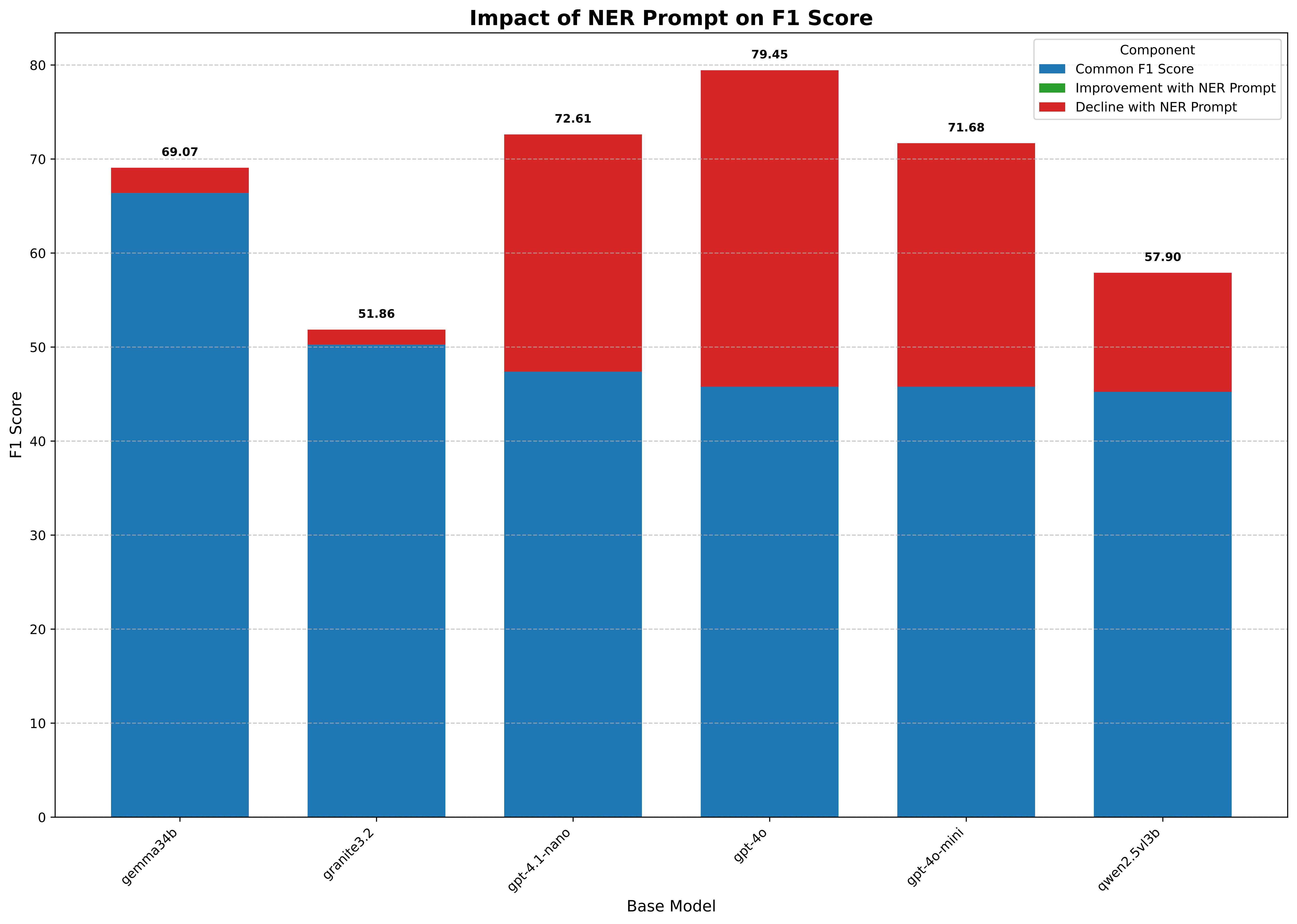


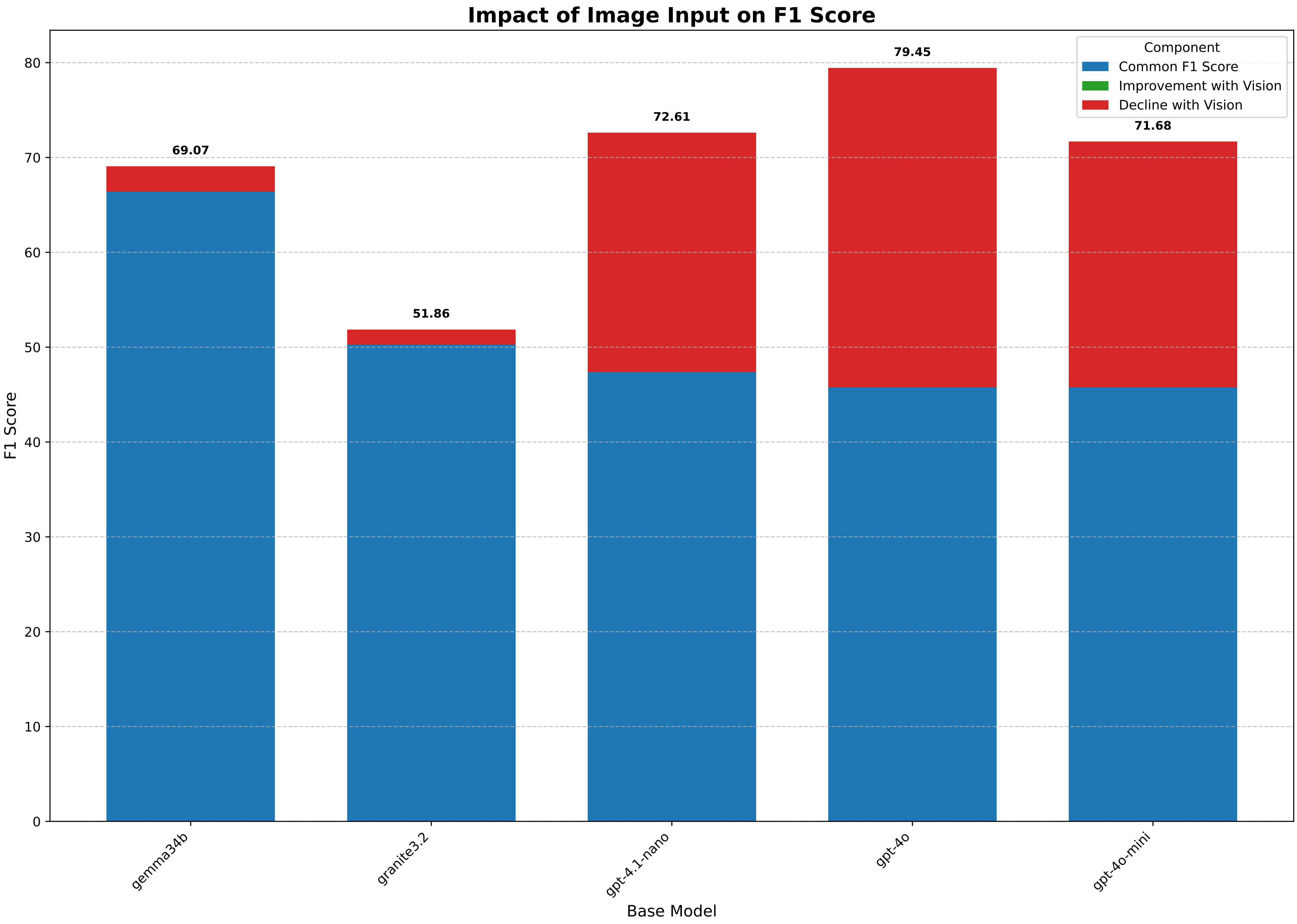
Average F1 Score: Family vs Input Type
Family
GLiNER
GPT
Gemma/Gemini
Granite
Llama
NuExtract
Qwen
Text-Only
Vision
Input_Type
18.70
74.58
46.32
62.24
66.40
51.86
50.26
61.62
65.87
71.03
51.58
20
30
40
50
60
70
Average Accuracy: Family vs Hospital
Family
GLiNER
GPT
Gemma/Gemini
Granite
Llama
NuExtract
Qwen
hospital1
hospital2
Hospital
9.34
15.80
53.67
37.33
45.13
48.95
34.71
37.75
48.58
43.06
48.95
51.10
43.13
42.86
10
15
20
25
30
35
40
45
50


Error Distribution by Model Family
Error_Type
False Positives
False Negatives
Error Count
0
10
20
30
40
50
60
70
Model Family
NuExtract
GLiNER
Llama
Qwen
Granite
Gemma/Gemini
GPT


Average F1 Score by Source
74.58 ± 7.38
(n=12)
65.87 ± 11.34
(n=8)
61.21 ± 11.04
(n=28)
53.97 ± 23.30
(n=12)
46.32 ± 13.30
(n=12)
18.70 ± 6.38
(n=4)
F1 Score
0
10
20
30
40
50
60
70
80
openai
huggingface
ollama
ollama_vision
openai_vision
gliner
Source


Overall F1 Score Comparison by Prompt Type
65.33
45.65
Average F1 Score
0
10
20
30
40
50
60
Normal
NER
Prompt Type


Impact of NER Prompt on F1 Score
Component
Common F1 Score
Improvement with NER Prompt
Decline with NER Prompt
F1 Score
0
10
20
30
40
50
60
70
80
69.07
51.86
72.61
79.45
71.68
57.90
gemma34b
granite3.2
gpt-4.1-nano
gpt-4o
gpt-4o-mini
qwen2.5vl3b
Base Model


Impact of Image Input on F1 Score
F1 Score
0
10
20
30
40
50
60
70
80
Base Model
gemma34b
granite3.2
gpt-4.1-nano
gpt-4o
gpt-4o-mini
69.07
51.86
72.61
79.45
71.68
Component
Common F1 Score
Improvement with Vision
Decline with Vision