COMPREHENSIVE LLM BENCHMARKING ANALYSIS - SUMMARY REPORT
==

DATASET OVERVIEW:
- Total Records: 114
- Model Families: NuExtract, GLiNER, Llama, Qwen, Granite, Gemma/Gemini, GPT
- Hospitals: hospital1, hospital2
- Vision-enabled Models: 42
- Text-only Models: 72

TOP PERFORMERS:
- qwen2.5vl3b (Qwen, 3.0B, Vision)
  F1: 89.08, Accuracy: 80.30
- llama3.21b (Llama, 3.21B, Text-only)
  F1: 88.14, Accuracy: 78.79
- granite3.2*ImageInput* (Granite, Unknown, Vision)
  F1: 87.18, Accuracy: 77.27
- gpt-4o (GPT, 200.0B, Text-only)
  F1: 87.18, Accuracy: 77.27
- qwen31.7b (Qwen, 31.7B, Text-only)
  F1: 86.21, Accuracy: 75.76

FAMILY PERFORMANCE RANKING (by F1 Score):
- Qwen: 65.90 (n=18.0)
- Llama: 65.26 (n=12.0)
- Gemma/Gemini: 59.70 (n=18.0)
- GPT: 58.97 (n=36.0)
- Granite: 58.94 (n=12.0)
- NuExtract: 55.72 (n=12.0)
- GLiNER: 28.12 (n=6.0)

KEY INSIGHTS:
- Vision Models Avg F1: 56.96
- Text-only Models Avg F1: 59.99
- Hospital 1 Avg F1: 62.01
- Hospital 2 Avg F1: 55.74

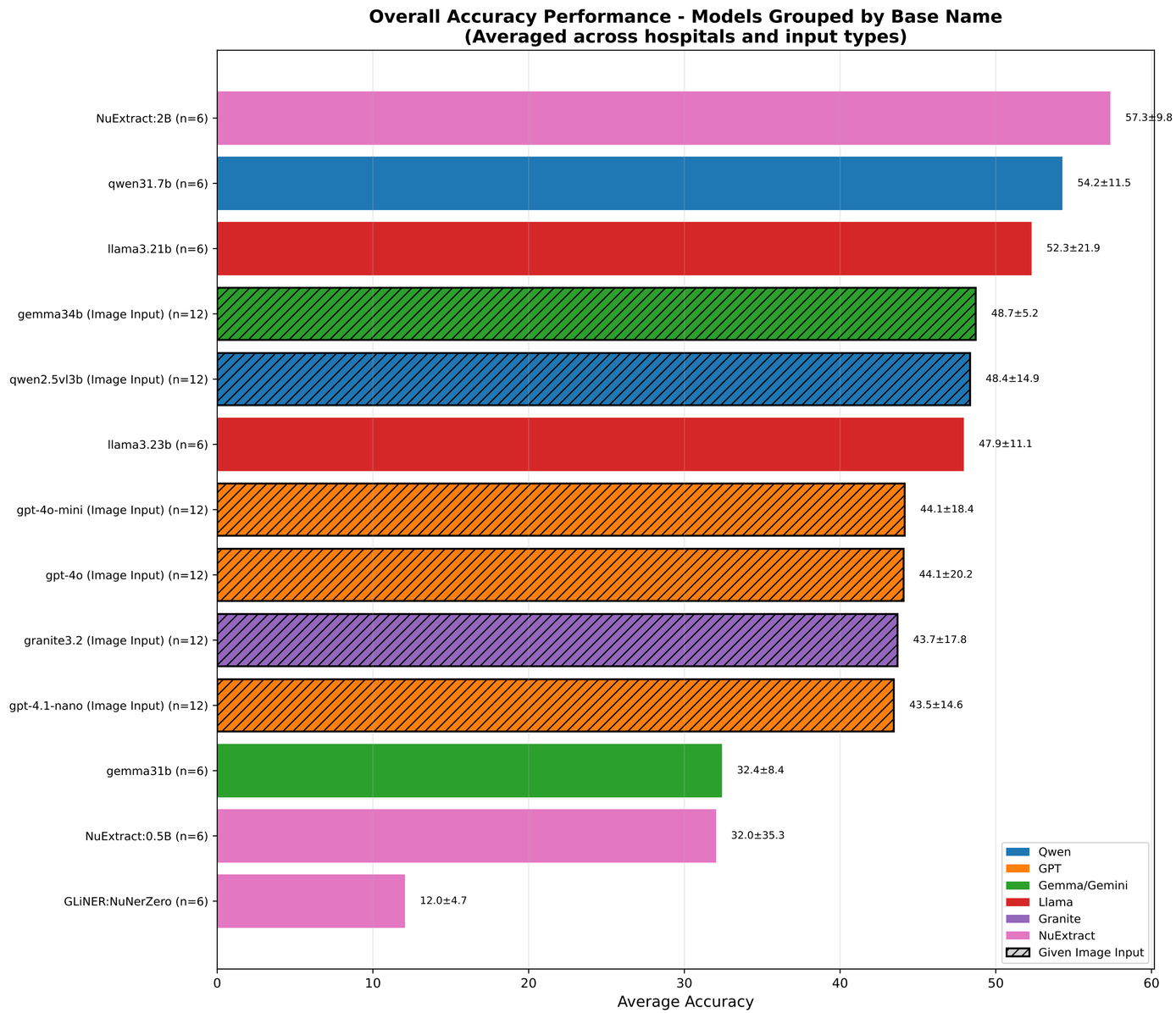
Grouped Model F1 Score Statistics:
- Unique Base Models: 13
- Total Test Instances: 114
- Best Performing Model: NuExtract:2B (F1: 72.48)
- Worst Performing Model: GLiNER:NuNerZero (F1: 28.12)
- Overall Average F1: 57.96
- Models with Vision: 6

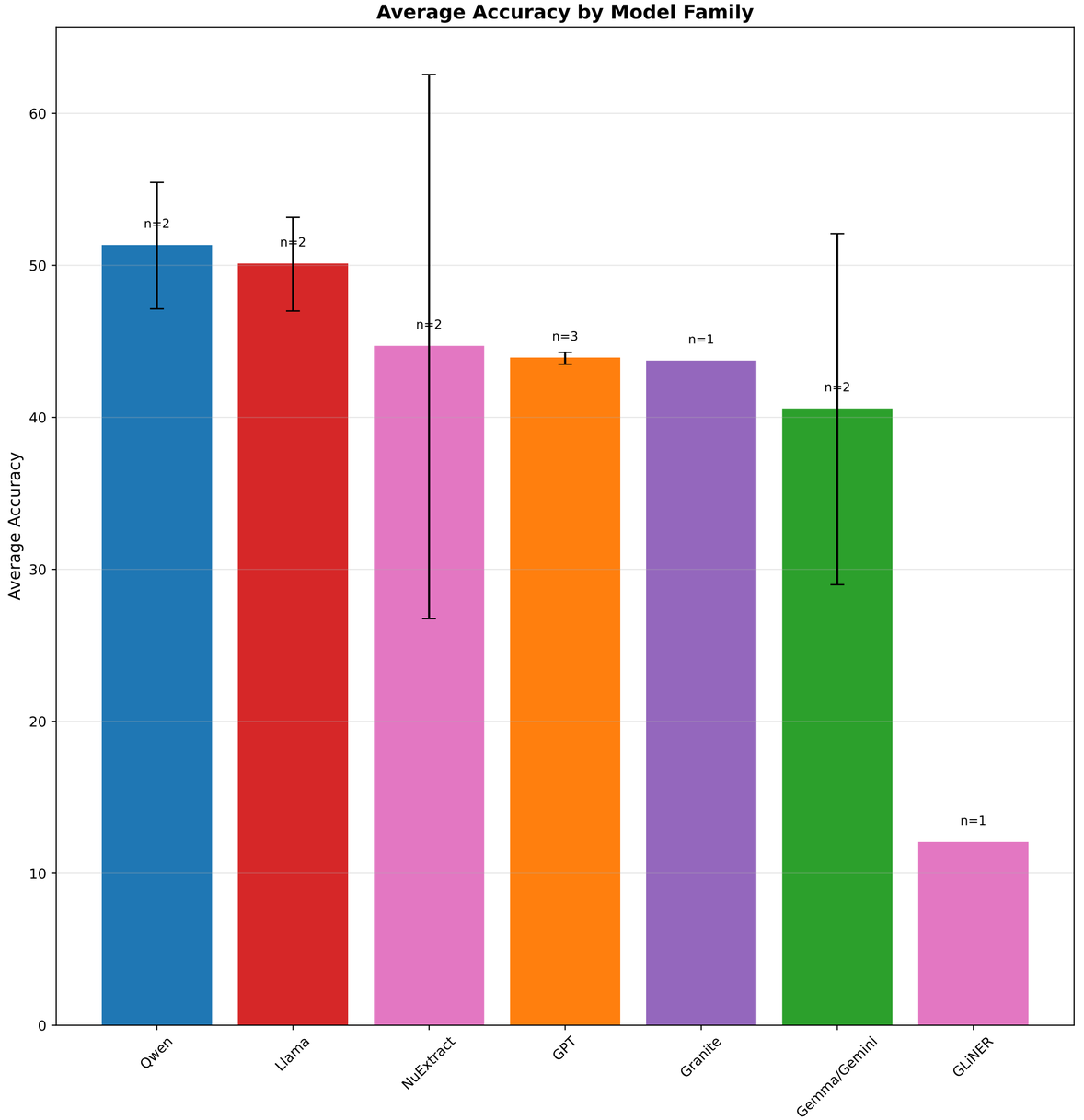
Top 5 Performers:
- llama3.23b (Llama, Text-only): F1 = 64.16 ± 9.91
- gemma34b (Gemma/Gemini, with Vision): F1 = 65.36 ± 4.69
- llama3.21b (Llama, Text-only): F1 = 66.35 ± 19.18
- qwen31.7b (Qwen, Text-only): F1 = 69.77 ± 9.07
- NuExtract:2B (NuExtract, Text-only): F1 = 72.48 ± 7.69

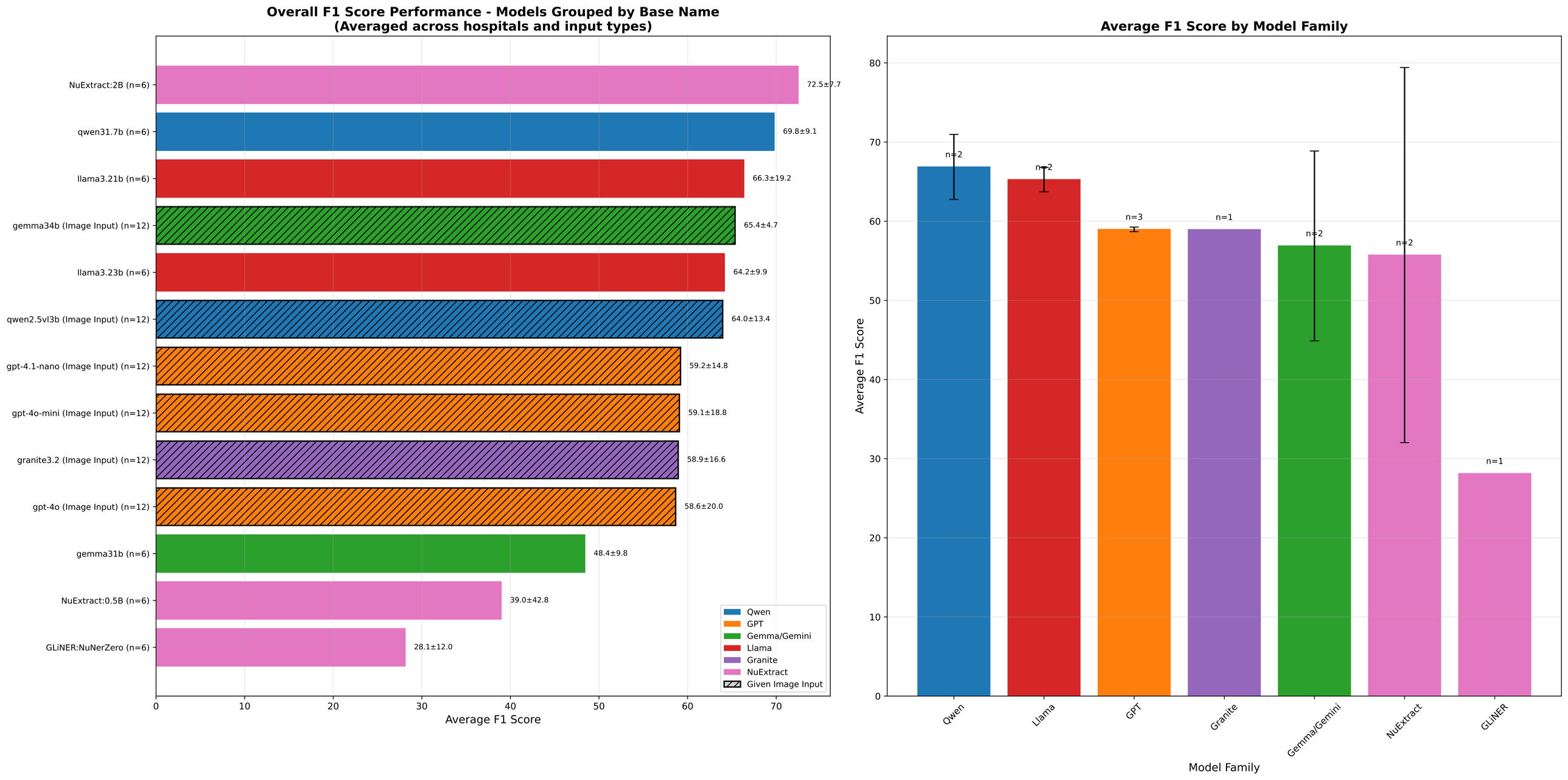
Bottom 5 Performers:
- GLiNER:NuNerZero (GLiNER, Text-only): F1 = 28.12 ± 11.97
- NuExtract:0.5B (NuExtract, Text-only): F1 = 38.96 ± 42.78
- gemma31b (Gemma/Gemini, Text-only): F1 = 48.40 ± 9.80
- gpt-4o (GPT, with Vision): F1 = 58.65 ± 19.97
- granite3.2 (Granite, with Vision): F1 = 58.94 ± 16.63

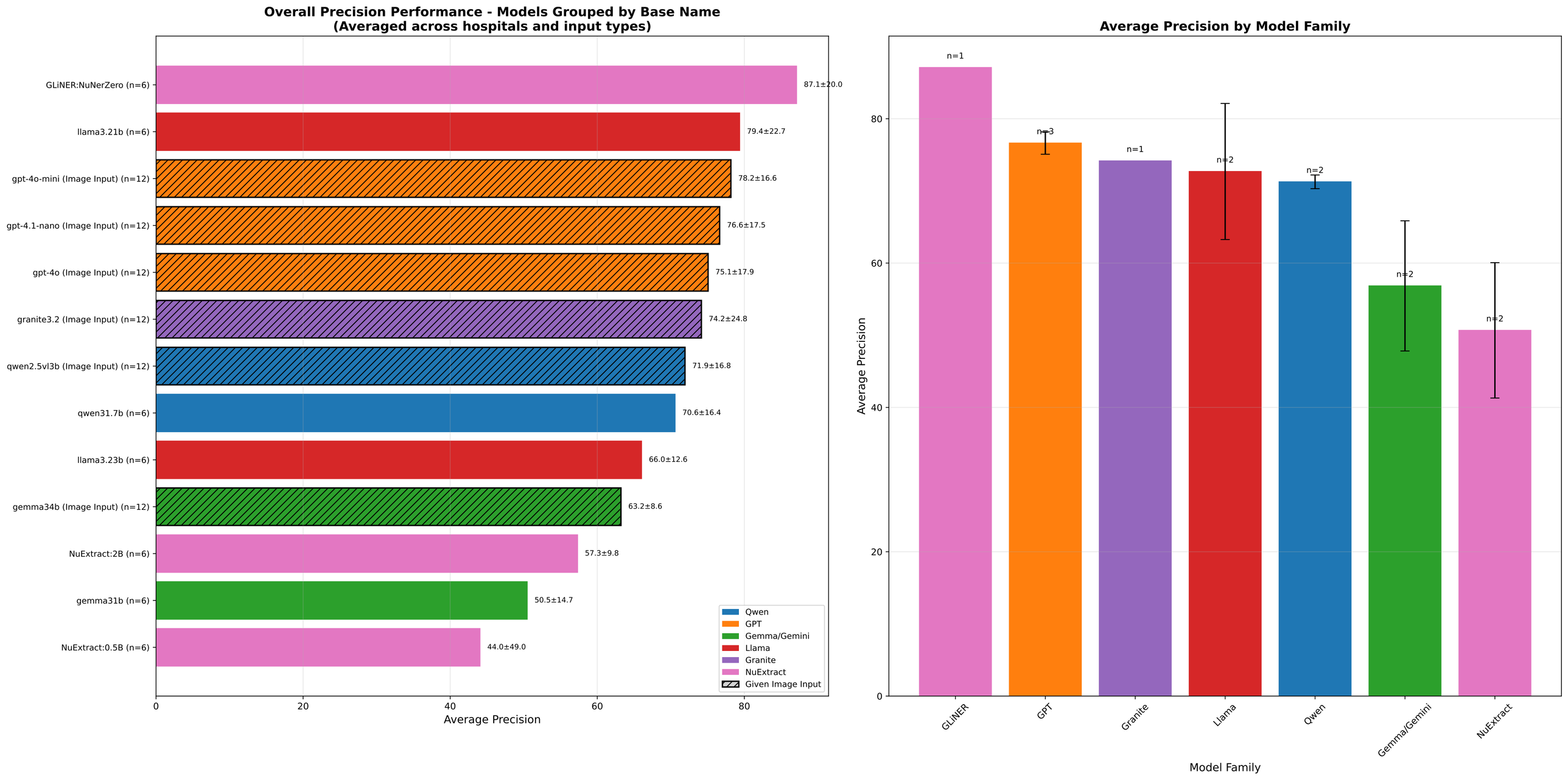
Error Analysis Summary:
- Average False Positives: 4.23
- Average False Negatives: 17.66
- Models with more FP than FN: 32
- Models with more FN than FP: 81

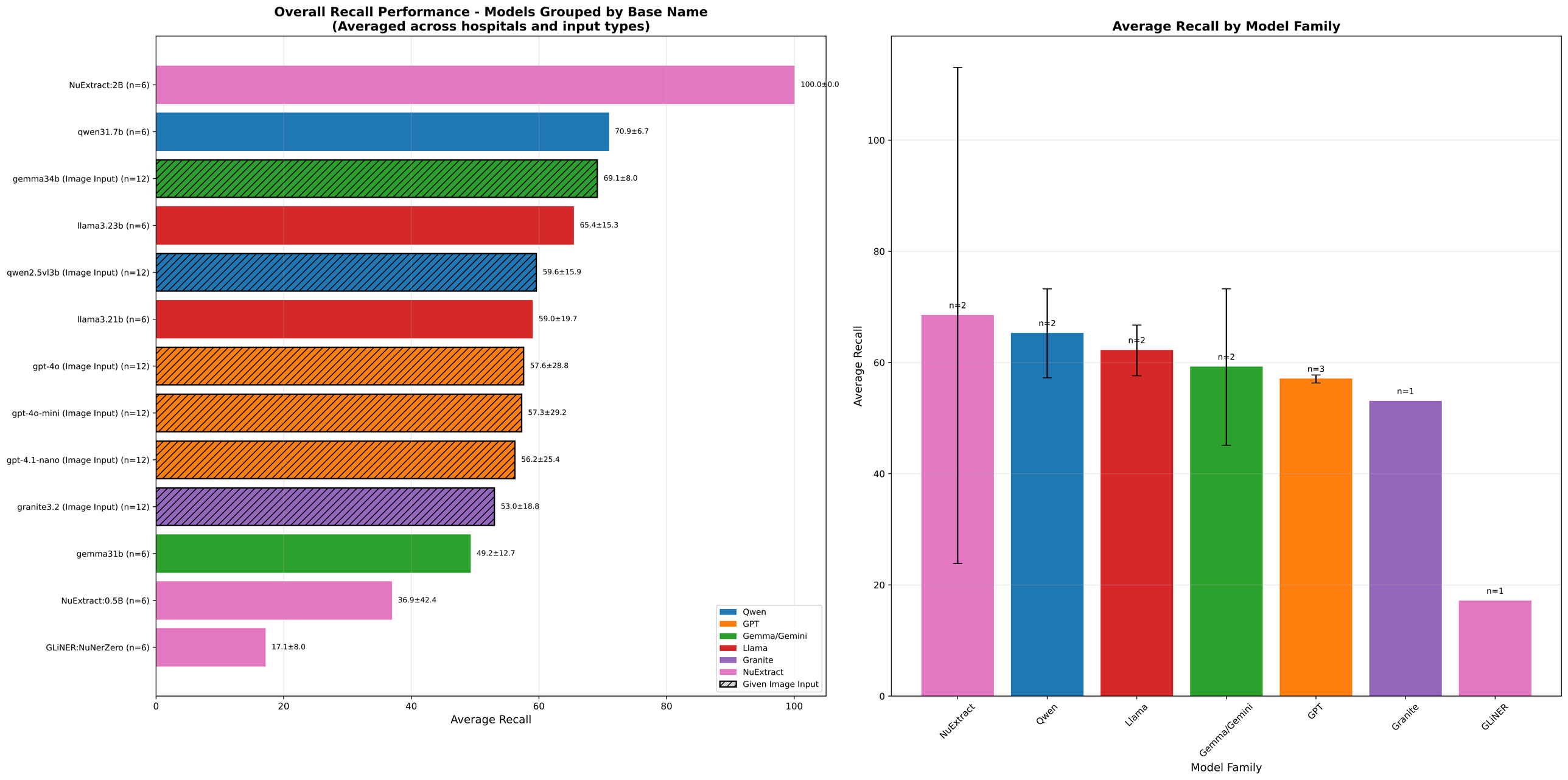
==

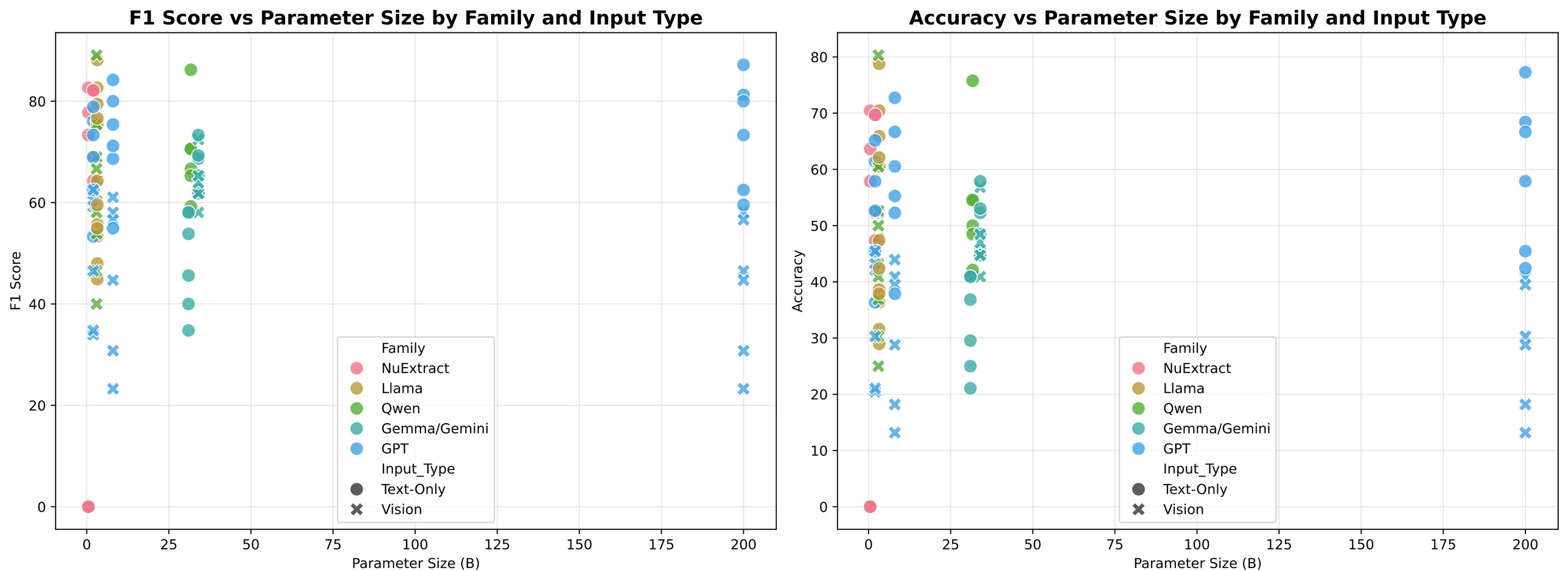

Overall Accuracy Performance - Models Grouped by Base Name
(Averaged across hospitals and input types)
NuExtract:2B (n=6)
qwen31.7b (n=6)
llama3.21b (n=6)
gemma34b (Image Input) (n=12)
qwen2.5vl3b (Image Input) (n=12)
llama3.23b (n=6)
gpt-4o-mini (Image Input) (n=12)
gpt-4o (Image Input) (n=12)
granite3.2 (Image Input) (n=12)
gpt-4.1-nano (Image Input) (n=12)
gemma31b (n=6)
NuExtract:0.5B (n=6)
GLiNER:NuNerZero (n=6)
57.3±9.8
54.2±11.5
52.3±21.9
48.7±5.2
48.4±14.9
47.9±11.1
44.1±18.4
44.1±20.2
43.7±17.8
43.5±14.6
32.4±8.4
32.0±35.3
12.0±4.7
Qwen
GPT
Gemma/Gemini
Llama
Granite
NuExtract
Given Image Input
0
10
20
30
40
50
60
Average Accuracy

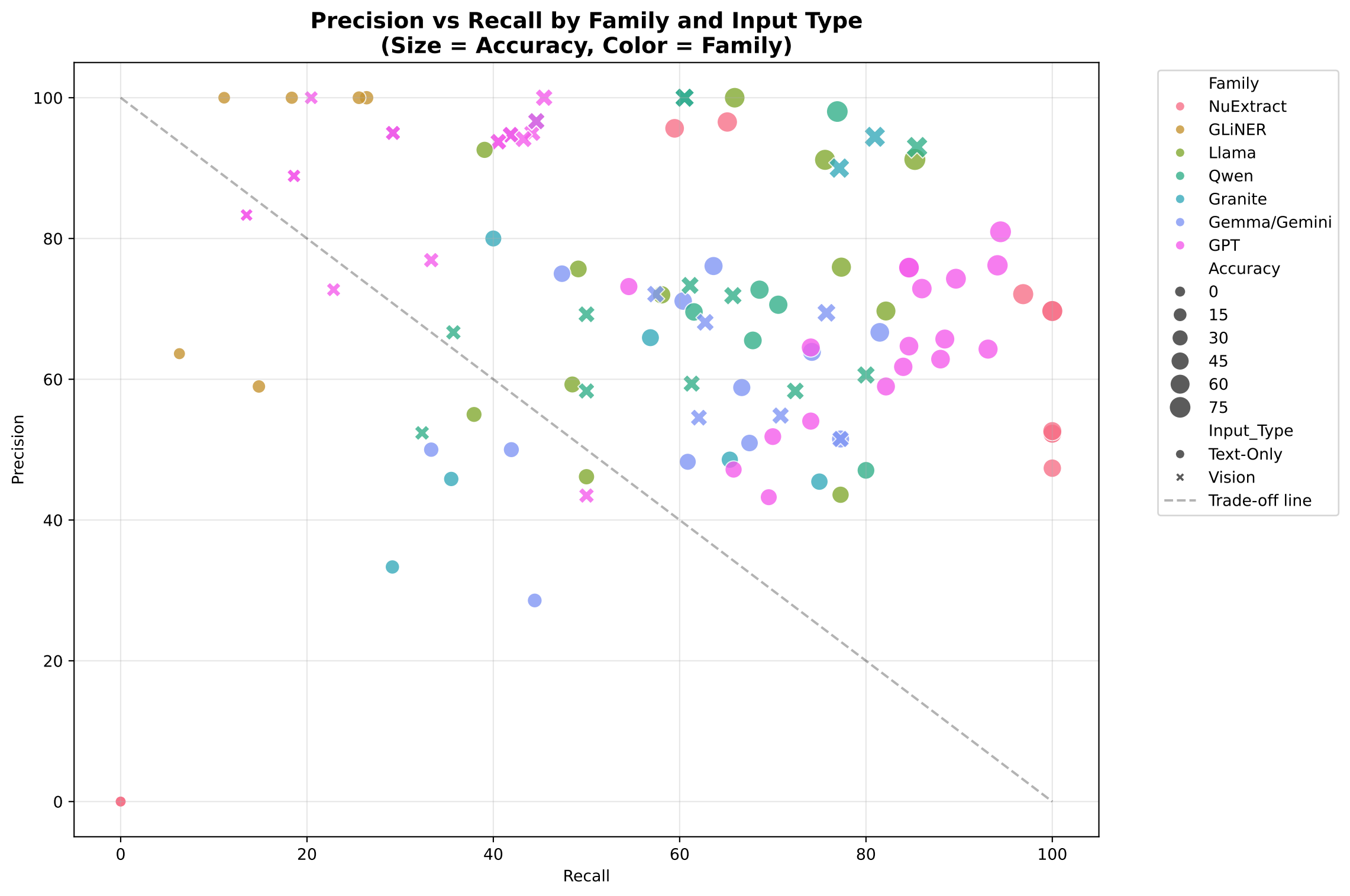

Average Accuracy by Model Family
n=2
n=2
n=2
n=3
n=1
n=2
n=1
0
10
20
30
40
50
60
Average Accuracy
Qwen
Llama
NuExtract
GPT
Granite
Gemma/Gemini
GLiNER
Model Family

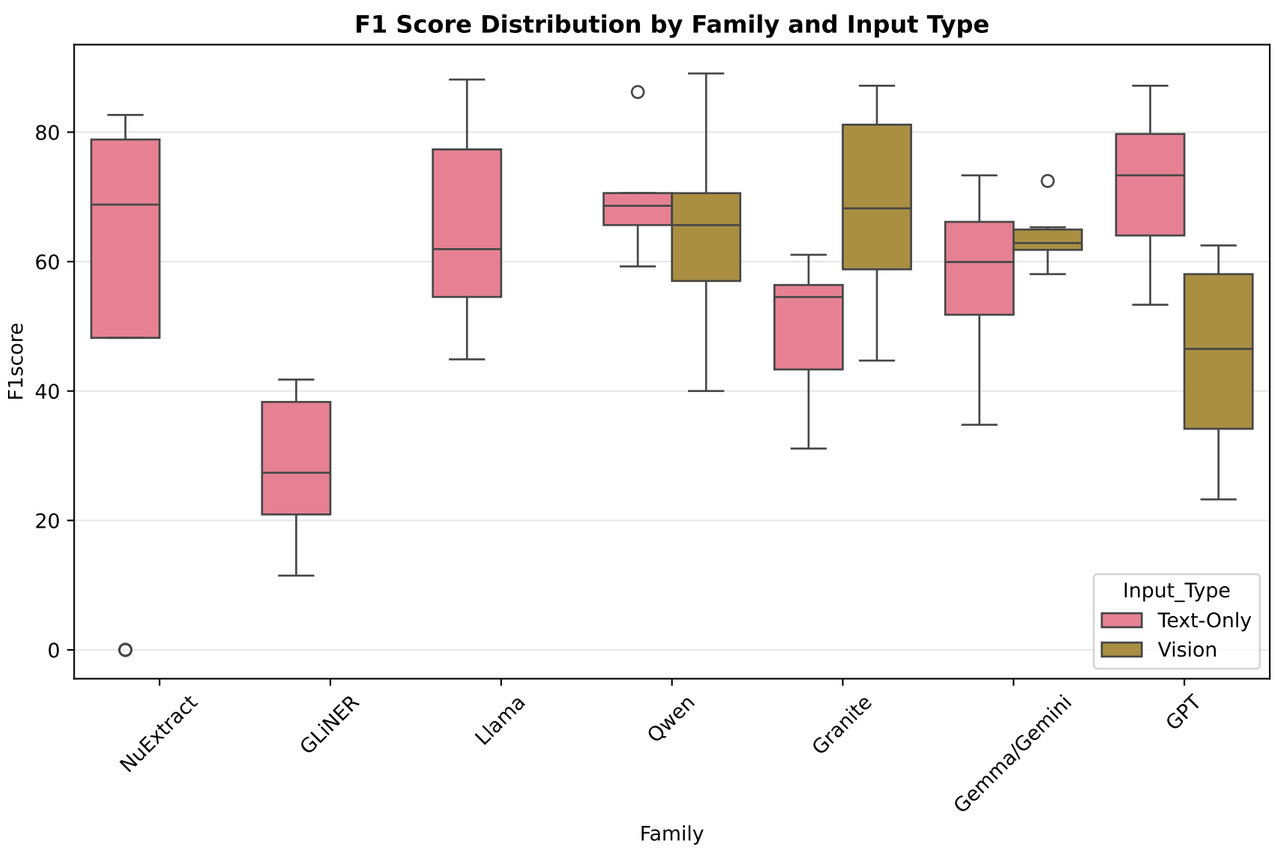

Overall F1 Score Performance - Models Grouped by Base Name
(Averaged across hospitals and input types)
NuExtract:2B (n=6)
qwen31.7b (n=6)
llama3.21b (n=6)
gemma34b (Image Input) (n=12)
llama3.23b (n=6)
qwen2.5vl3b (Image Input) (n=12)
gpt-4.1-nano (Image Input) (n=12)
gpt-4o-mini (Image Input) (n=12)
granite3.2 (Image Input) (n=12)
gpt-4o (Image Input) (n=12)
gemma31b (n=6)
NuExtract:0.5B (n=6)
GLiNER:NuNerZero (n=6)
72.5±7.7
69.8±9.1
66.3±19.2
65.4±4.7
64.2±9.9
64.0±13.4
59.2±14.8
59.1±18.8
58.9±16.6
58.6±20.0
48.4±9.8
39.0±42.8
28.1±12.0
Qwen
GPT
Gemma/Gemini
Llama
Granite
NuExtract
Given Image Input
0
10
20
30
40
50
60
70
Average F1 Score
Average F1 Score by Model Family
Average F1 Score
0
10
20
30
40
50
60
70
80
n=2
n=2
n=3
n=1
n=2
n=2
n=1
Qwen
Llama
GPT
Granite
Gemma/Gemini
NuExtract
GLiNER
Model Family

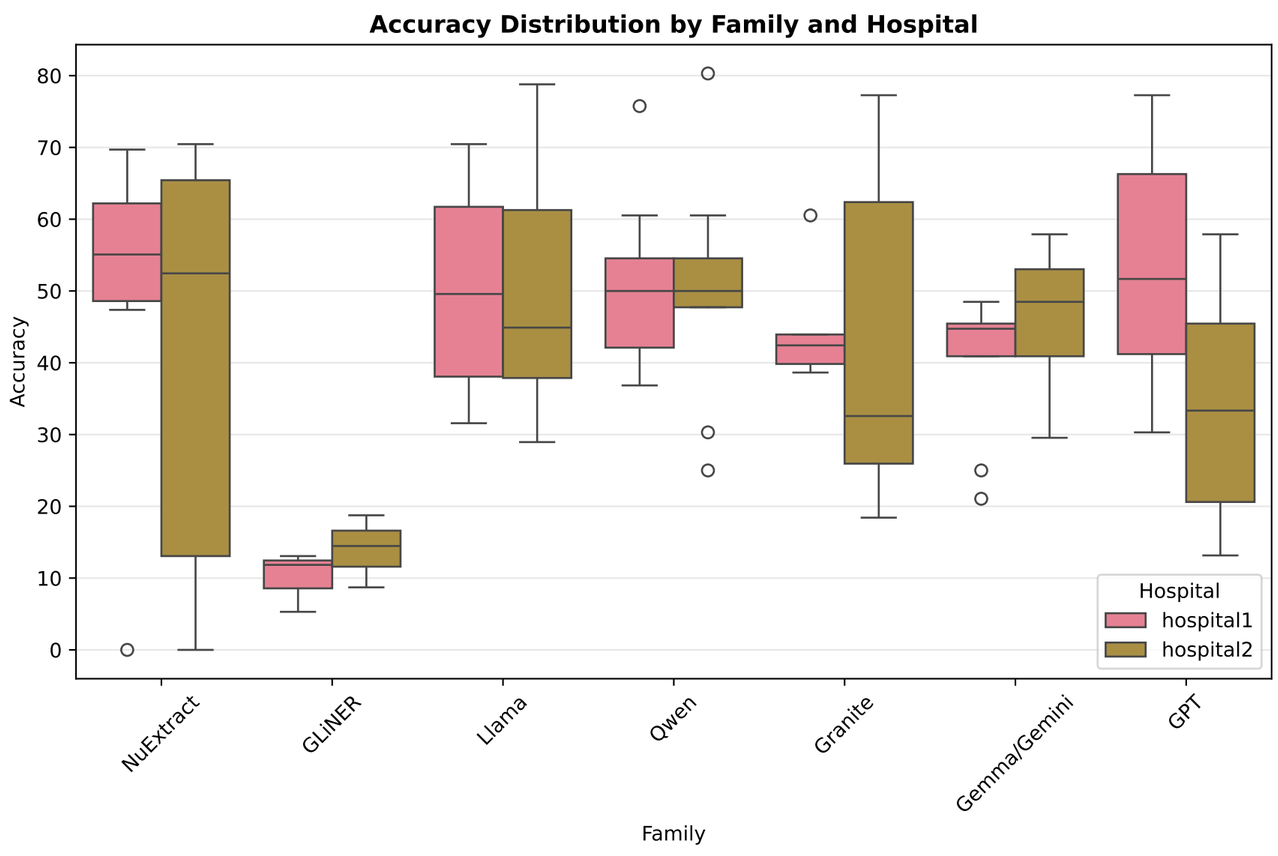

Overall Precision Performance - Models Grouped by Base Name
(Averaged across hospitals and input types)
GLiNER:NuNerZero (n=6)
llama3.21b (n=6)
gpt-4o-mini (Image Input) (n=12)
gpt-4.1-nano (Image Input) (n=12)
gpt-4o (Image Input) (n=12)
granite3.2 (Image Input) (n=12)
qwen2.5vl3b (Image Input) (n=12)
qwen31.7b (n=6)
llama3.23b (n=6)
gemma34b (Image Input) (n=12)
NuExtract:2B (n=6)
gemma31b (n=6)
NuExtract:0.5B (n=6)
87.1±20.0
79.4±22.7
78.2±16.6
76.6±17.5
75.1±17.9
74.2±24.8
71.9±16.8
70.6±16.4
66.0±12.6
63.2±8.6
57.3±9.8
50.5±14.7
44.0±49.0
Qwen
GPT
Gemma/Gemini
Llama
Granite
NuExtract
Given Image Input
0
20
40
60
80
Average Precision
Average Precision by Model Family
n=1
n=3
n=1
n=2
n=2
n=2
n=2
Average Precision
0
20
40
60
80
GLiNER
GPT
Granite
Llama
Qwen
Gemma/Gemini
NuExtract
Model Family

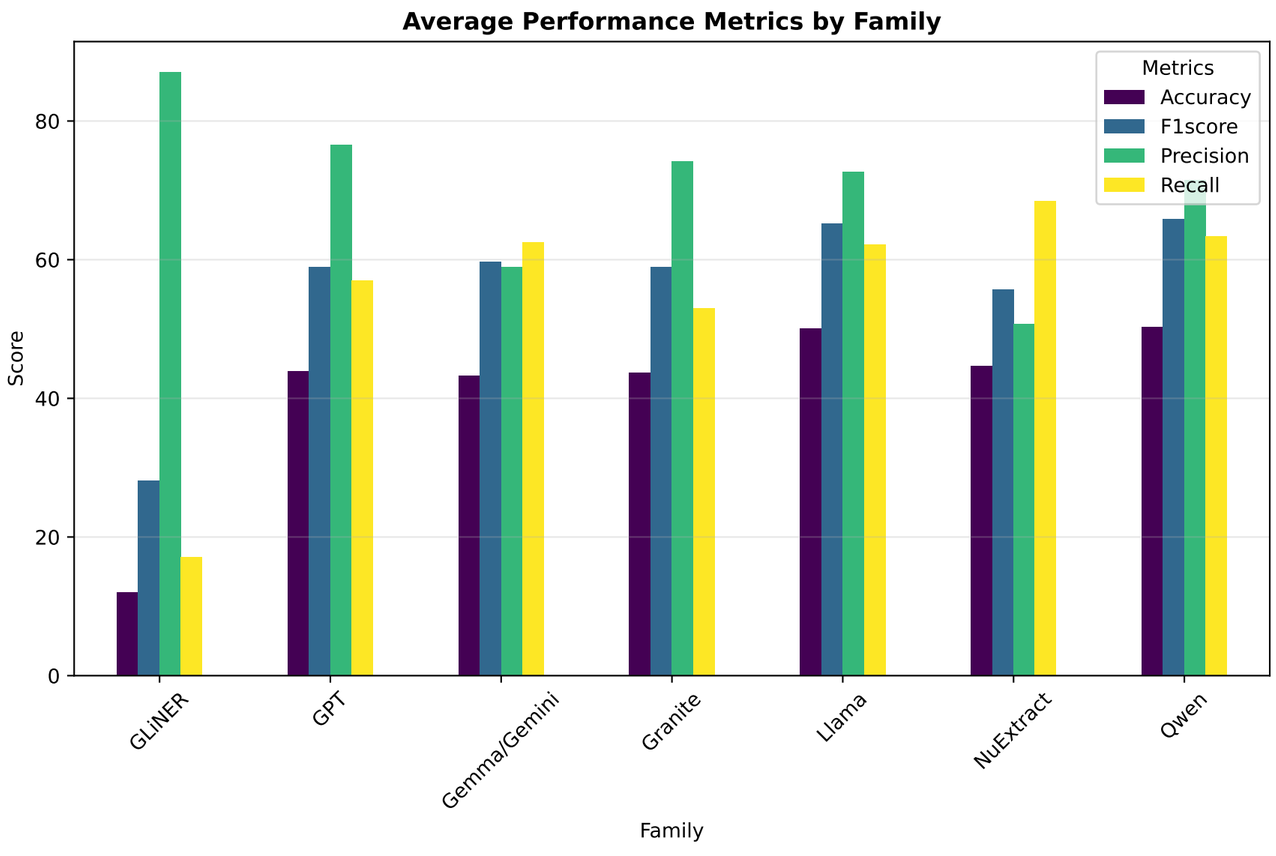

Overall Recall Performance - Models Grouped by Base Name
(Averaged across hospitals and input types)
NuExtract:2B (n=6)
qwen31.7b (n=6)
gemma34b (Image Input) (n=12)
llama3.23b (n=6)
qwen2.5vl3b (Image Input) (n=12)
llama3.21b (n=6)
gpt-4o (Image Input) (n=12)
gpt-4o-mini (Image Input) (n=12)
gpt-4.1-nano (Image Input) (n=12)
granite3.2 (Image Input) (n=12)
gemma31b (n=6)
NuExtract:0.5B (n=6)
GLiNER:NuNerZero (n=6)
100.0±0.0
70.9±6.7
69.1±8.0
65.4±15.3
59.6±15.9
59.0±19.7
57.6±28.8
57.3±29.2
56.2±25.4
53.0±18.8
49.2±12.7
36.9±42.4
17.1±8.0
Qwen
GPT
Gemma/Gemini
Llama
Granite
NuExtract
Given Image Input
0
20
40
60
80
100
Average Recall
Average Recall by Model Family
Average Recall
n=2
n=2
n=2
n=2
n=3
n=1
n=1
NuExtract
Qwen
Llama
Gemma/Gemini
GPT
Granite
GLiNER
Model Family

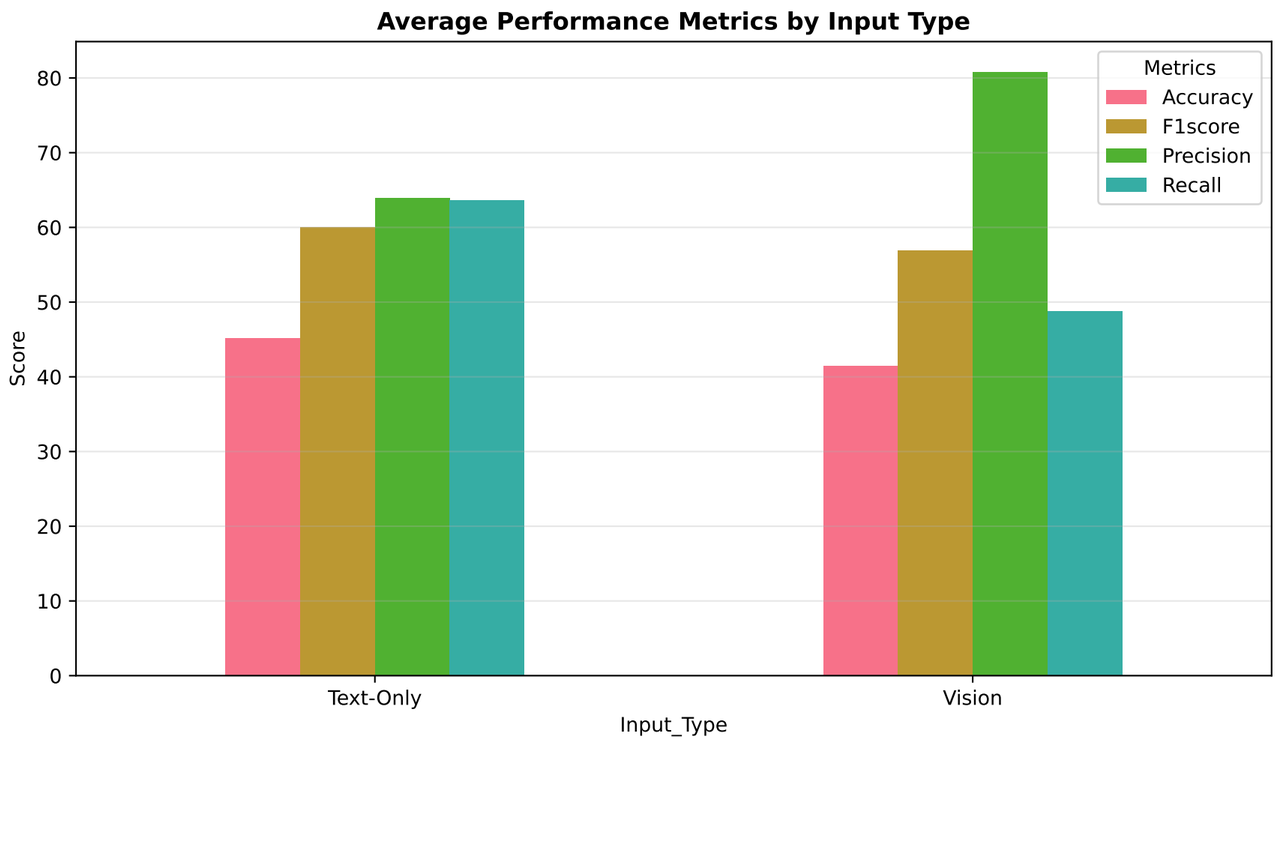

F1 Score vs Parameter Size by Family and Input Type
F1 Score
Parameter Size (B)
Family
NuExtract
Llama
Qwen
Gemma/Gemini
GPT
Input_Type
Text-Only
Vision
Accuracy vs Parameter Size by Family and Input Type
Accuracy
Parameter Size (B)

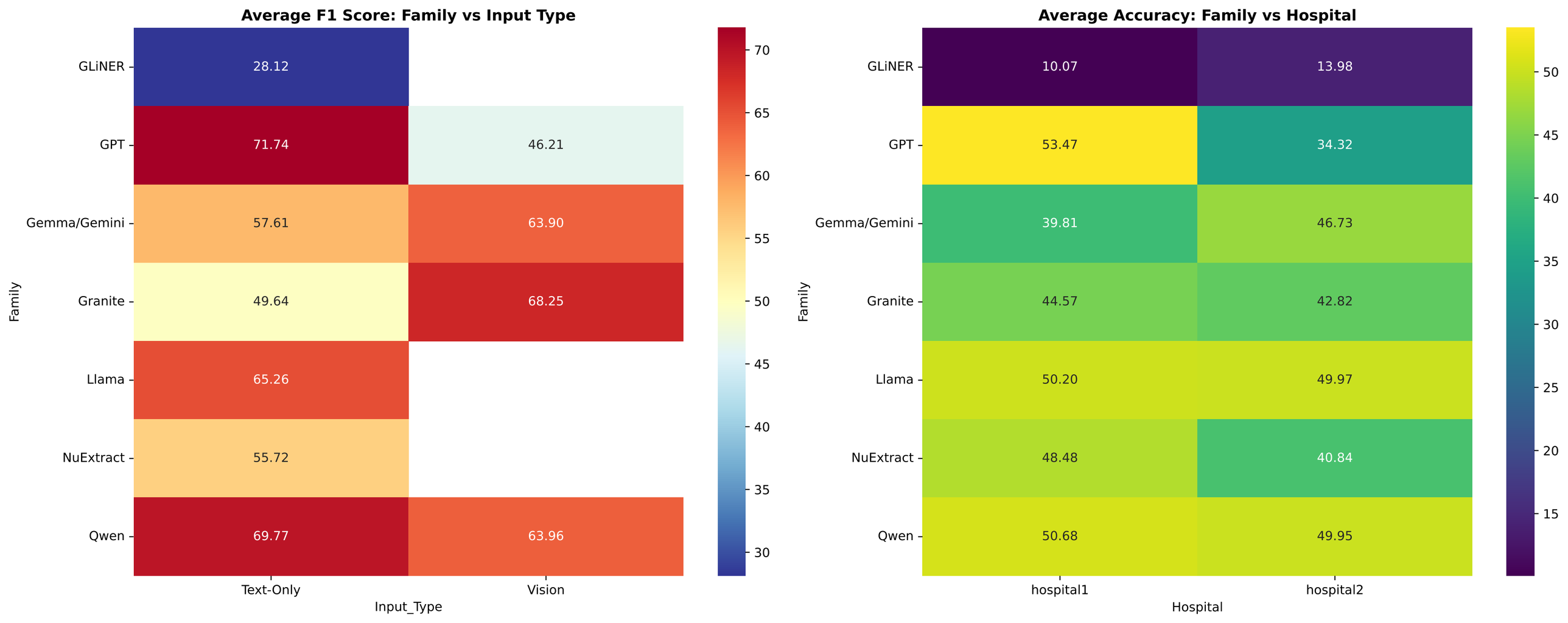

Precision vs Recall by Family and Input Type
(Size = Accuracy, Color = Family)
Precision
Recall
0
20
40
60
80
100
Family
NuExtract
GLiNER
Llama
Qwen
Granite
Gemma/Gemini
GPT
Accuracy
0
15
30
45
60
75
Input_Type
Text-Only
Vision
Trade-off line

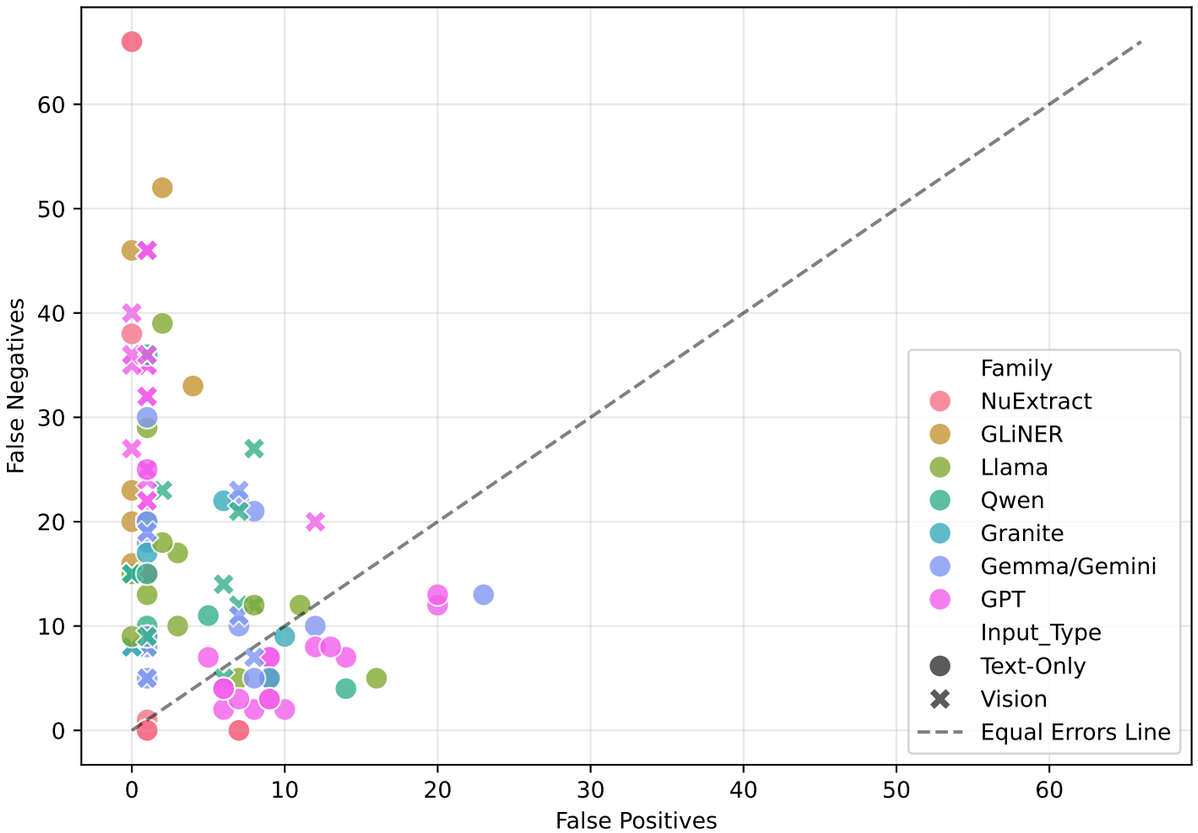

F1 Score Distribution by Family and Input Type
F1score
Family
NuExtract
GLiNER
Llama
Qwen
Granite
Gemma/Gemini
GPT
Input_Type
Text-Only
Vision

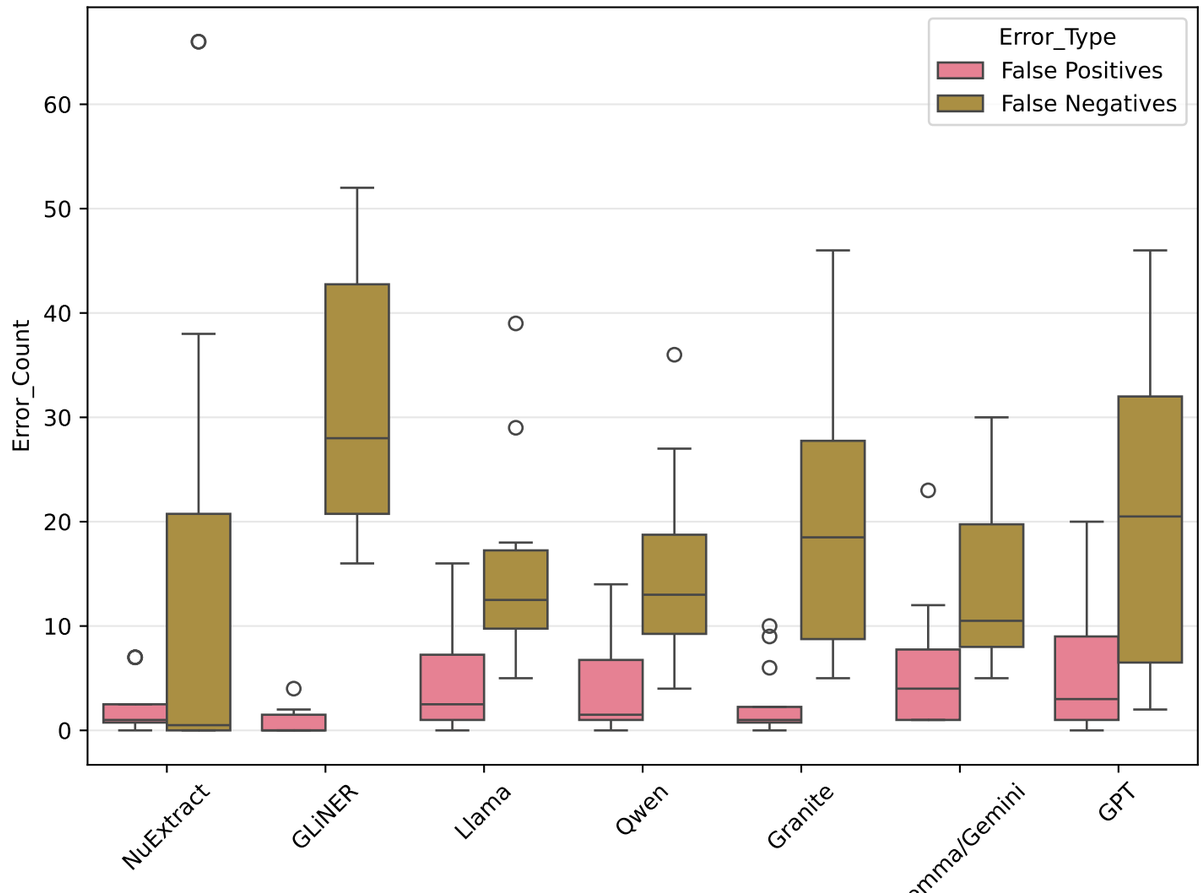

Accuracy Distribution by Family and Hospital
Accuracy
Family
NuExtract
GLiNER
Llama
Qwen
Granite
Gemma/Gemini
GPT
Hospital
hospital1
hospital2

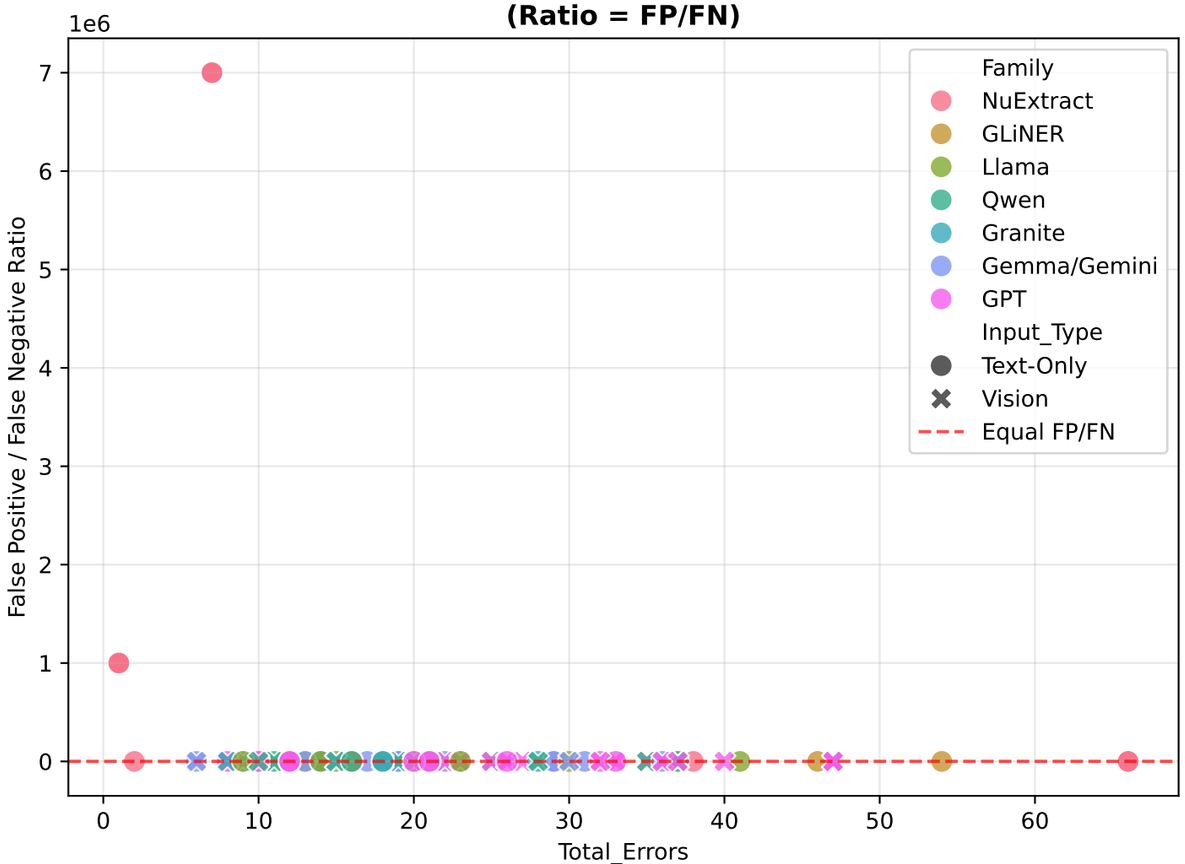

Average Performance Metrics by Family
Score
Family
GLiNER
GPT
Gemma/Gemini
Granite
Llama
NuExtract
Qwen
Metrics
Accuracy
F1score
Precision
Recall

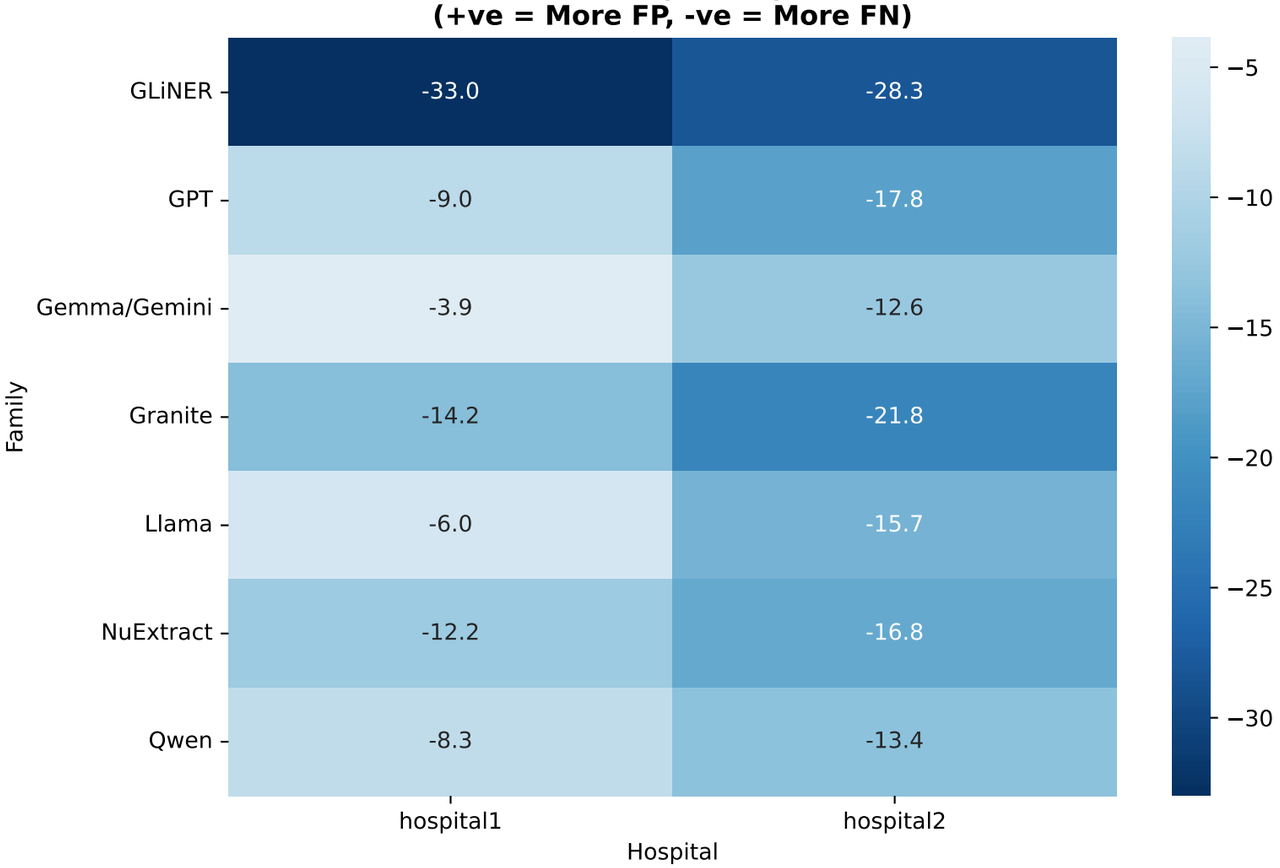

Average Performance Metrics by Input Type
Score
Input_Type
Text-Only
Vision
Metrics
Accuracy
F1score
Precision
Recall

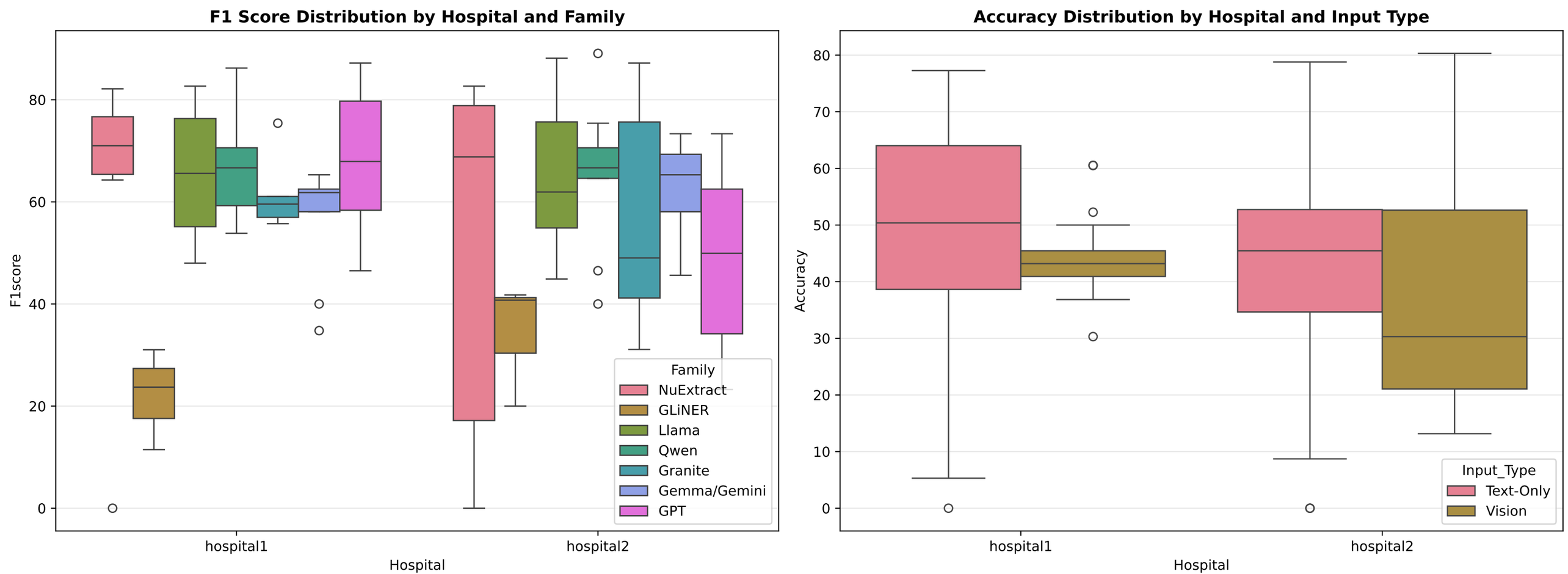

Average F1 Score: Family vs Input Type
GLiNER
GPT
Gemma/Gemini
Granite
Llama
NuExtract
Qwen
Family
28.12
71.74
46.21
57.61
63.90
49.64
68.25
65.26
55.72
69.77
63.96
Text-Only
Vision
Input_Type
70
65
60
55
50
45
40
35
30
Average Accuracy: Family vs Hospital
GLiNER
GPT
Gemma/Gemini
Granite
Llama
NuExtract
Qwen
Family
10.07
13.98
53.47
34.32
39.81
46.73
44.57
42.82
50.20
49.97
48.48
40.84
50.68
49.95
hospital1
hospital2
Hospital
50
45
40
35
30
25
20
15


False Positives vs False Negatives by Family and Input Type
False Negatives
False Positives
Family
NuExtract
GLiNER
Llama
Qwen
Granite
Gemma/Gemini
GPT
Input_Type
Text-Only
Vision
Equal Errors Line

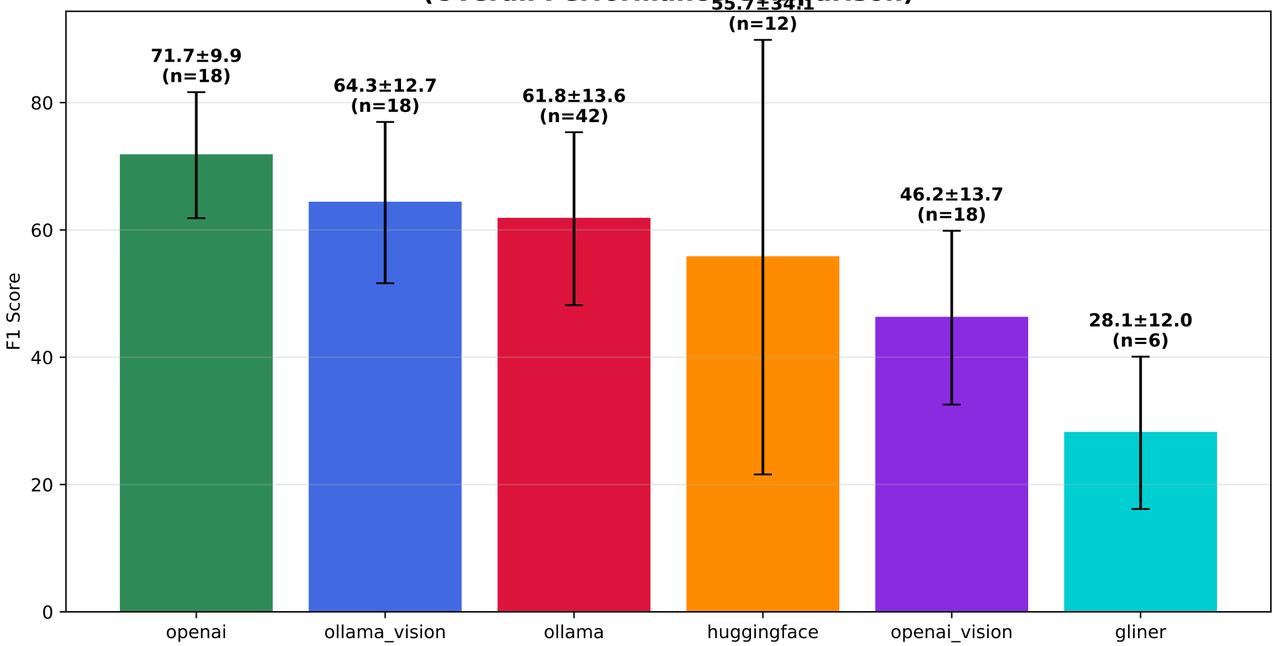

Error Distribution by Model Family
Error_Count
Error_Type
False Positives
False Negatives
NuExtract
GLiNER
Llama
Qwen
Granite
Gemma/Gemini
GPT
Family


Error Ratio vs Total Errors
(Ratio = FP/FN)
1e6
False Positive / False Negative Ratio
Total_Errors
Family
NuExtract
GLiNER
Llama
Qwen
Granite
Gemma/Gemini
GPT
Input_Type
Text-Only
Vision
Equal FP/FN

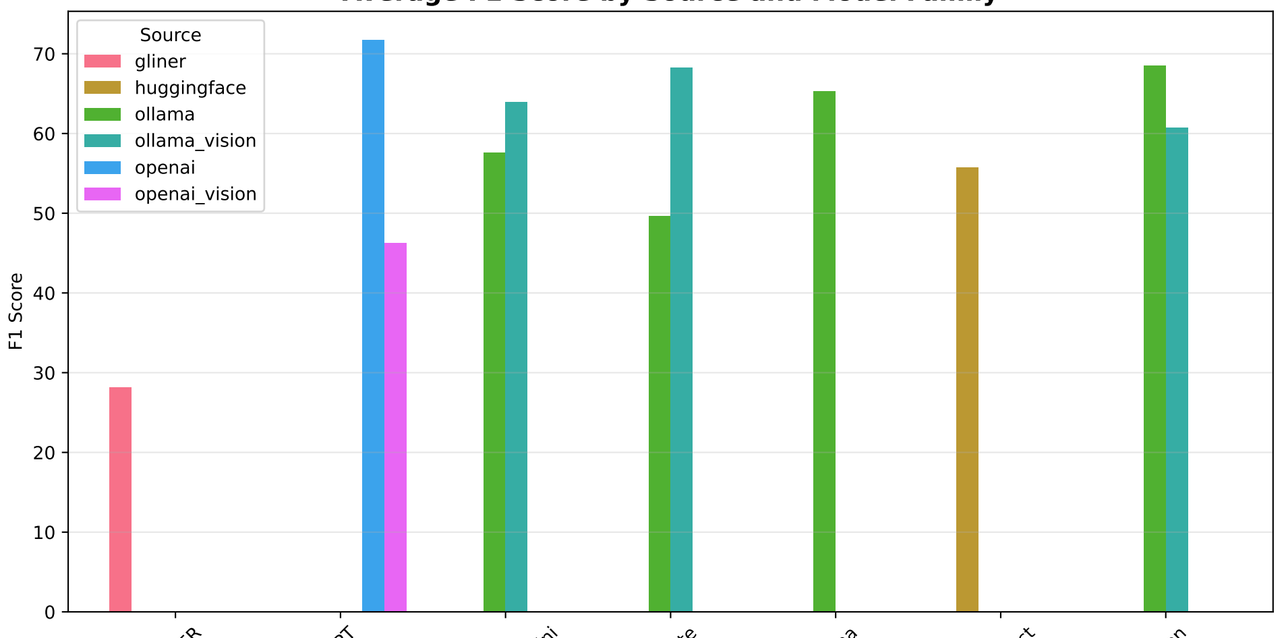

Error Bias: FP - FN by Family and Hospital
(+ve = More FP, -ve = More FN)
Family
GLiNER
GPT
Gemma/Gemini
Granite
Llama
NuExtract
Qwen
-33.0
-28.3
-9.0
-17.8
-3.9
-12.6
-14.2
-21.8
-6.0
-15.7
-12.2
-16.8
-8.3
-13.4
hospital1
hospital2
Hospital


F1 Score Distribution by Hospital and Family
F1score
Hospital
hospital1
hospital2
Family
NuExtract
GLiNER
Llama
Qwen
Granite
Gemma/Gemini
GPT
Accuracy Distribution by Hospital and Input Type
Accuracy
Hospital
hospital1
hospital2
Input_Type
Text-Only
Vision

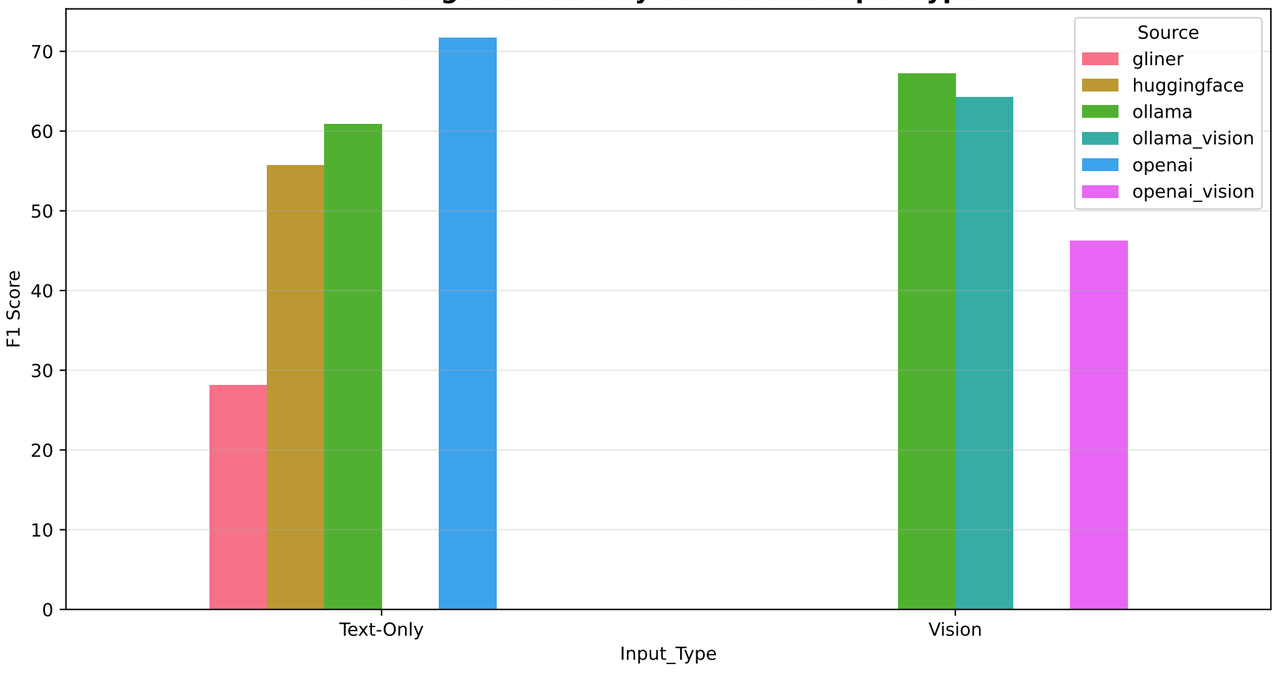
Average F1 Score by Source (Overall Performance Comparison)



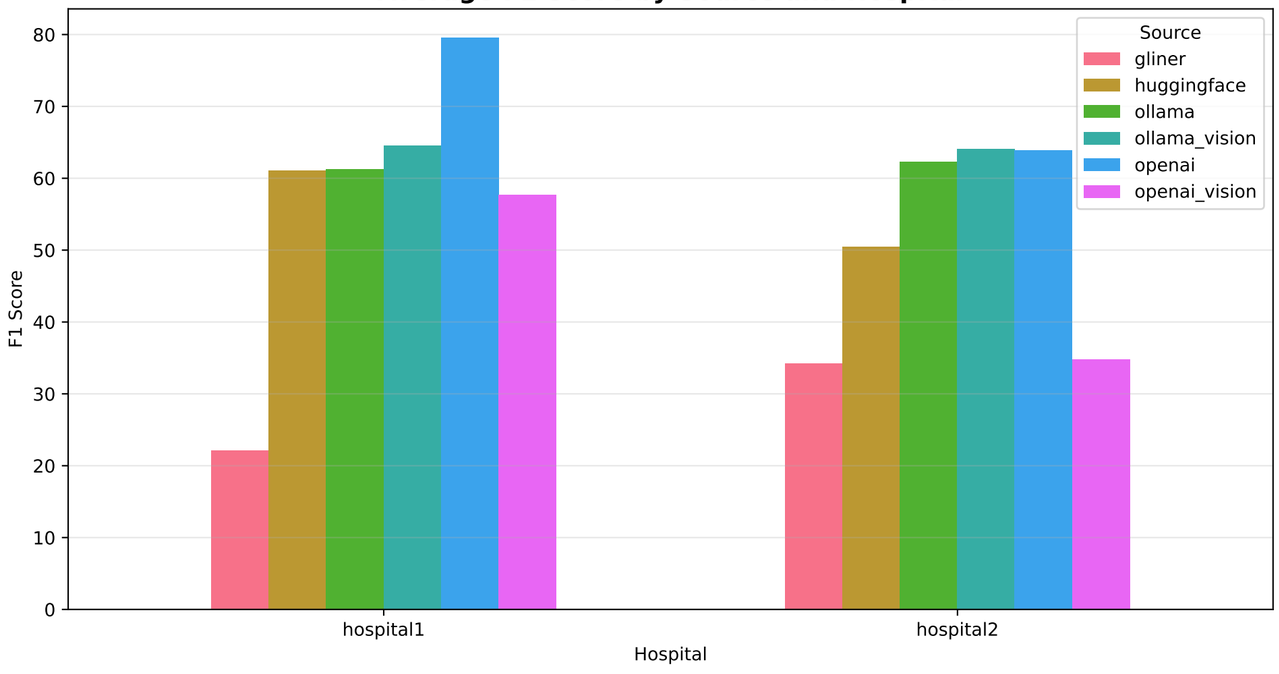
Average F1 Score by Source and Model Family

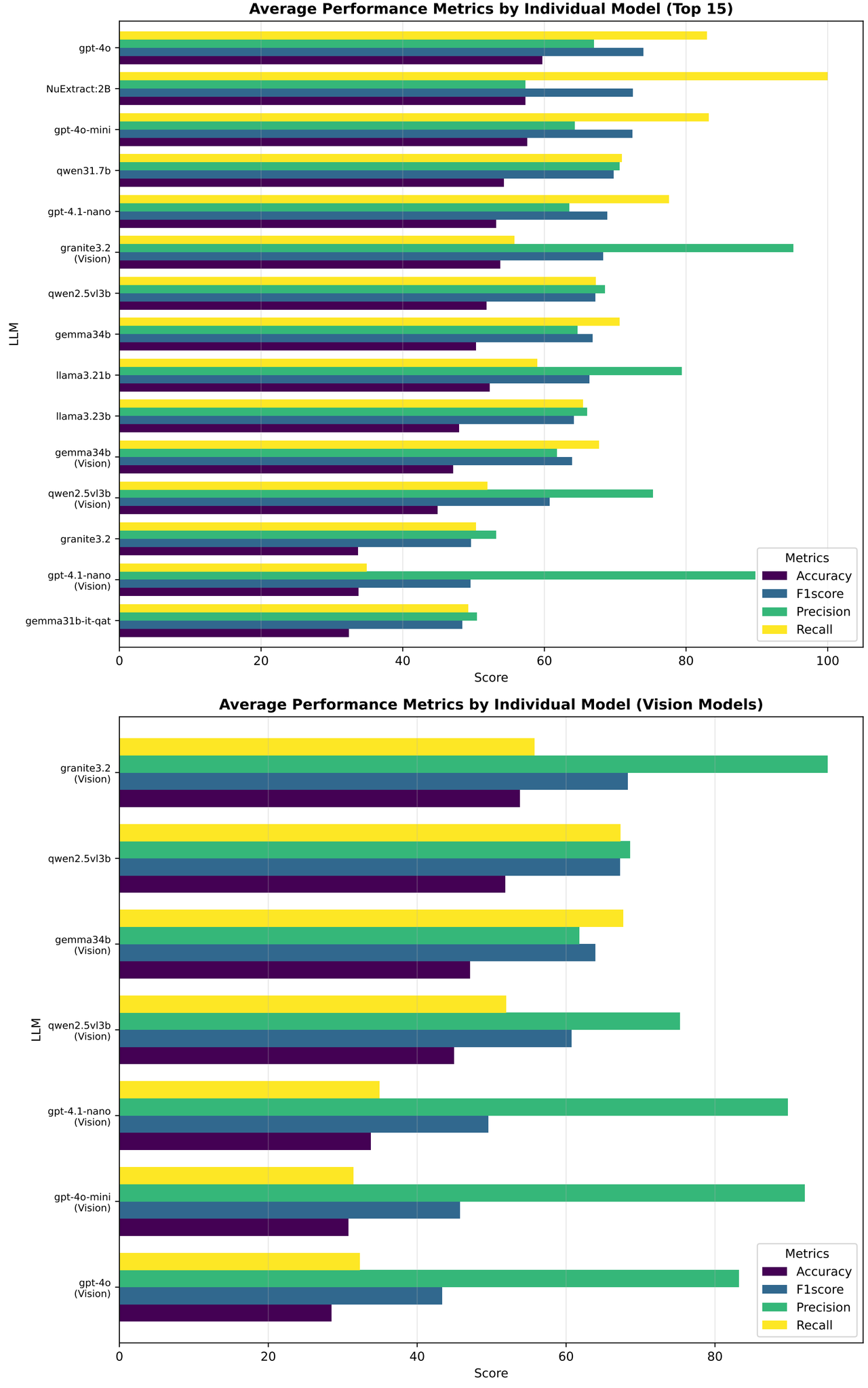


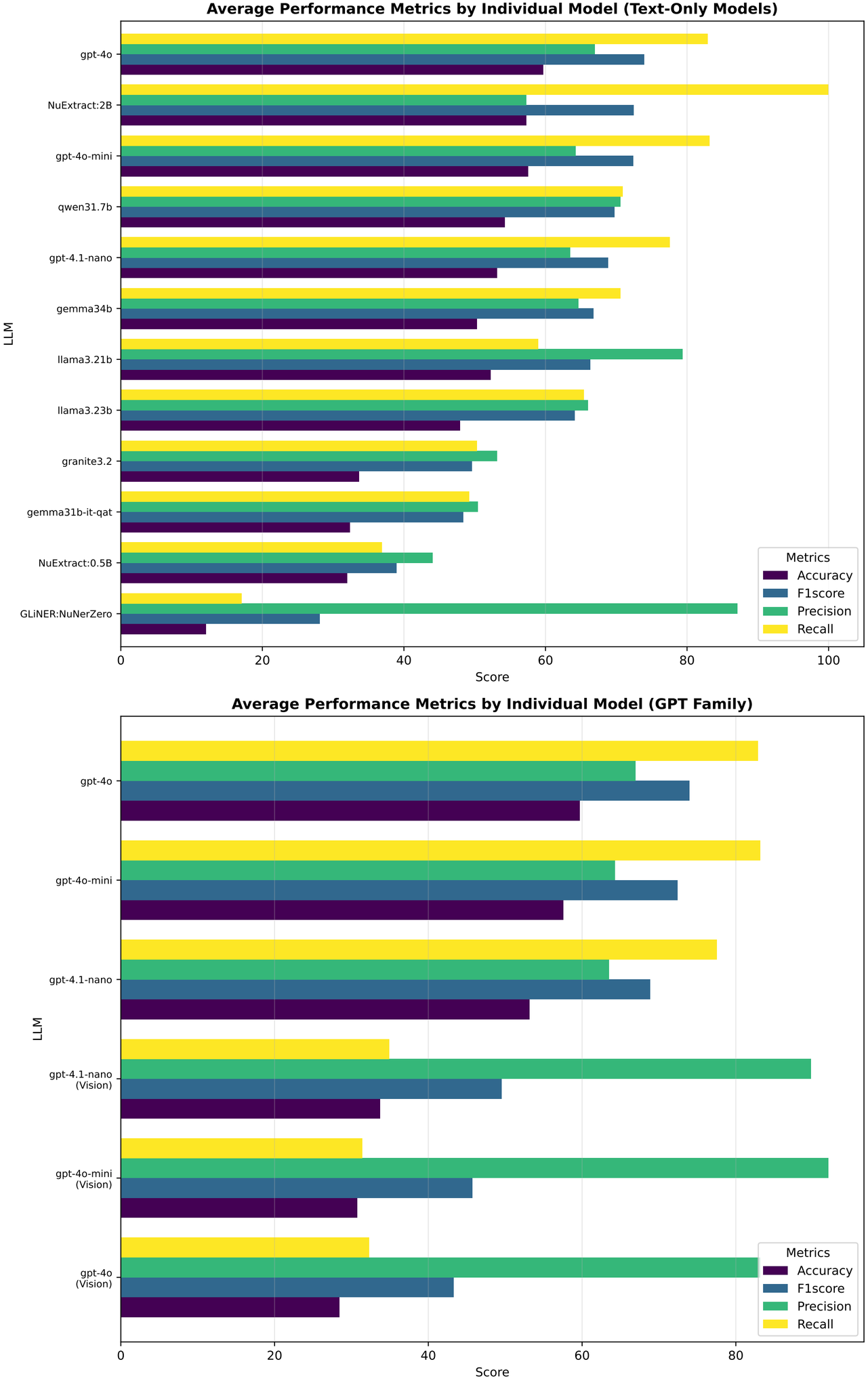
Average F1 Score by Source and Input Type



Average F1 Score by Source and Hospital




Average Performance Metrics by Individual Model (Top 15)
gpt-4o
NuExtract:2B
gpt-4o-mini
qwen31.7b
gpt-4.1-nano
granite3.2 (Vision)
qwen2.5vl3b
gemma34b
llama3.21b
llama3.23b
gemma34b (Vision)
qwen2.5vl3b (Vision)
granite3.2
gpt-4.1-nano (Vision)
gemma31b-it-qat
LLM
0
20
40
60
80
100
Score
Metrics
Accuracy
F1score
Precision
Recall


Average Performance Metrics by Individual Model (Text-Only Models)
gpt-4o
NuExtract:2B
gpt-4o-mini
qwen31.7b
gpt-4.1-nano
gemma34b
llama3.21b
llama3.23b
granite3.2
gemma31b-it-qat
NuExtract:0.5B
GLiNER:NuNerZero
LLM
0
20
40
60
80
100
Score
Metrics
Accuracy
F1score
Precision
Recall

Average Performance Metrics by Individual Model (Vision Models)
granite3.2 (Vision)
qwen2.5vl3b
gemma34b (Vision)
qwen2.5vl3b (Vision)
gpt-4.1-nano (Vision)
gpt-4o-mini (Vision)
gpt-4o (Vision)
LLM
0
20
40
60
80
Score
Metrics
Accuracy
F1score
Precision
Recall

Average Performance Metrics by Individual Model (GPT Family)
gpt-4o
gpt-4o-mini
gpt-4.1-nano
gpt-4.1-nano (Vision)
gpt-4o-mini (Vision)
gpt-4o (Vision)
LLM
0
20
40
60
80
Score
Metrics
Accuracy
F1score
Precision
Recall

Performance Metrics for All Individual Models
(Sorted by F1 Score)