

## 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:

- qwen2.5v3b (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:

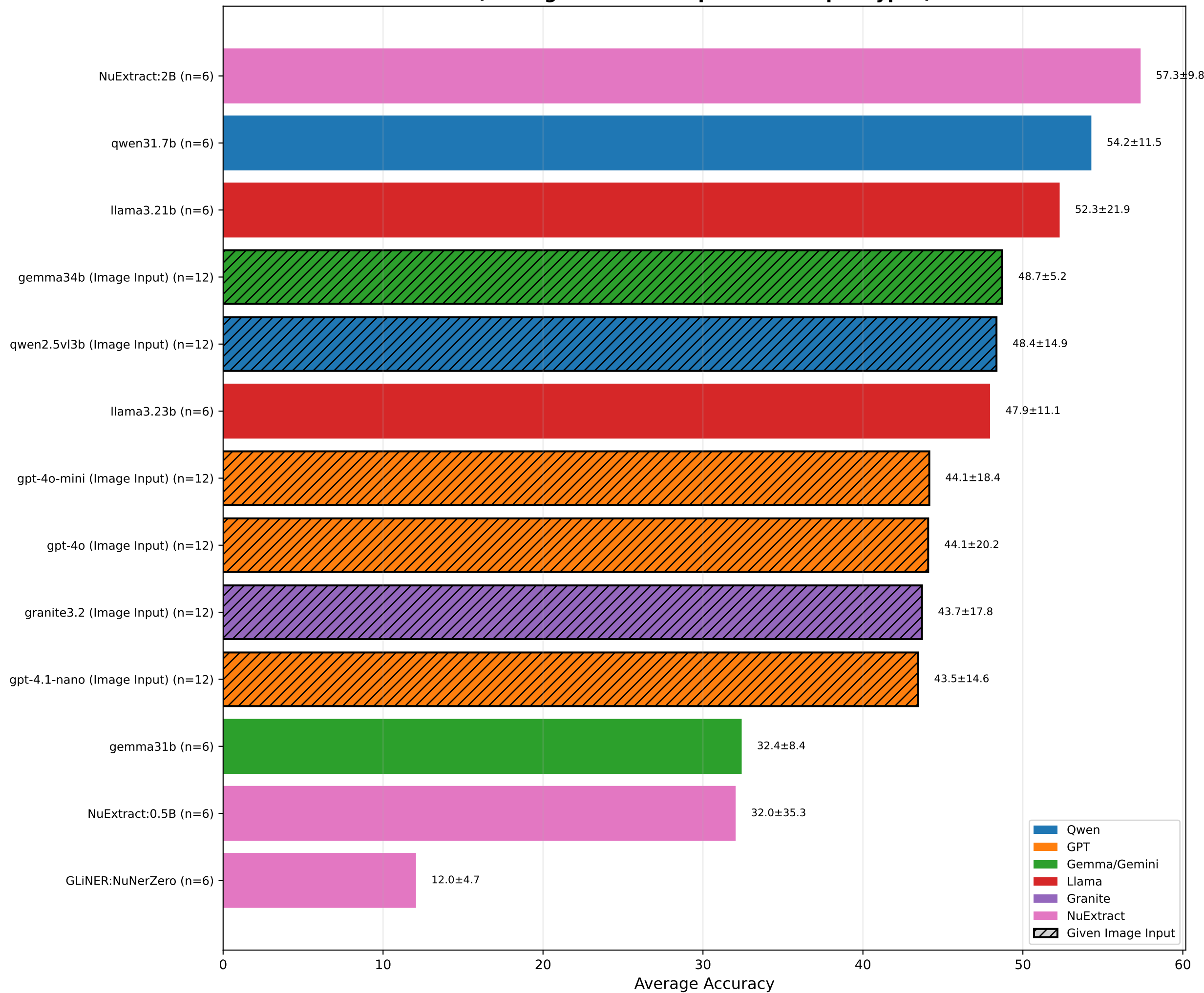
- GLiNER:NuNerZero (GLiNER, Text-only): F1 = 28.12 ± 11.97
- NuExtract:0.5B (NuExtract, Text-only): F1 = 38.96 ± 42.78
- gemma31b (Gemma/Gemini, Text-only): F1 = 48.40 ± 9.80
- gpt-4o (GPT, with Vision): F1 = 58.65 ± 19.97
- granite3.2 (Granite, with Vision): F1 = 58.94 ± 16.63

## Error Analysis Summary:

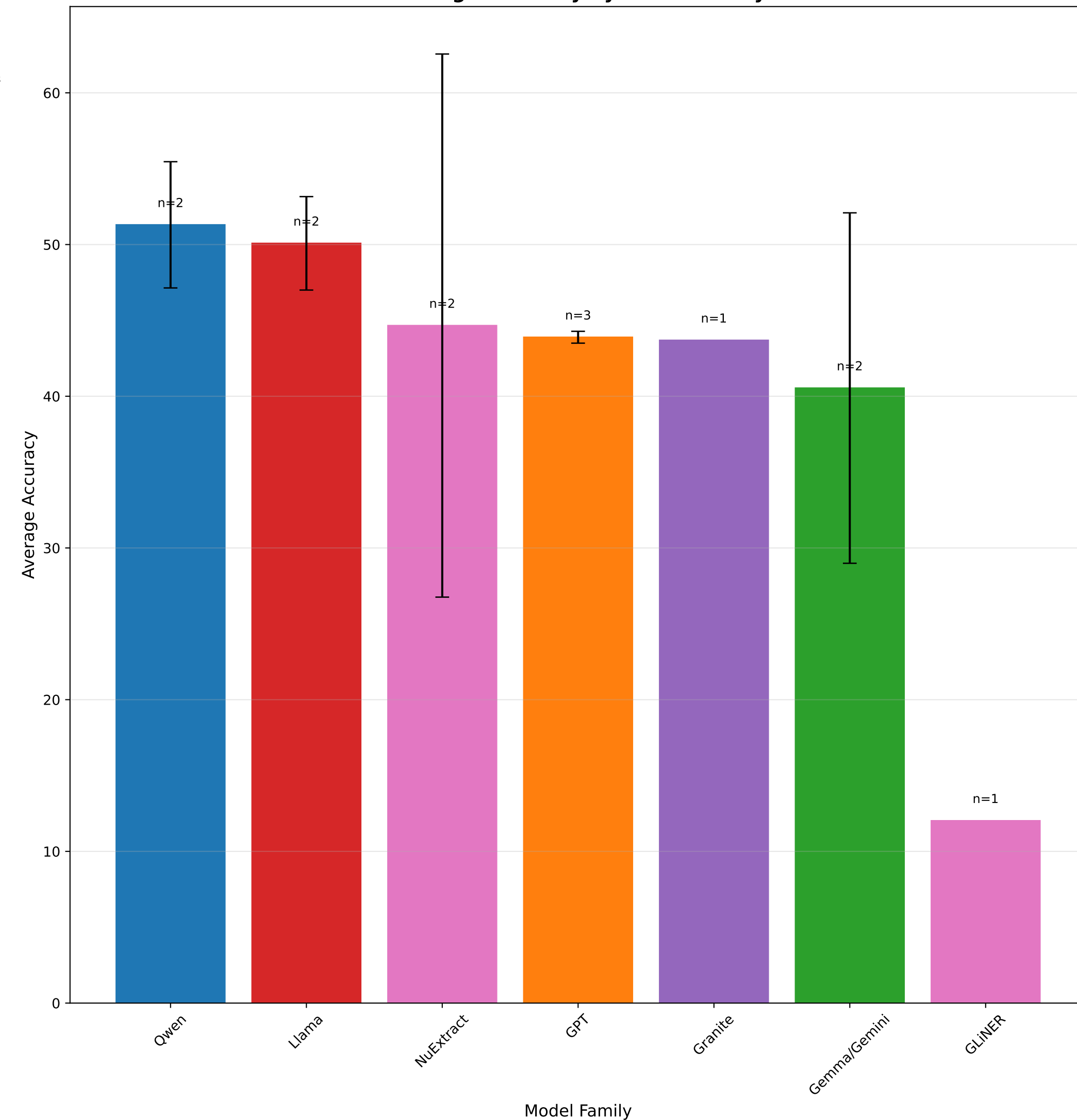
- Average False Positives: 4.23
- Average False Negatives: 17.66
- Models with more FP than FN: 32
- Models with more FN than FP: 81

=====

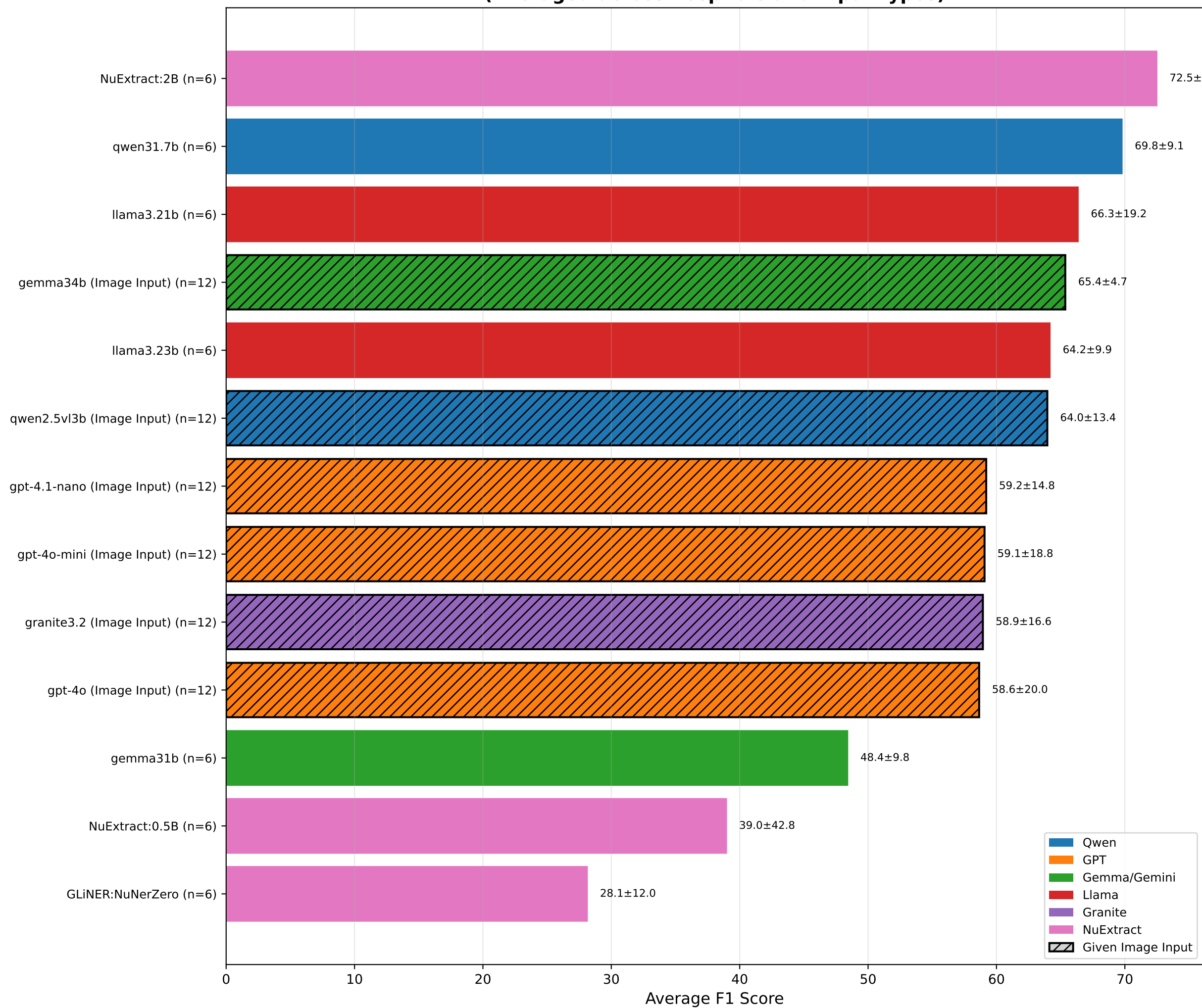
Overall Accuracy Performance - Models Grouped by Base Name  
(Averaged across hospitals and input types)



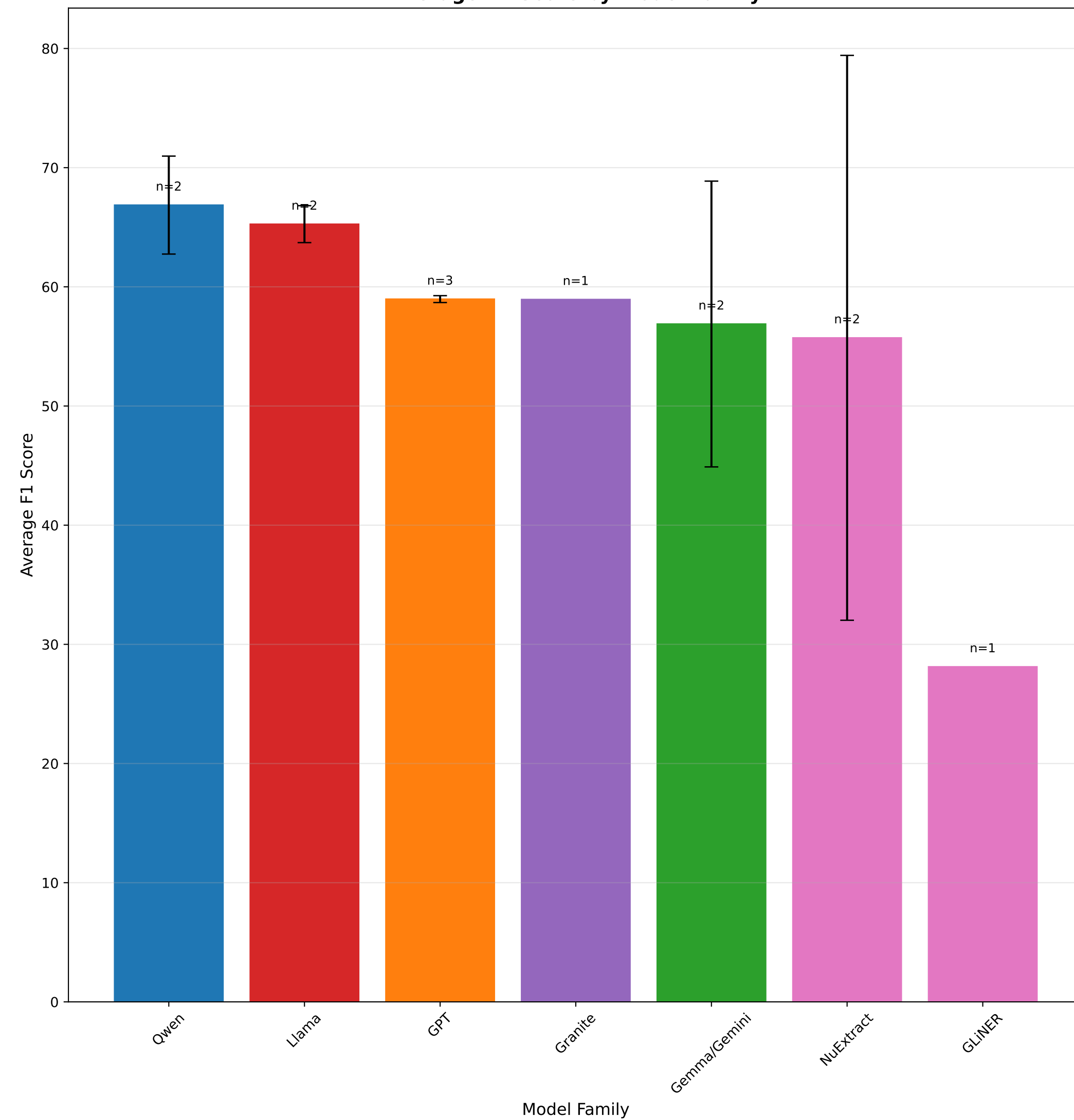
Average Accuracy by Model Family



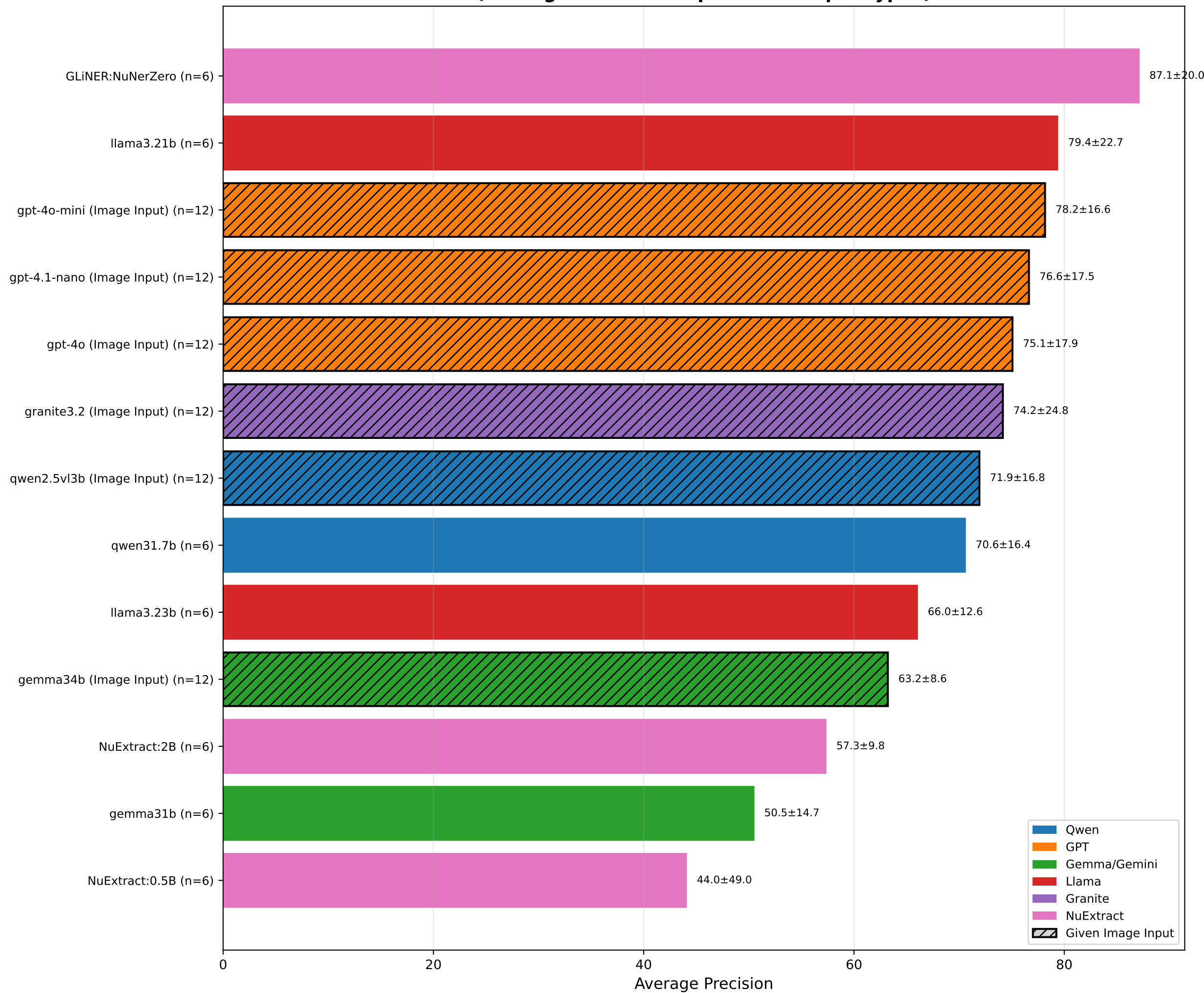
Overall F1 Score Performance - Models Grouped by Base Name  
(Averaged across hospitals and input types)



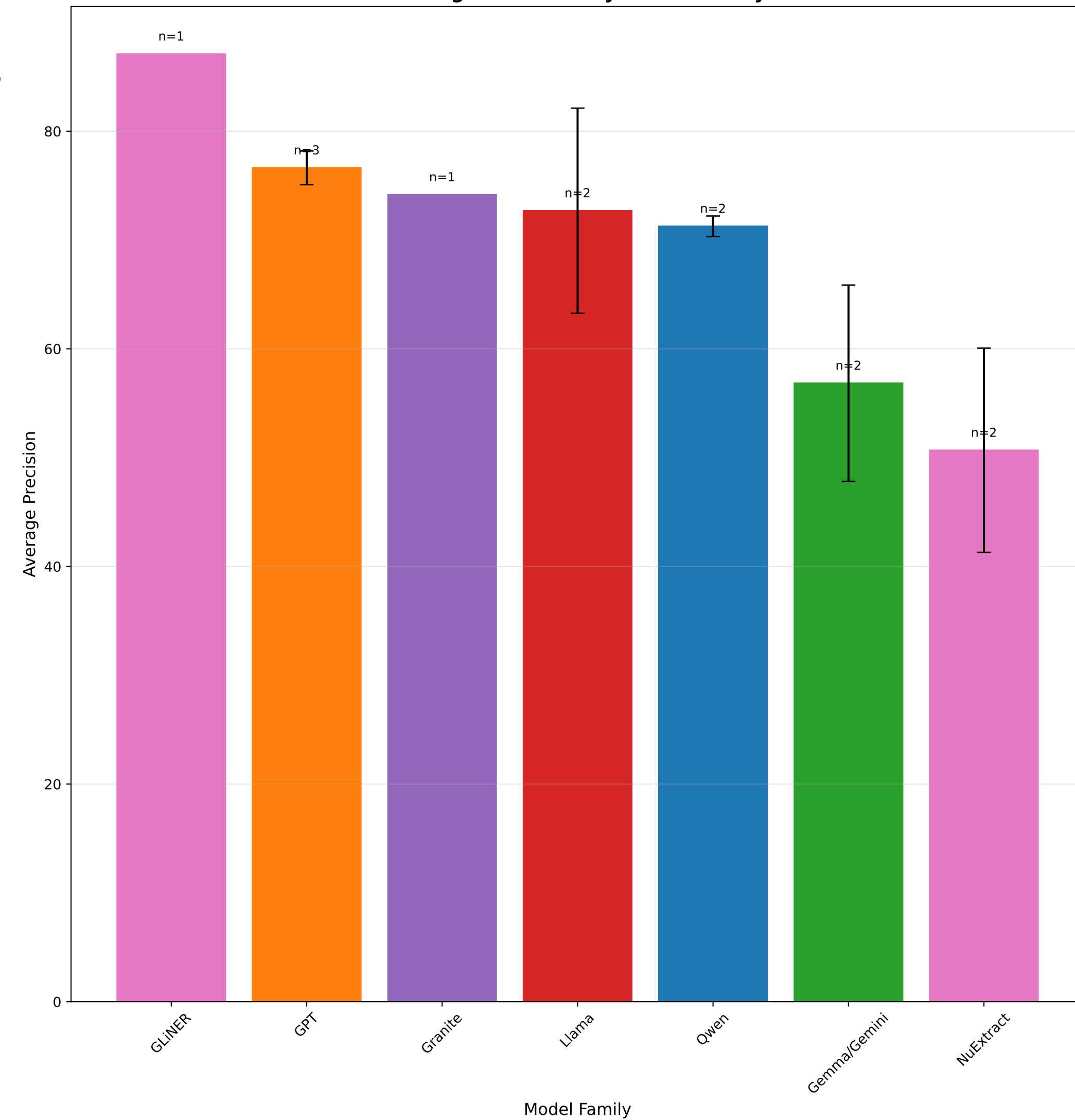
Average F1 Score by Model Family



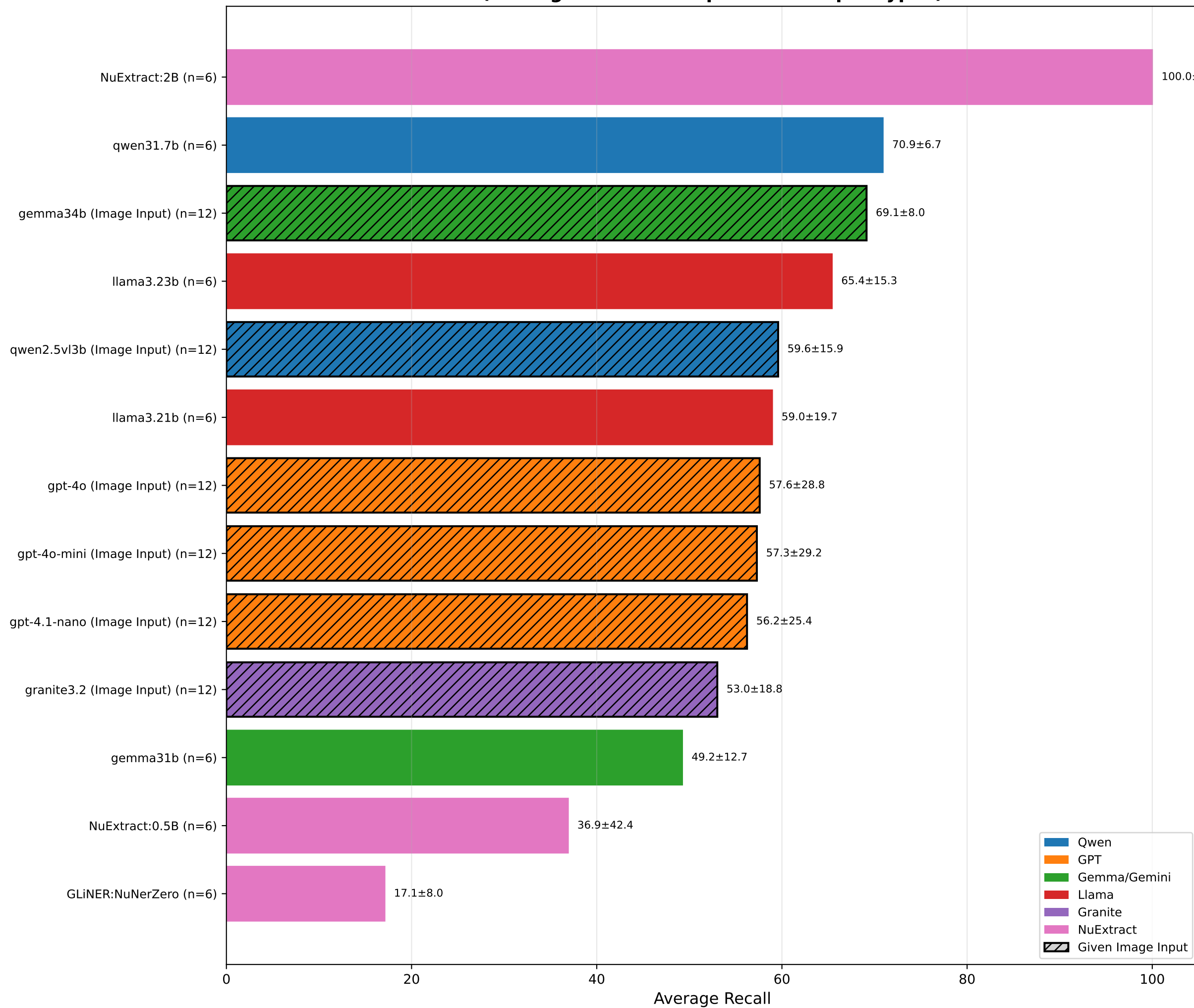
Overall Precision Performance - Models Grouped by Base Name  
(Averaged across hospitals and input types)



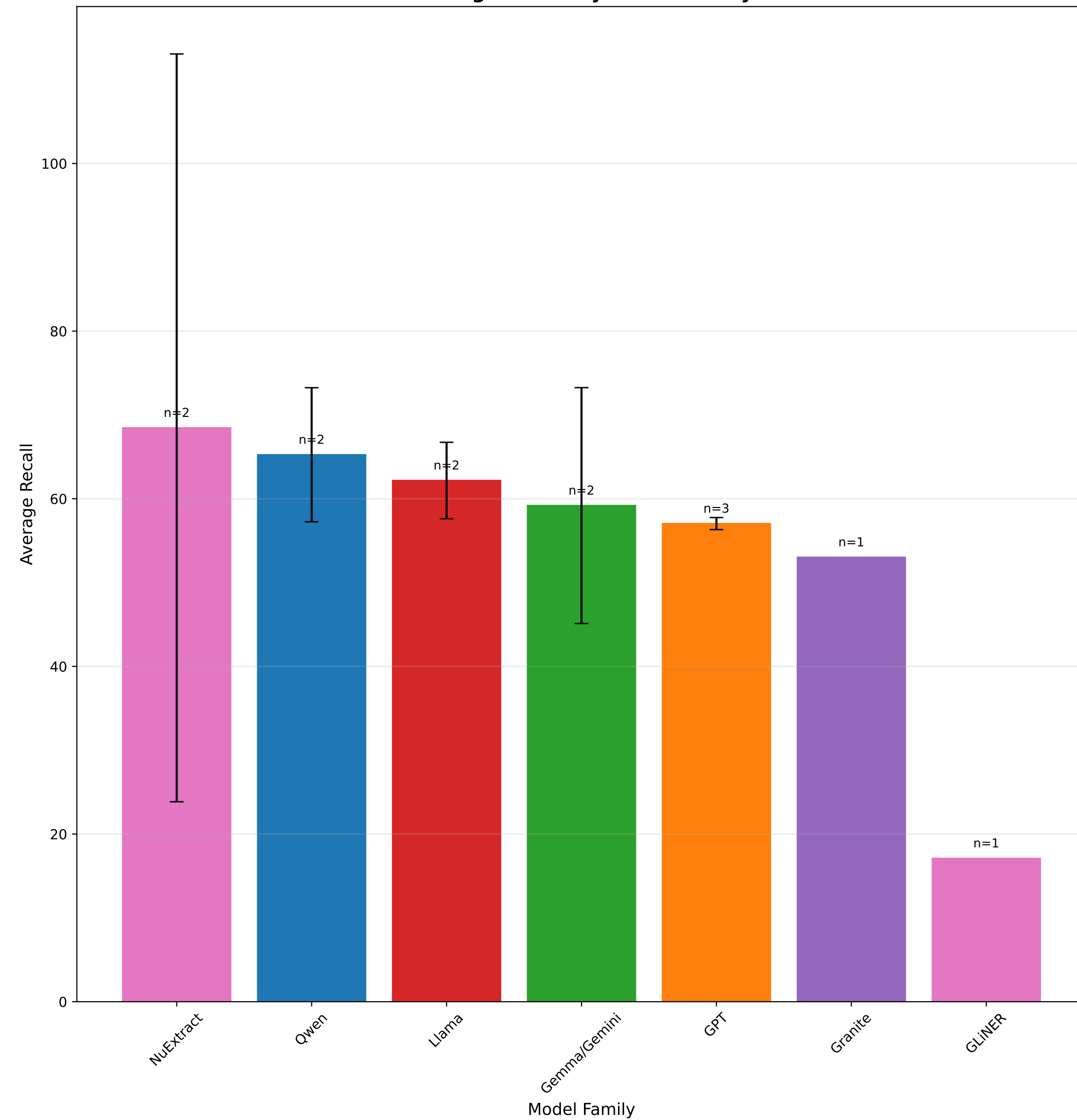
Average Precision by Model Family



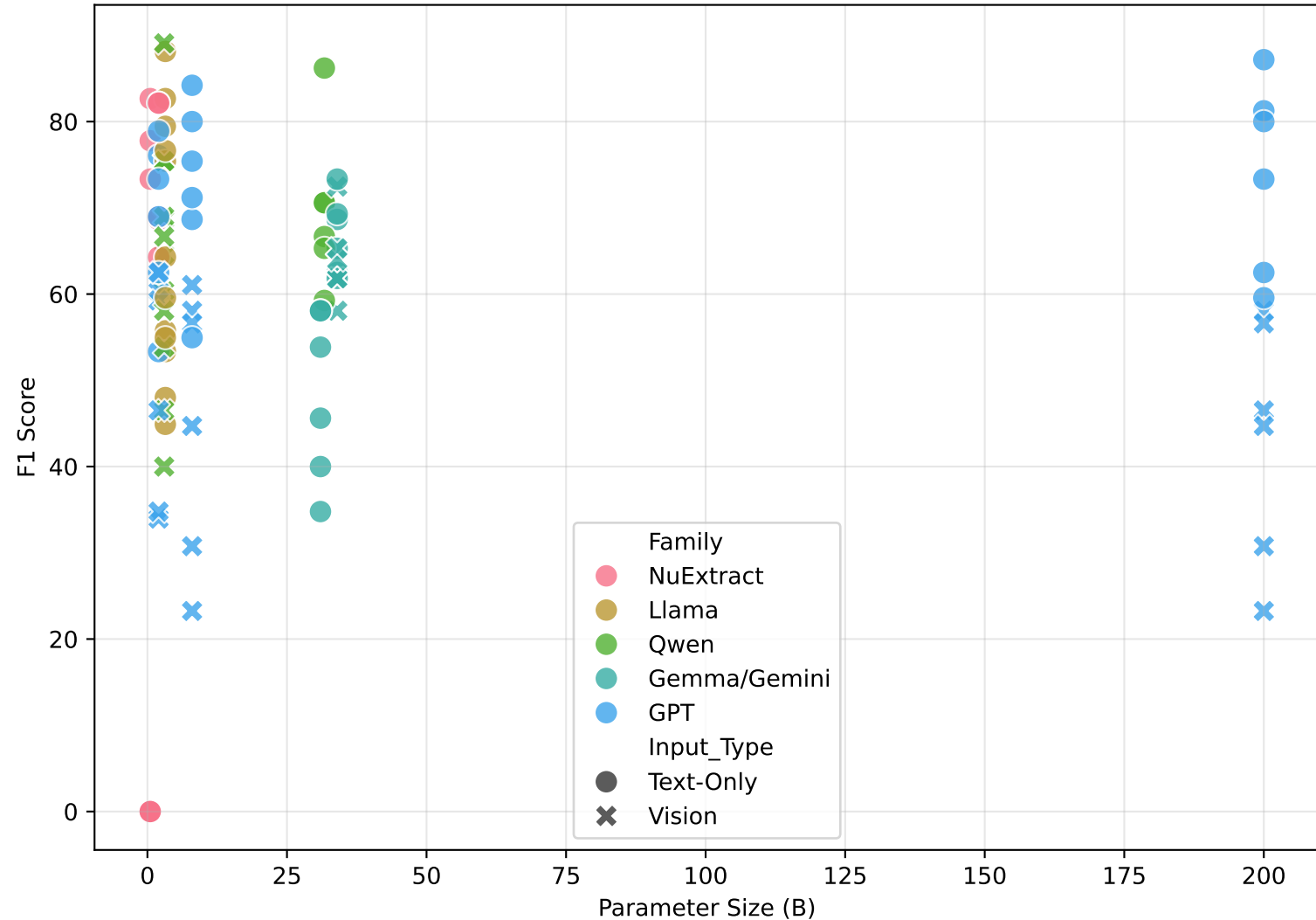
**Overall Recall Performance - Models Grouped by Base Name**  
(Averaged across hospitals and input types)



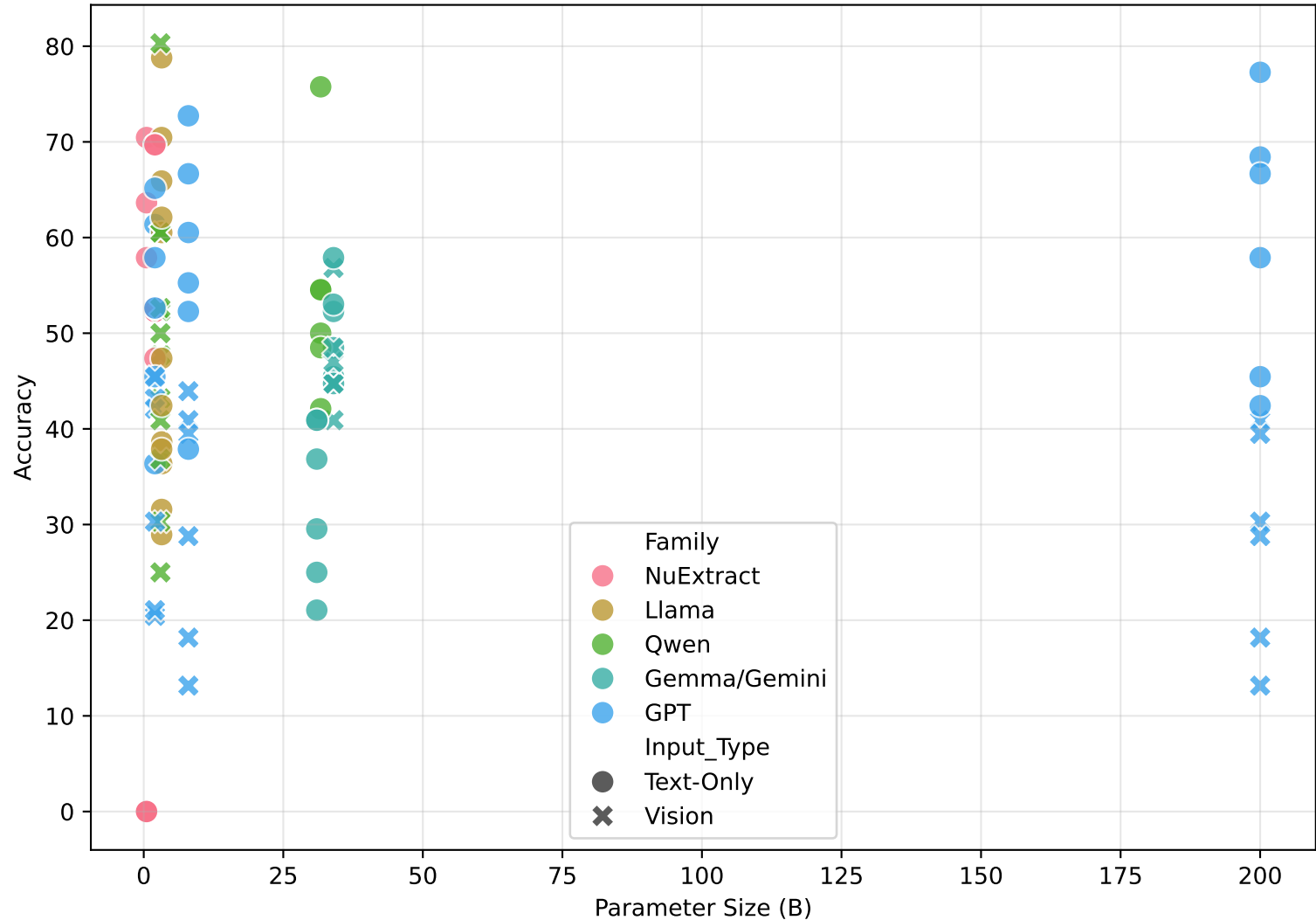
**Average Recall by Model Family**



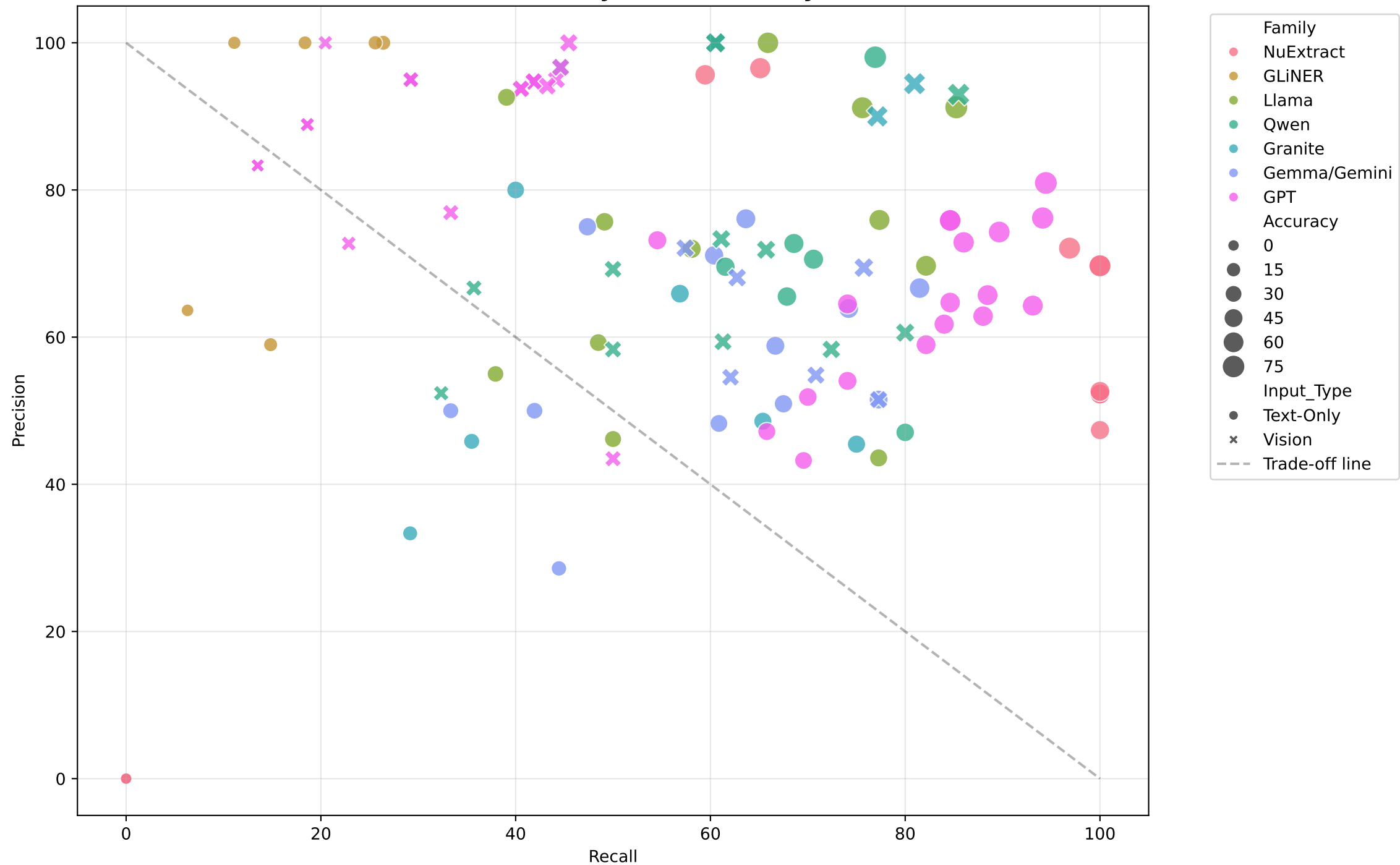
### F1 Score vs Parameter Size by Family and Input Type



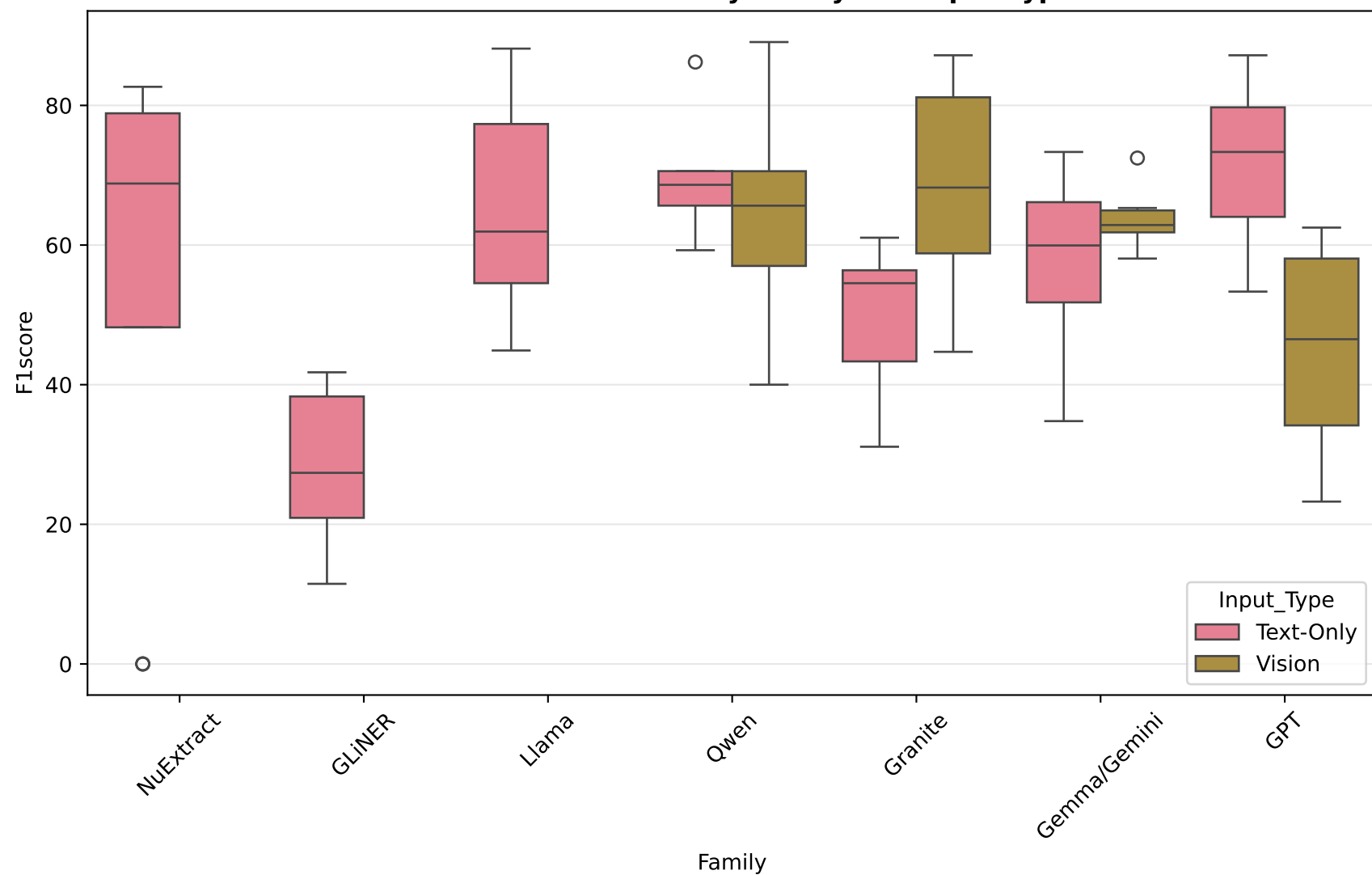
### Accuracy vs Parameter Size by Family and Input Type



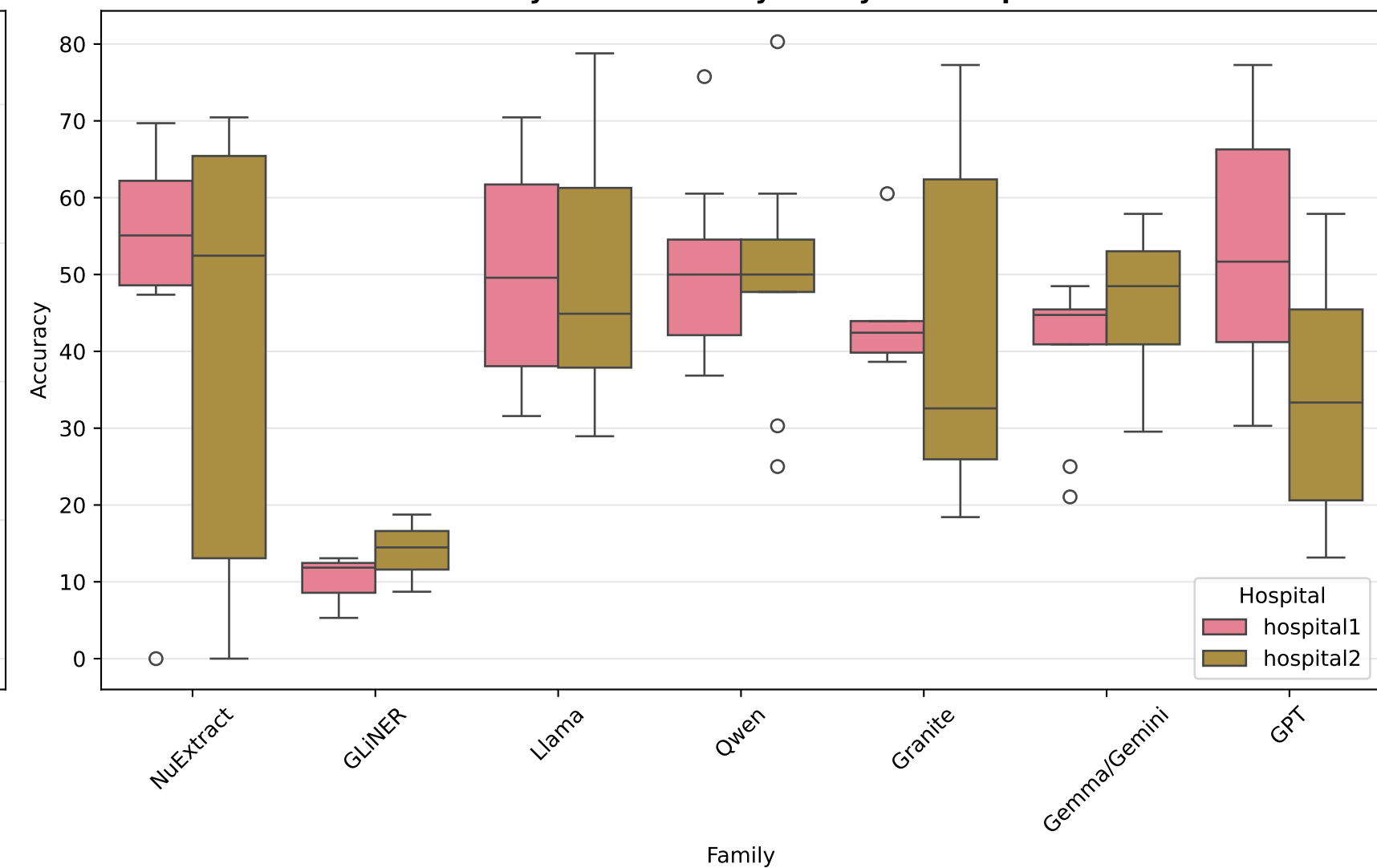
**Precision vs Recall by Family and Input Type**  
(Size = Accuracy, Color = Family)



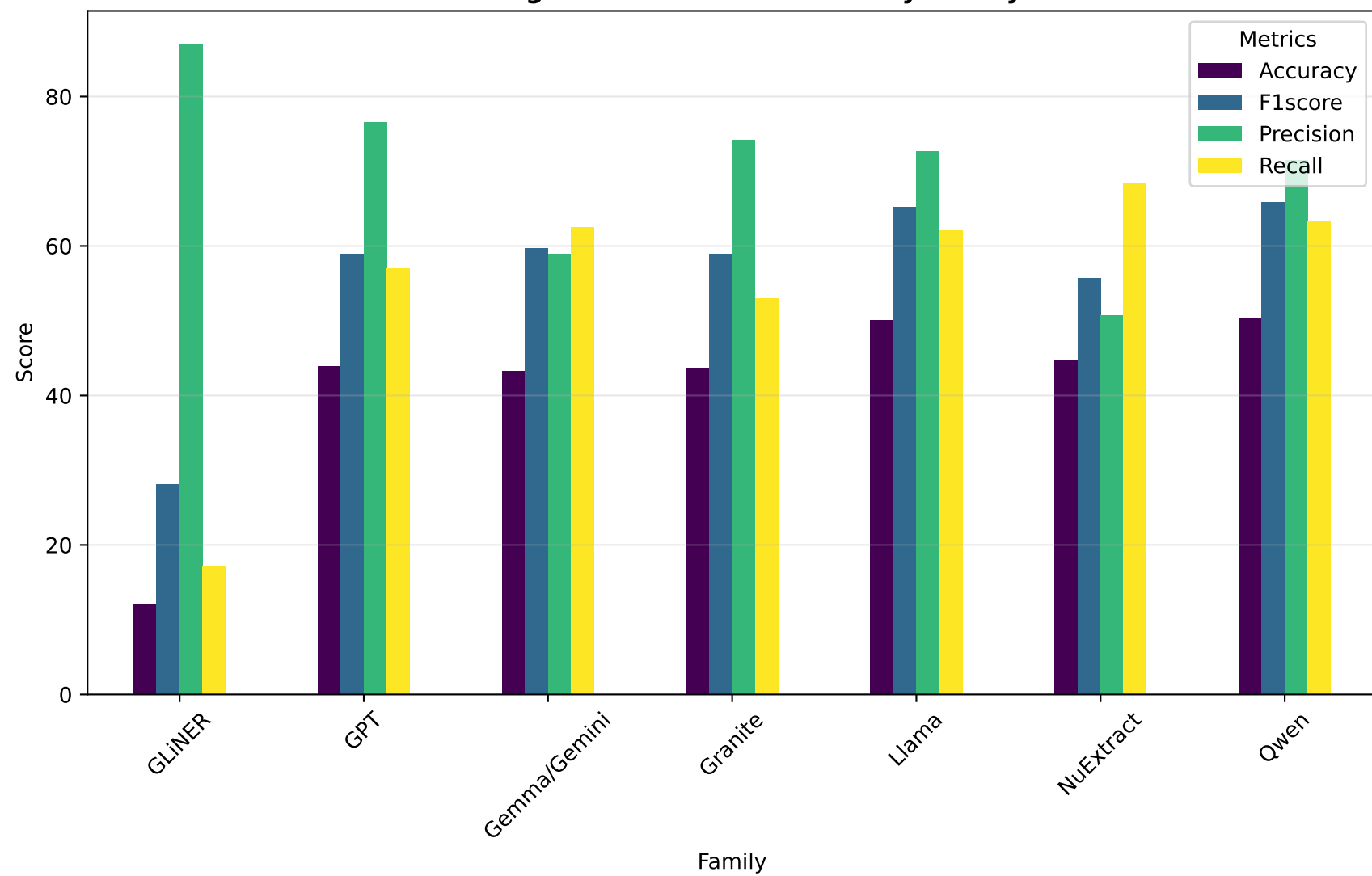
F1 Score Distribution by Family and Input Type



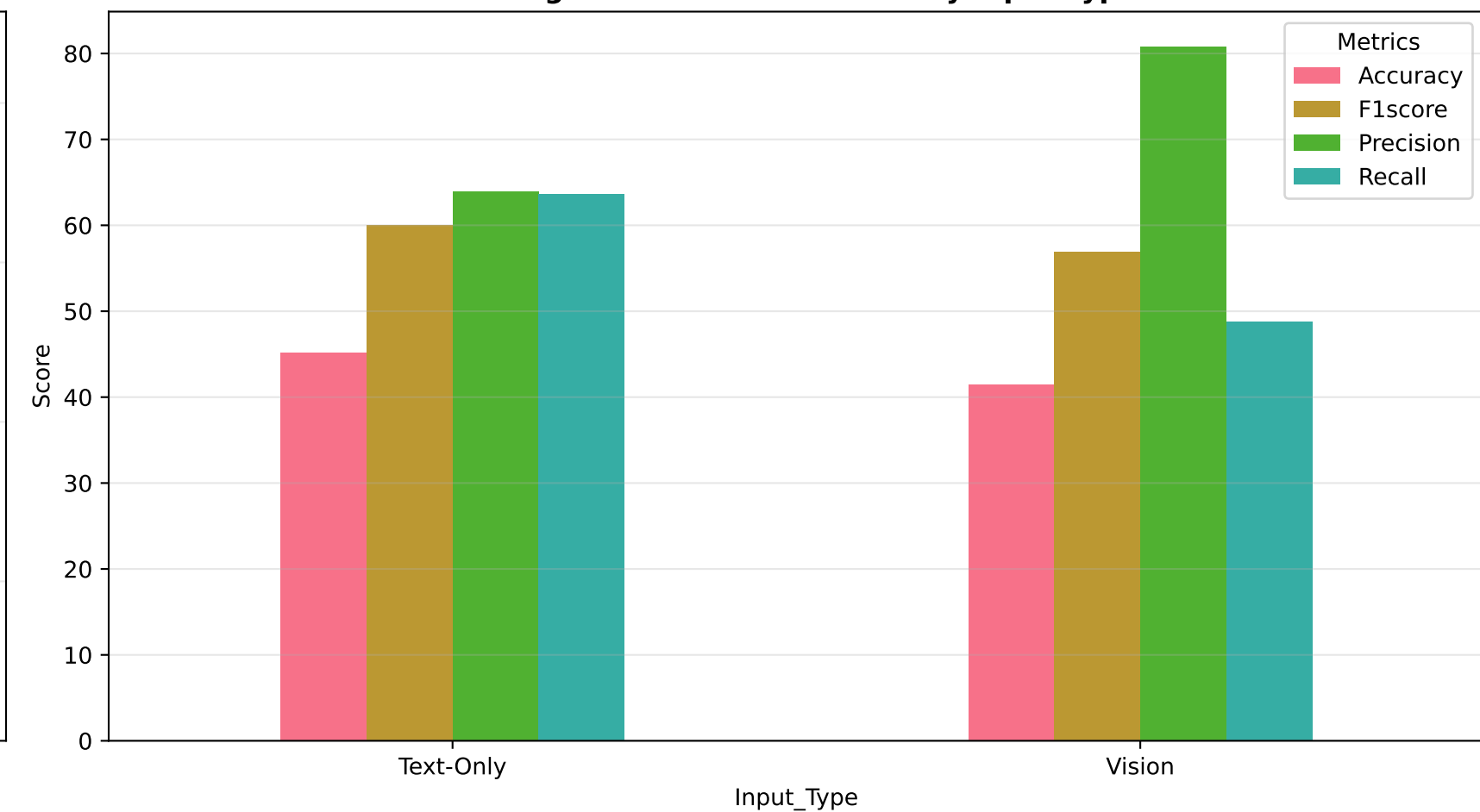
Accuracy Distribution by Family and Hospital



Average Performance Metrics by Family

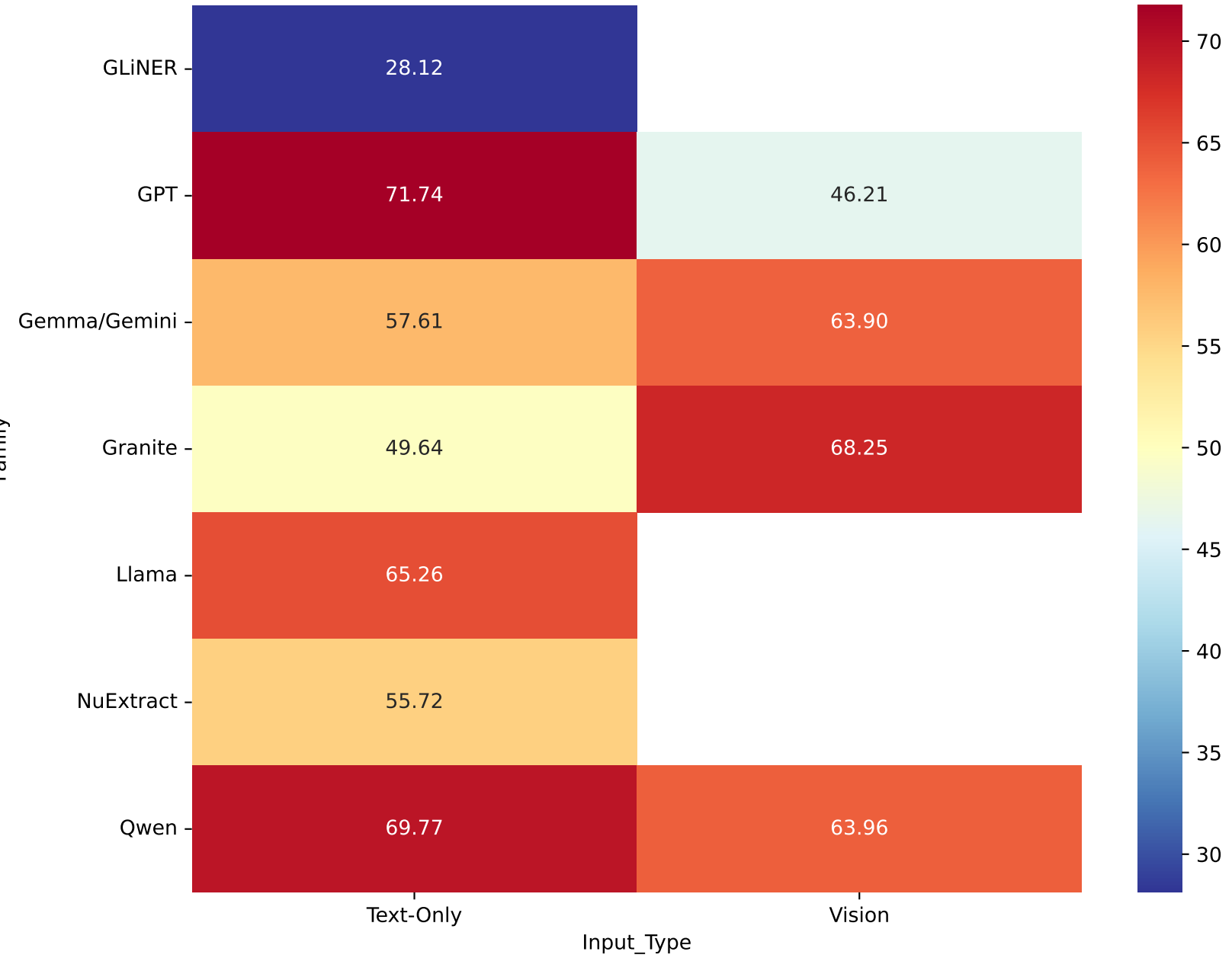


Average Performance Metrics by Input Type

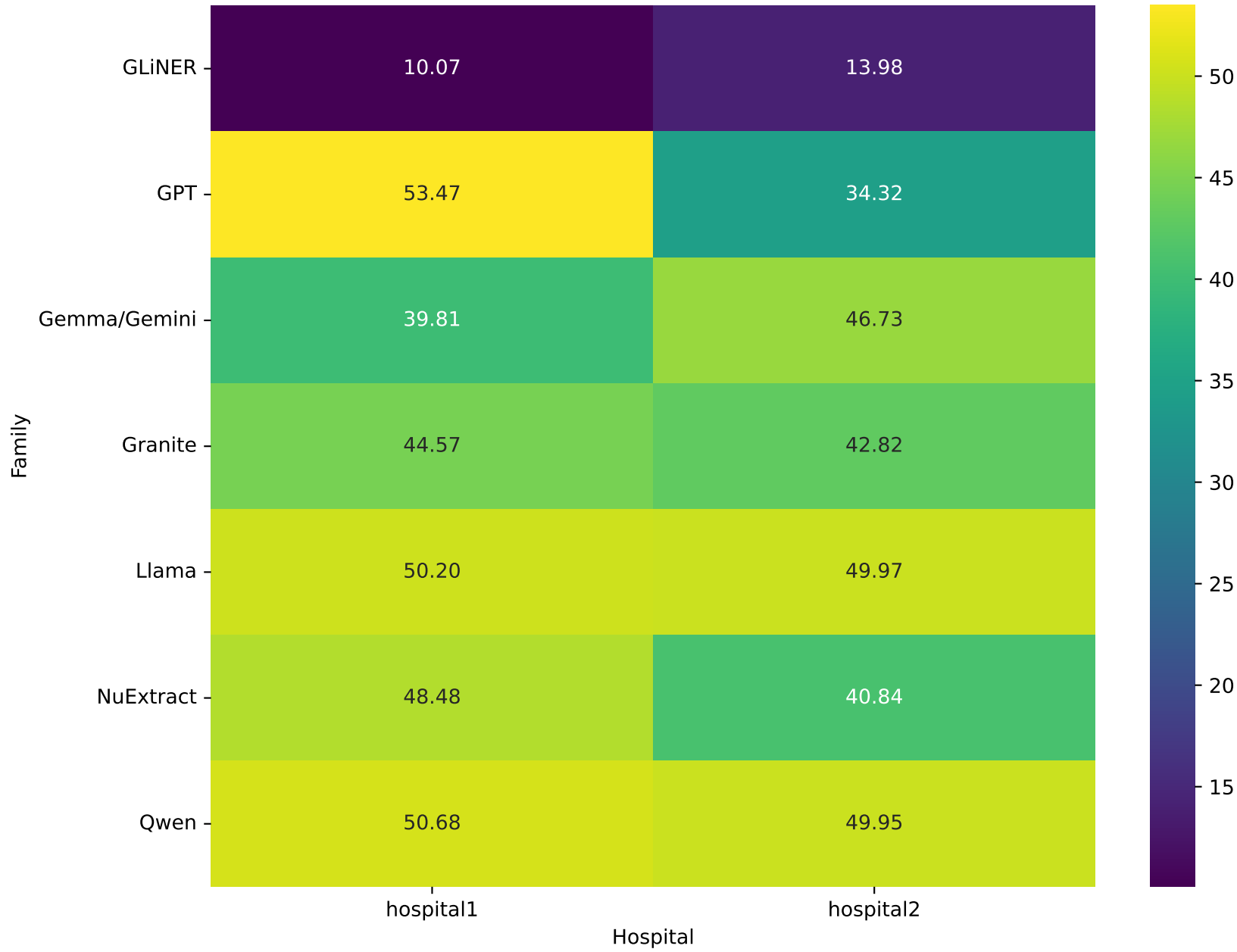




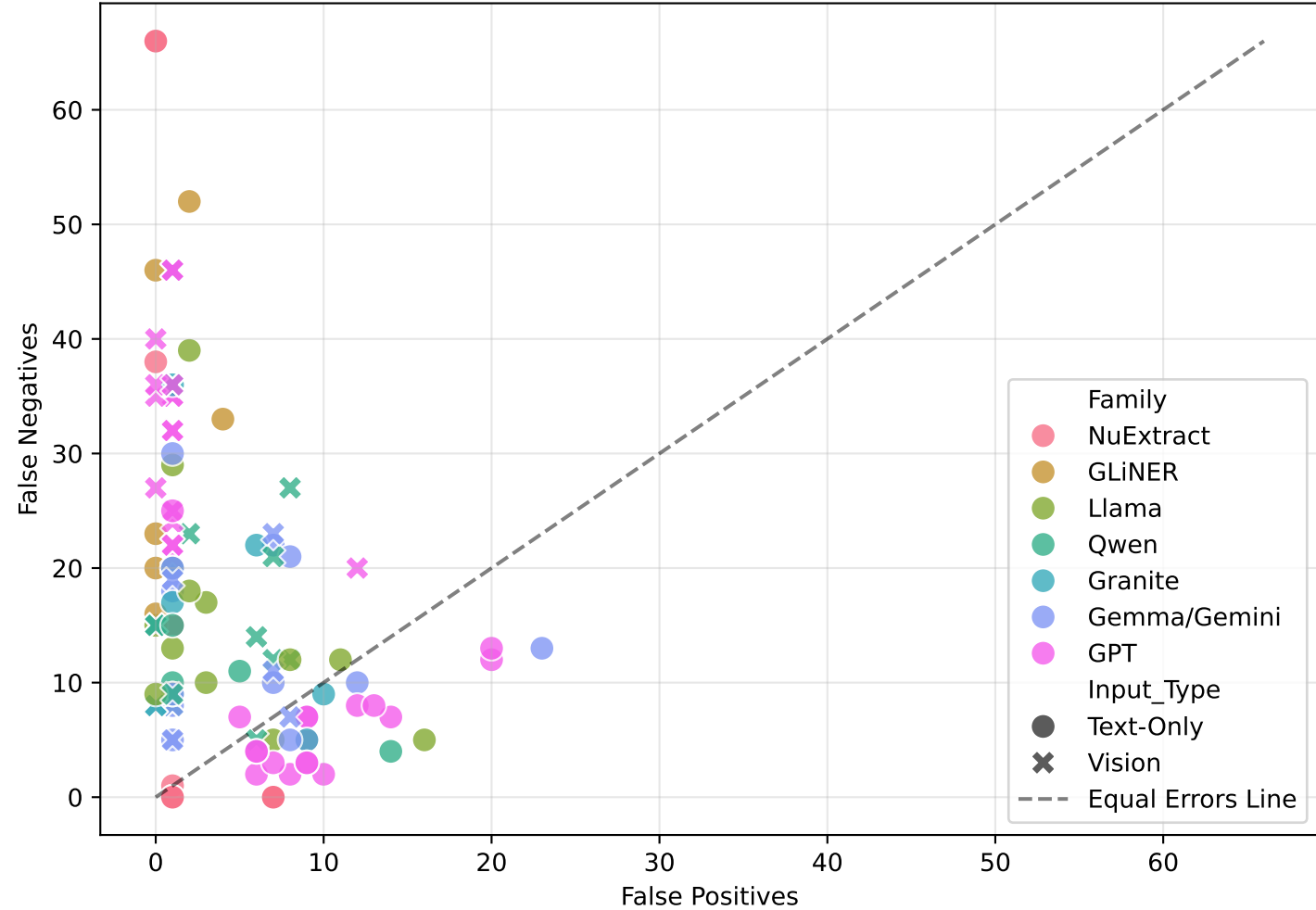
Average F1 Score: Family vs Input Type



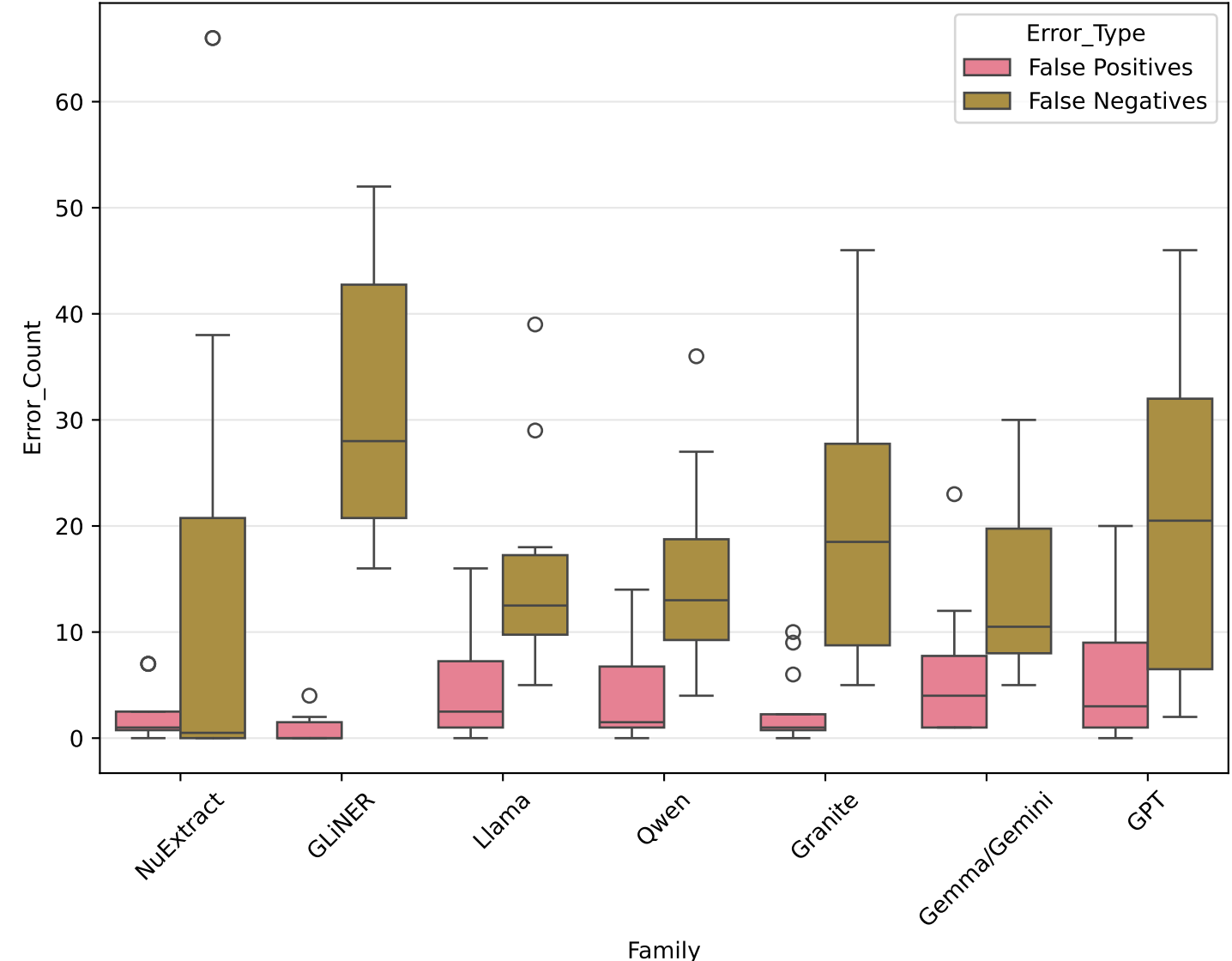
Average Accuracy: Family vs Hospital



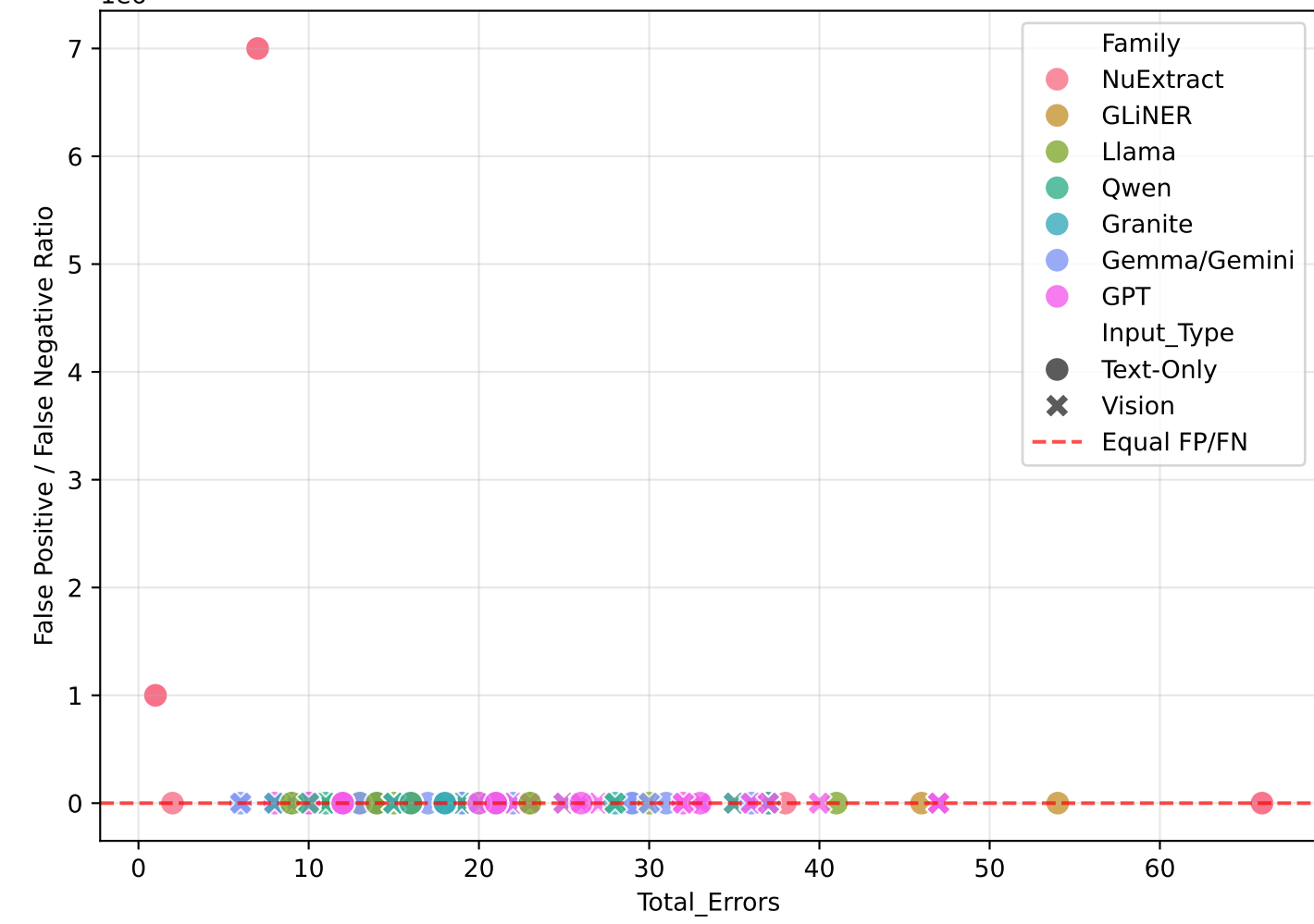
False Positives vs False Negatives by Family and Input Type



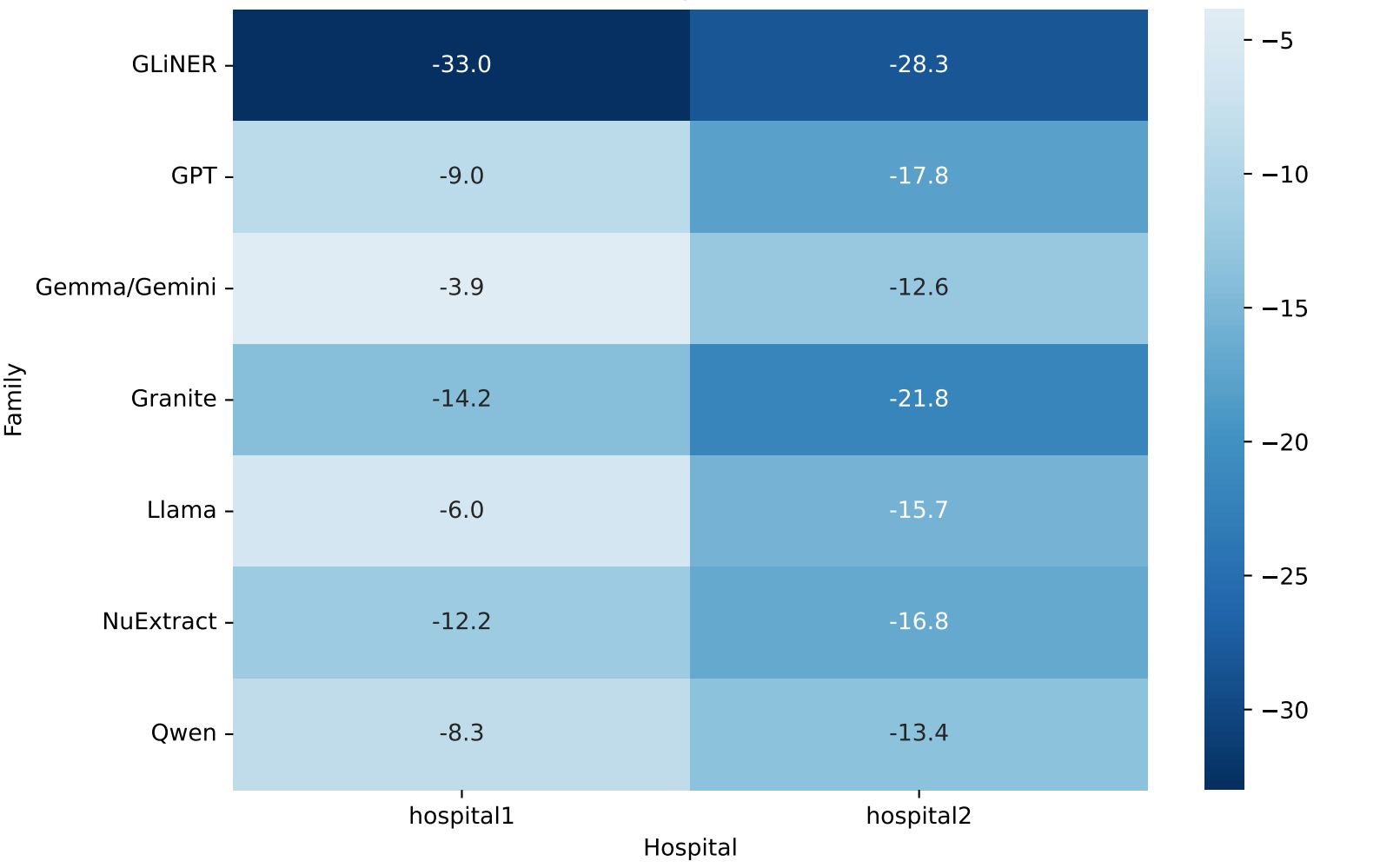
Error Distribution by Model Family



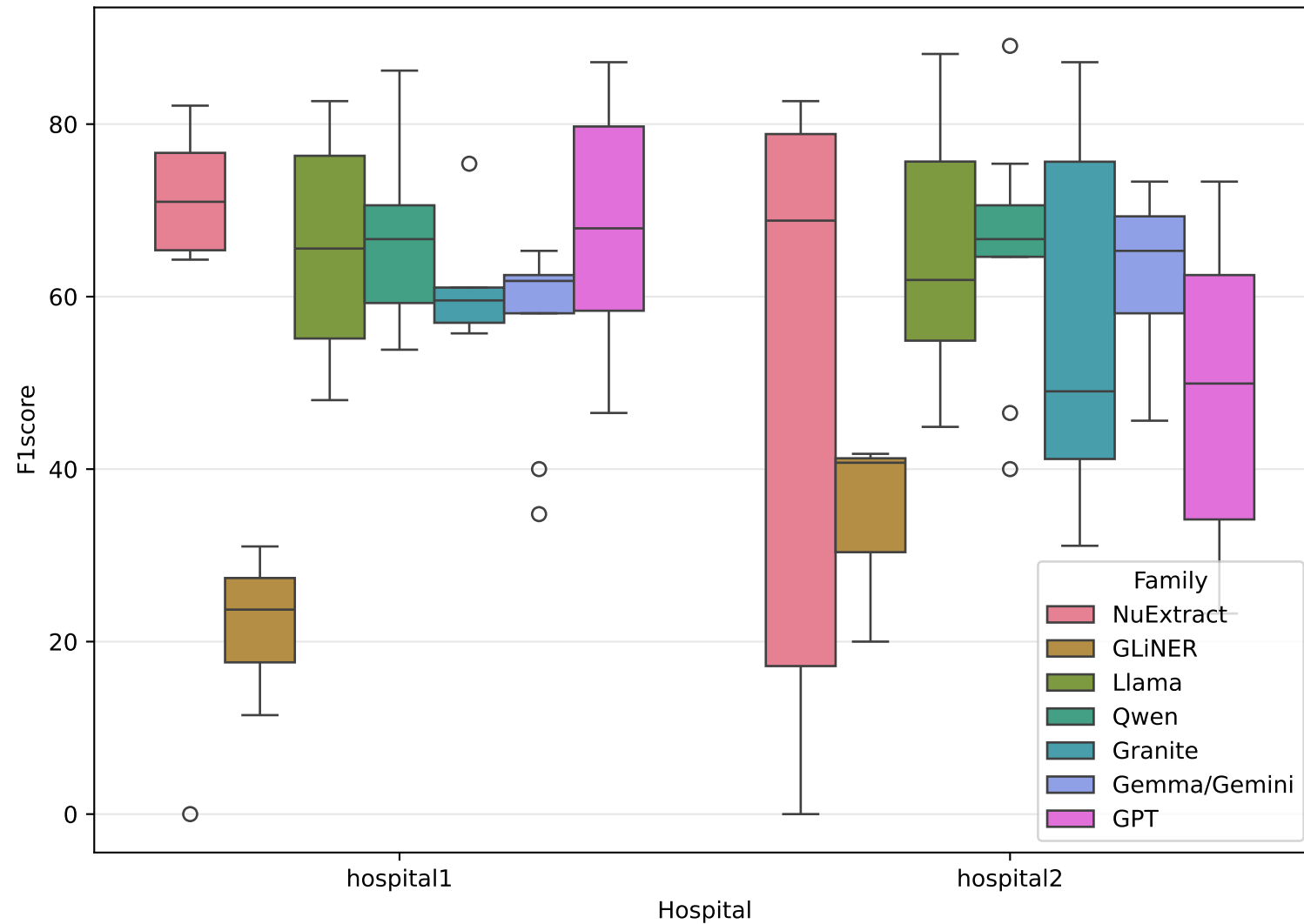
Error Ratio vs Total Errors  
(Ratio = FP/FN)



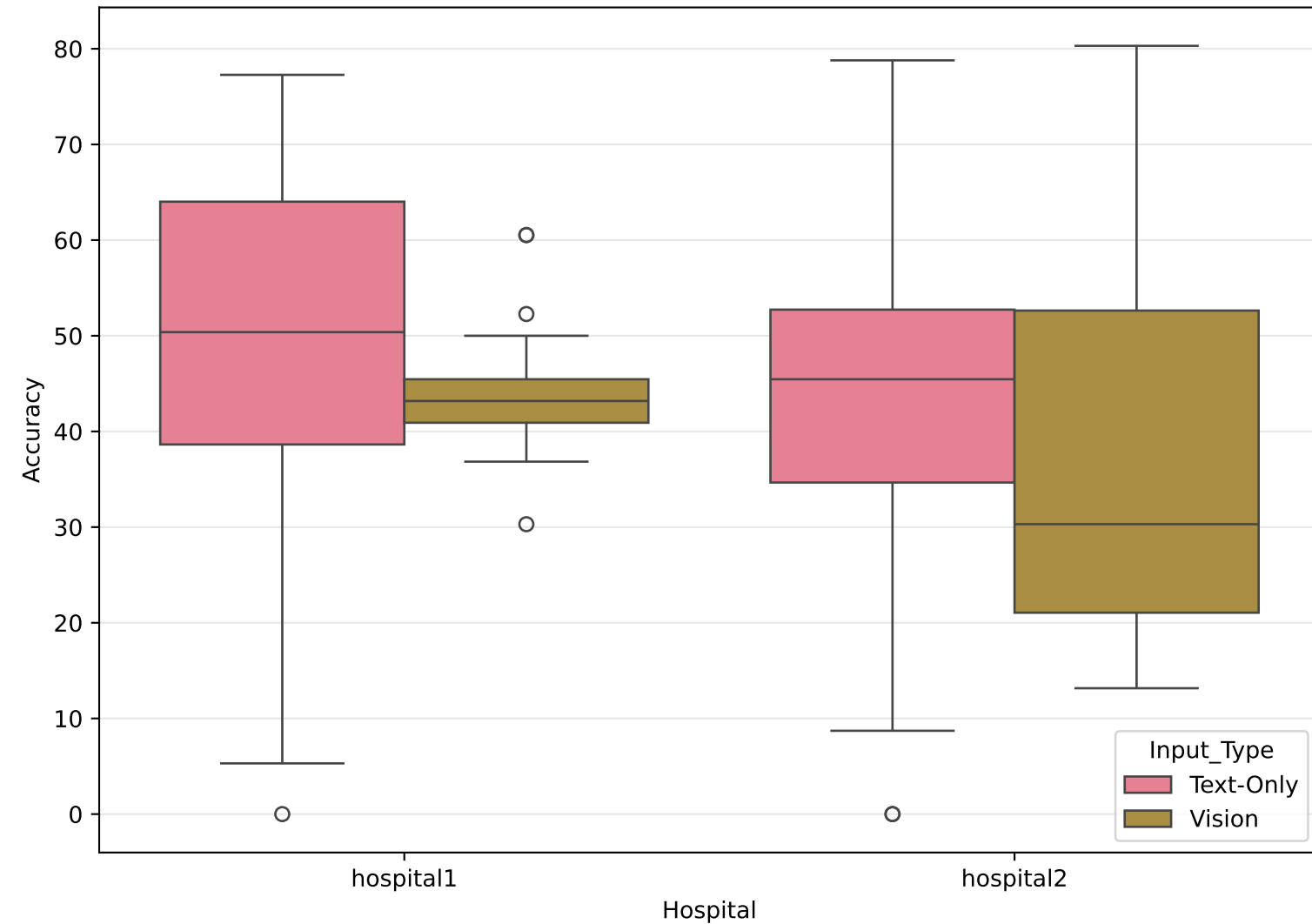
Error Bias: FP - FN by Family and Hospital  
(+ve = More FP, -ve = More FN)



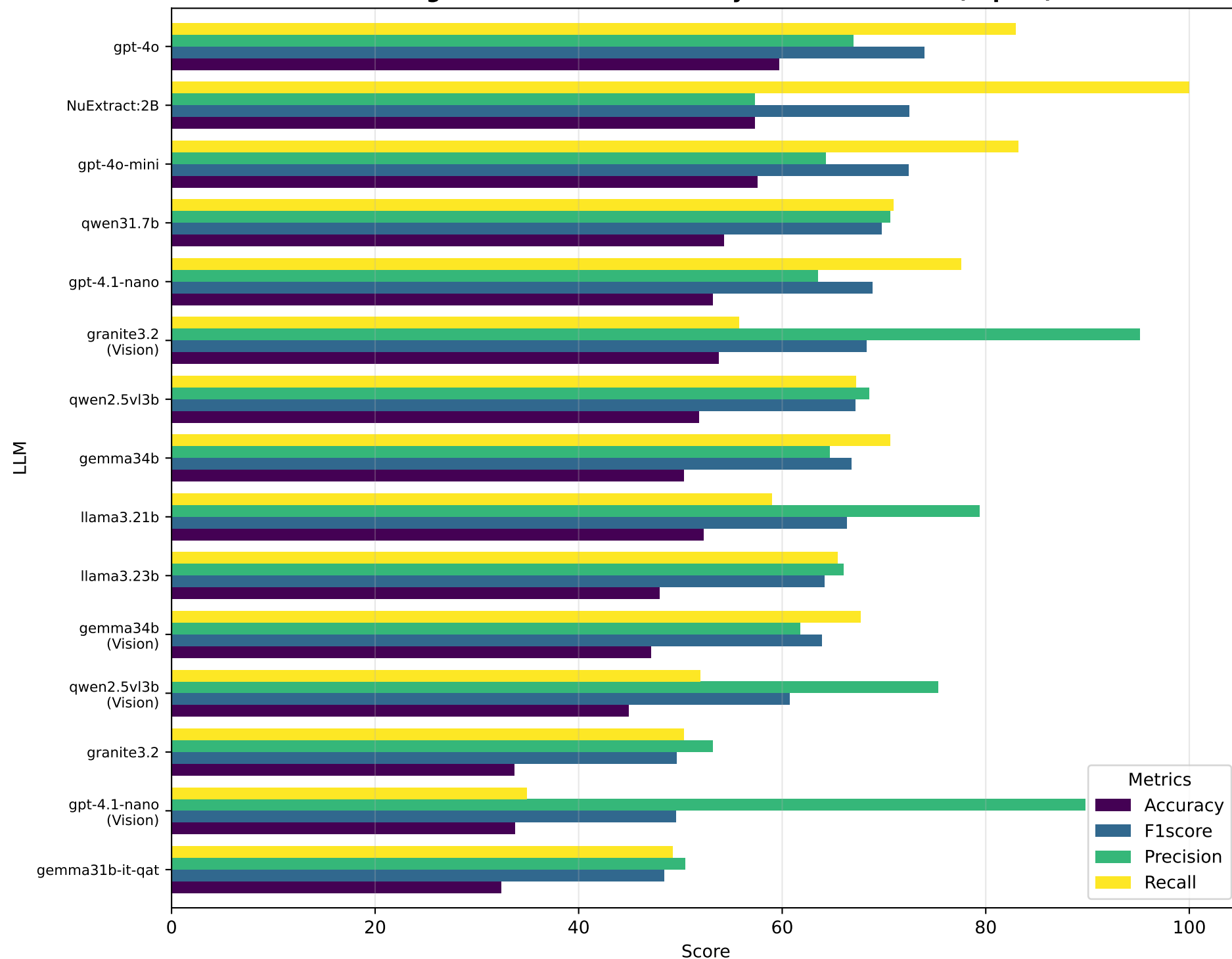
### F1 Score Distribution by Hospital and Family



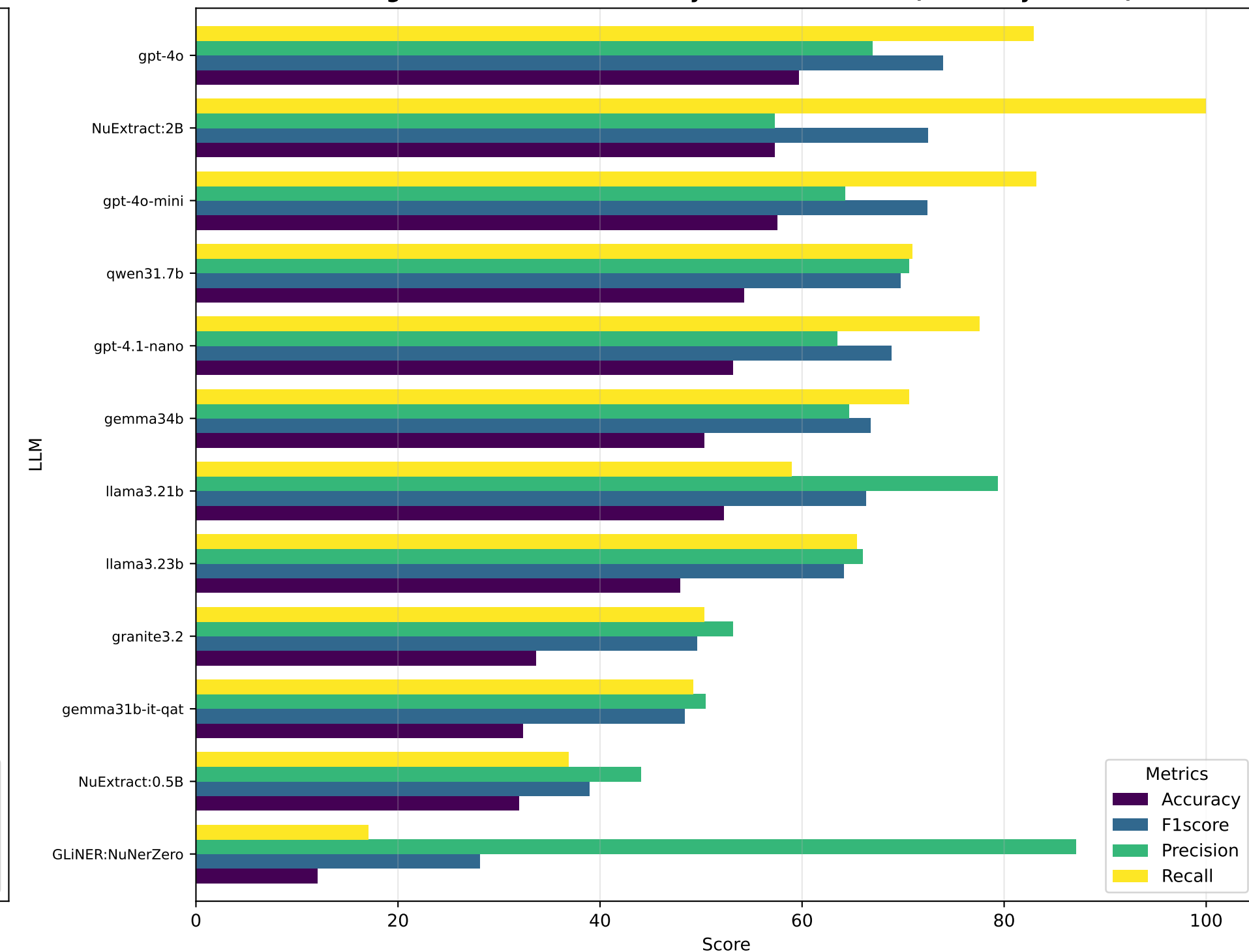
### Accuracy Distribution by Hospital and Input Type



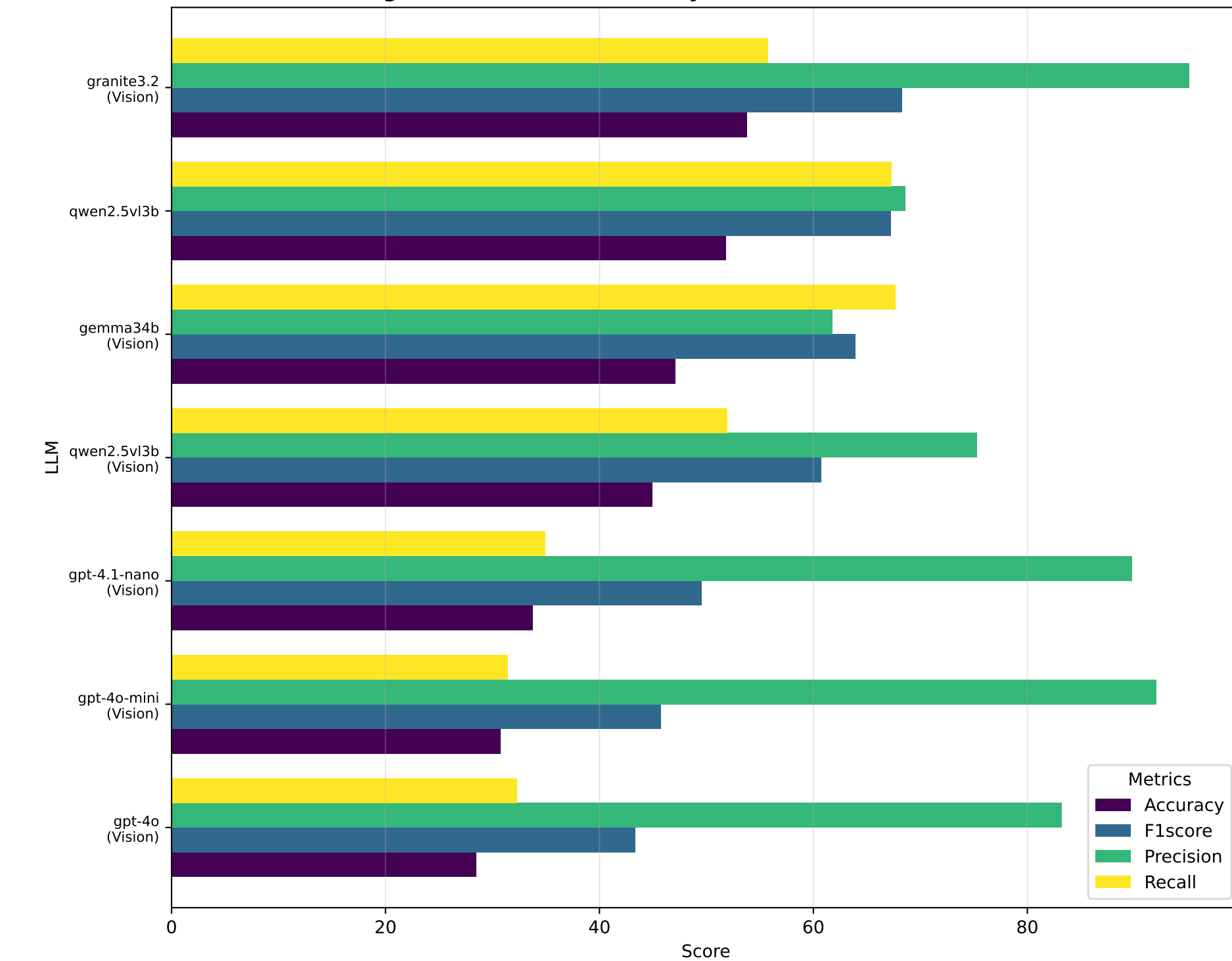
Average Performance Metrics by Individual Model (Top 15)



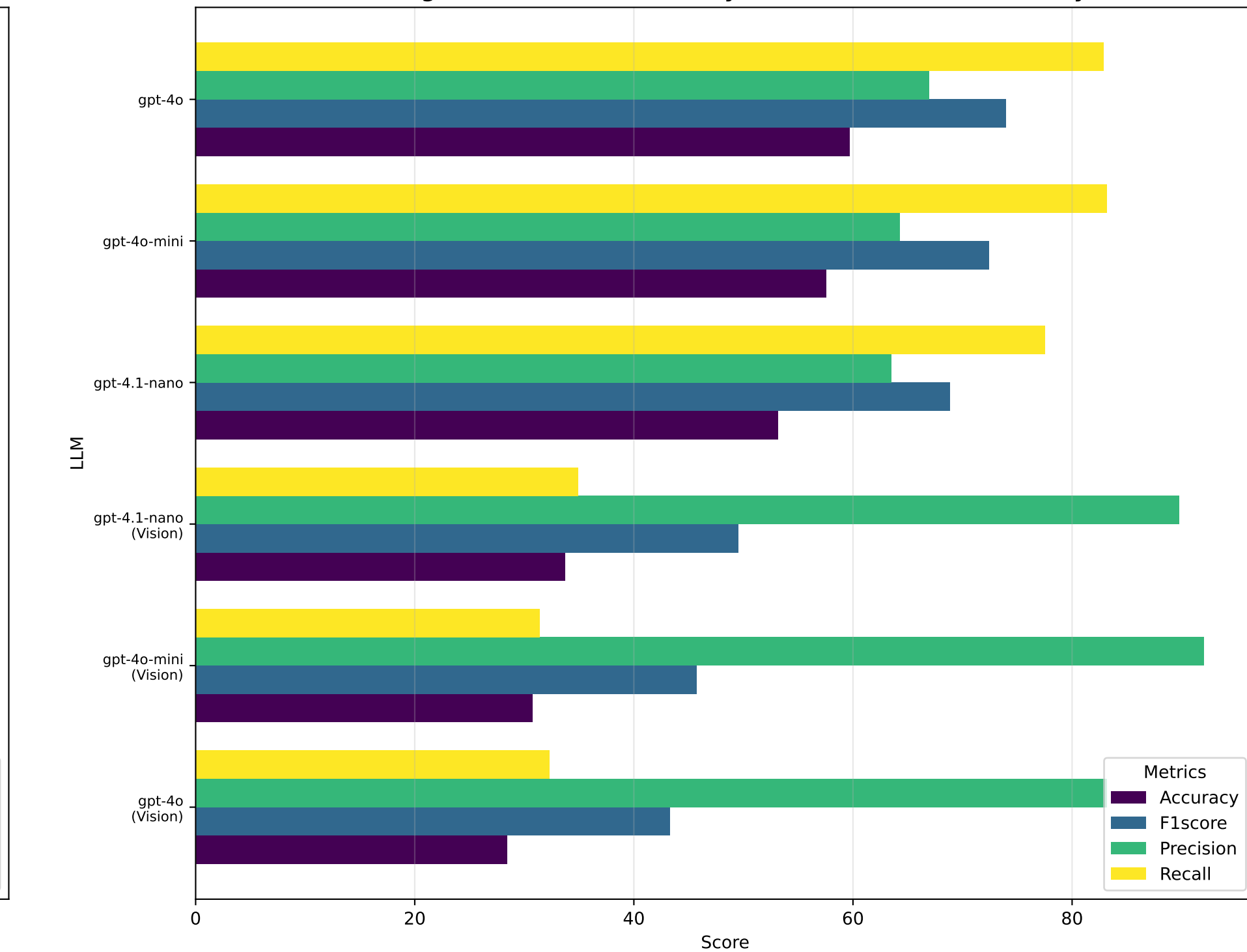
Average Performance Metrics by Individual Model (Text-Only Models)



Average Performance Metrics by Individual Model (Vision Models)



Average Performance Metrics by Individual Model (GPT Family)



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

LLM

