

COMPREHENSIVE LLM BENCHMARKING ANALYSIS - SUMMARY REPORT

DATASET OVERVIEW:

- Total Records: 162
- Model Families: Qwen, Gemma/Gemini, Granite, Llama, GPT, Mistral, Other
- Hospitals: hospital1, hospital2
- Vision-enabled Models: 63
- Text-only Models: 99
- Ollama Models: 91
- Commercial Models: 71

TOP PERFORMERS:

- qwen2.5v3b*ImageInput* (Qwen, 3.0B, Vision)
F1: 98.86, Accuracy: 98.86
- NuExtract:0.5B (Other, 0.5B, Text-only)
F1: 97.62, Accuracy: 97.62
- NuExtract:2B (Other, 2.0B, Text-only)
F1: 97.62, Accuracy: 97.62
- NuExtract:2B (Other, 2.0B, Text-only)
F1: 97.62, Accuracy: 97.62
- NuExtract:0.5B (Other, 0.5B, Text-only)
F1: 97.62, Accuracy: 97.62

FAMILY PERFORMANCE RANKING (by F1 Score):

- Other: 97.62 (n=8.0)
- Mistral: 82.95 (n=3.0)
- Gemma/Gemini: 61.69 (n=33.0)
- GPT: 59.17 (n=46.0)
- Llama: 58.40 (n=20.0)
- Qwen: 56.76 (n=33.0)
- Granite: 47.82 (n=19.0)

KEY INSIGHTS:

- Vision Models Avg F1: 53.33
- Text-only Models Avg F1: 64.42
- Hospital 1 Avg F1: 61.06
- Hospital 2 Avg F1: 59.09
- Ollama Models Avg F1: 53.22
- Commercial Models Avg F1: 68.94

Grouped Model F1 Score Statistics:

- Unique Base Models: 18
- Total Test Instances: 162
- Best Performing Model: NuExtract:0.5B (F1: 97.62)
- Worst Performing Model: granite3.32b (F1: 37.28)
- Overall Average F1: 66.10
- Models with Vision: 6

Top 5 Performers:

- devstral-small (Mistral, Text-only): F1 = 85.69 ± 6.34
- gemini-2.0 (Gemma/Gemini, Text-only): F1 = 87.53 ± 3.53
- qwen2.5.72b (Qwen, Text-only): F1 = 89.61 ± 5.11
- NuExtract:2B (Other, Text-only): F1 = 97.62 ± 0.00
- NuExtract:0.5B (Other, Text-only): F1 = 97.62 ± 0.00

Bottom 5 Performers:

- granite3.32b (Granite, Text-only): F1 = 37.28 ± 41.09
- qwen31.7b (Qwen, Text-only): F1 = 43.51 ± 26.89
- llama3.21b (Llama, Text-only): F1 = 46.06 ± 21.61
- gpt-4o (GPT, with Vision): F1 = 49.58 ± 23.24
- granite3.2 (Granite, with Vision): F1 = 49.80 ± 18.49

Error Analysis Summary:

- Average False Positives: 14.77
- Average False Negatives: 25.19
- Models with more FP than FN: 27
- Models with more FN than FP: 114

SOURCE CATEGORY DETAILED STATISTICS:

COMMERCIAL MODELS:

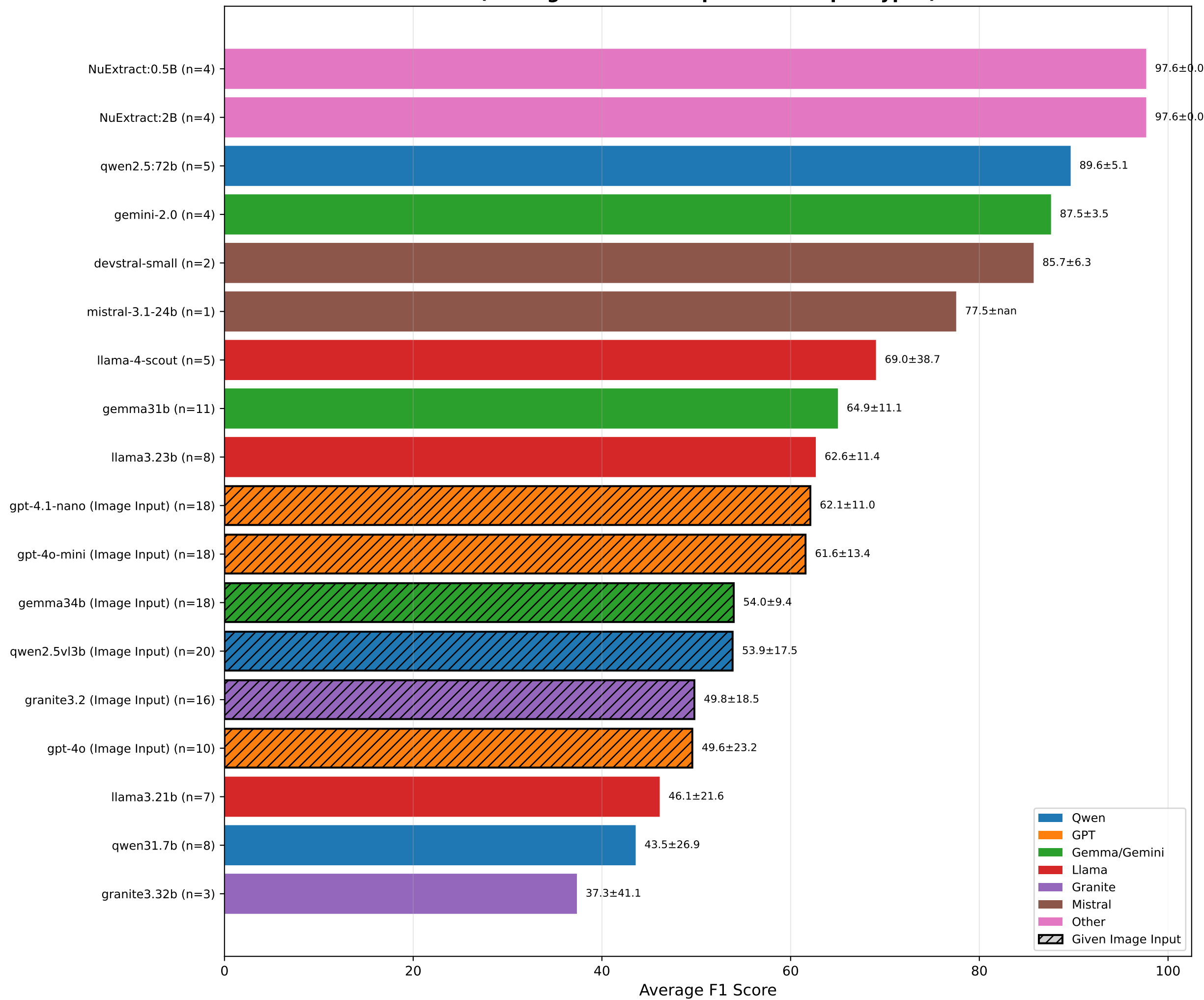
- GPT:
Accuracy: 47.2±14.8 (n=46)
F1 Score: 59.2±15.7
Precision: 75.6±21.1
False Positives: 9.5
False Negatives: 25.6
- Gemma/Gemini:
Accuracy: 82.1±4.9 (n=4)
F1 Score: 87.5±3.5
Precision: 87.0±3.0
False Positives: 10.8
False Negatives: 9.8
- Llama:
Accuracy: 61.7±34.8 (n=5)
F1 Score: 69.0±38.7
Precision: 71.0±39.9
False Positives: 23.8
False Negatives: 27.2
- Mistral:
Accuracy: 77.0±11.3 (n=3)
F1 Score: 83.0±6.5
Precision: 80.6±8.8
False Positives: 14.3
False Negatives: 10.0
- Other:
Accuracy: 97.6±0.0 (n=8)
F1 Score: 97.6±0.0
Precision: 97.6±0.0
False Positives: 1.0
False Negatives: 1.0
- Qwen:
Accuracy: 83.2±9.0 (n=5)
F1 Score: 89.6±5.1
Precision: 89.0±5.1
False Positives: 8.2
False Negatives: 7.0

OLLAMA MODELS:

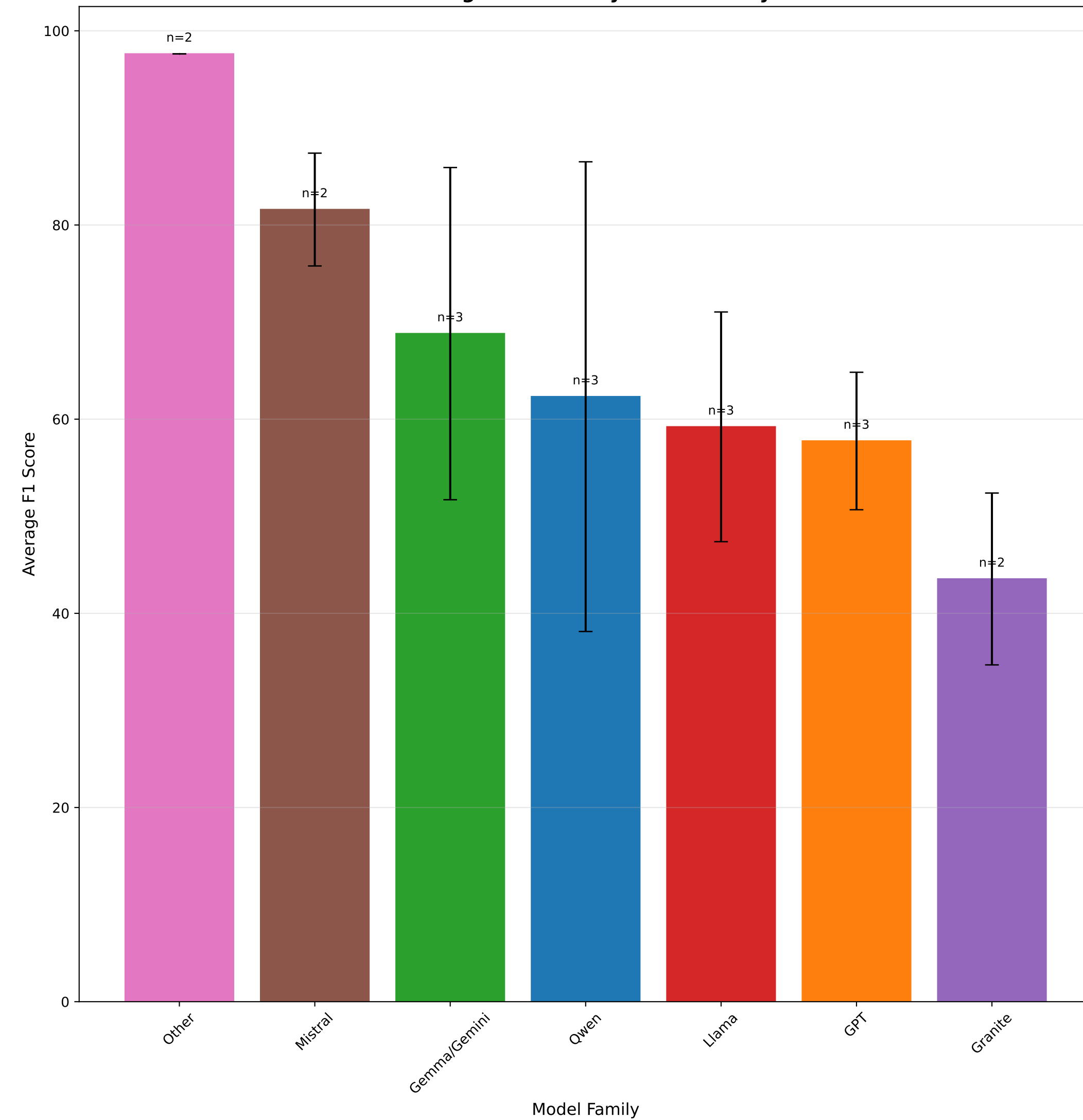
- Gemma/Gemini:
Accuracy: 48.5±11.0 (n=29)
F1 Score: 58.1±11.3
Precision: 63.9±15.4
False Positives: 16.8
False Negatives: 24.3
- Granite:
Accuracy: 35.2±22.9 (n=19)
F1 Score: 47.8±22.2
Precision: 60.6±37.0
False Positives: 17.2
False Negatives: 31.9
- Llama:
Accuracy: 42.7±15.9 (n=15)
F1 Score: 54.9±18.4
Precision: 69.9±22.6
False Positives: 15.0
False Negatives: 31.7
- Qwen:
Accuracy: 42.3±19.1 (n=28)
F1 Score: 50.9±20.6
Precision: 55.8±22.5
False Positives: 23.8
False Negatives: 31.0

=====

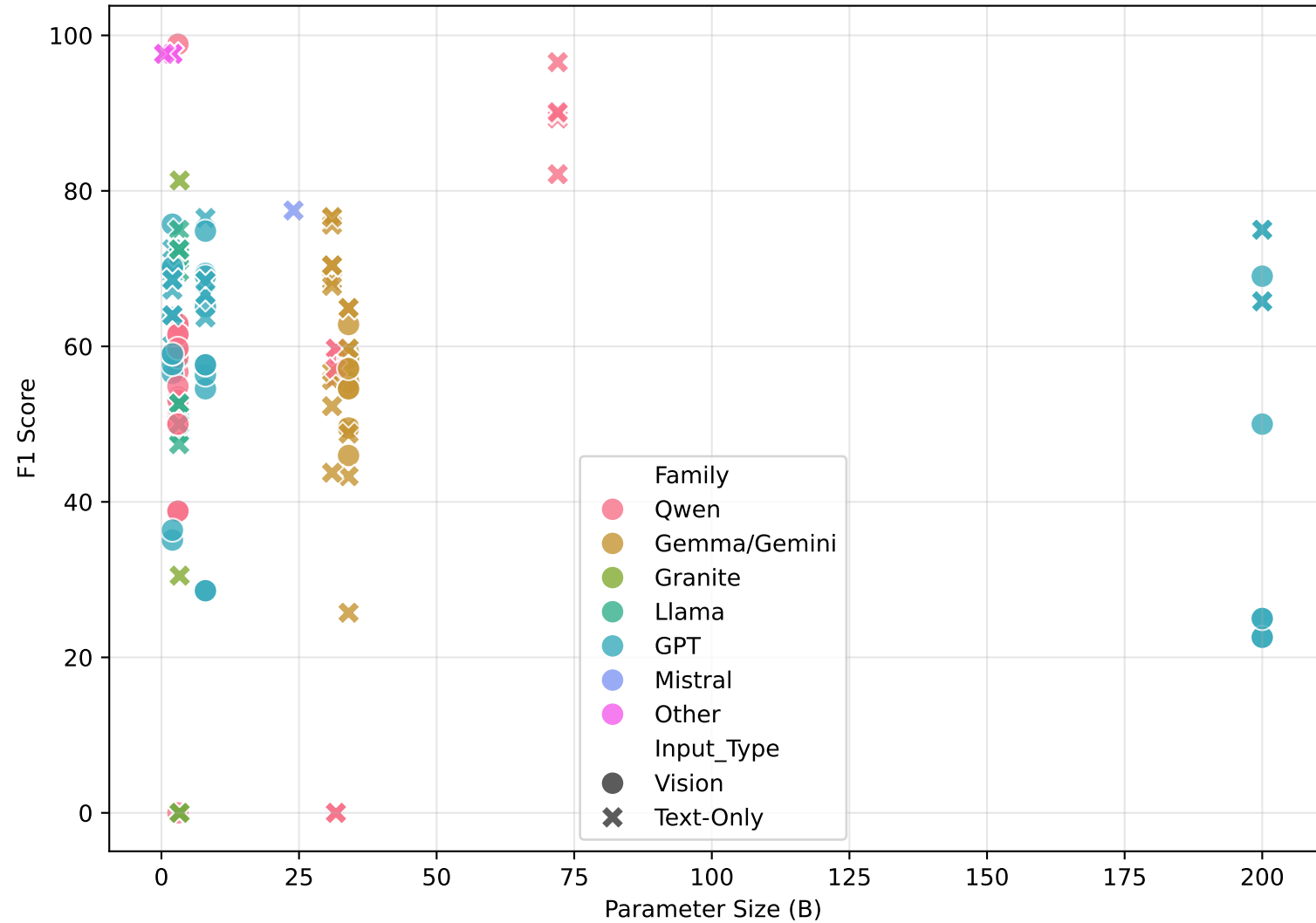
Overall F1 Score Performance - Models Grouped by Base Name
(Averaged across hospitals and input types)



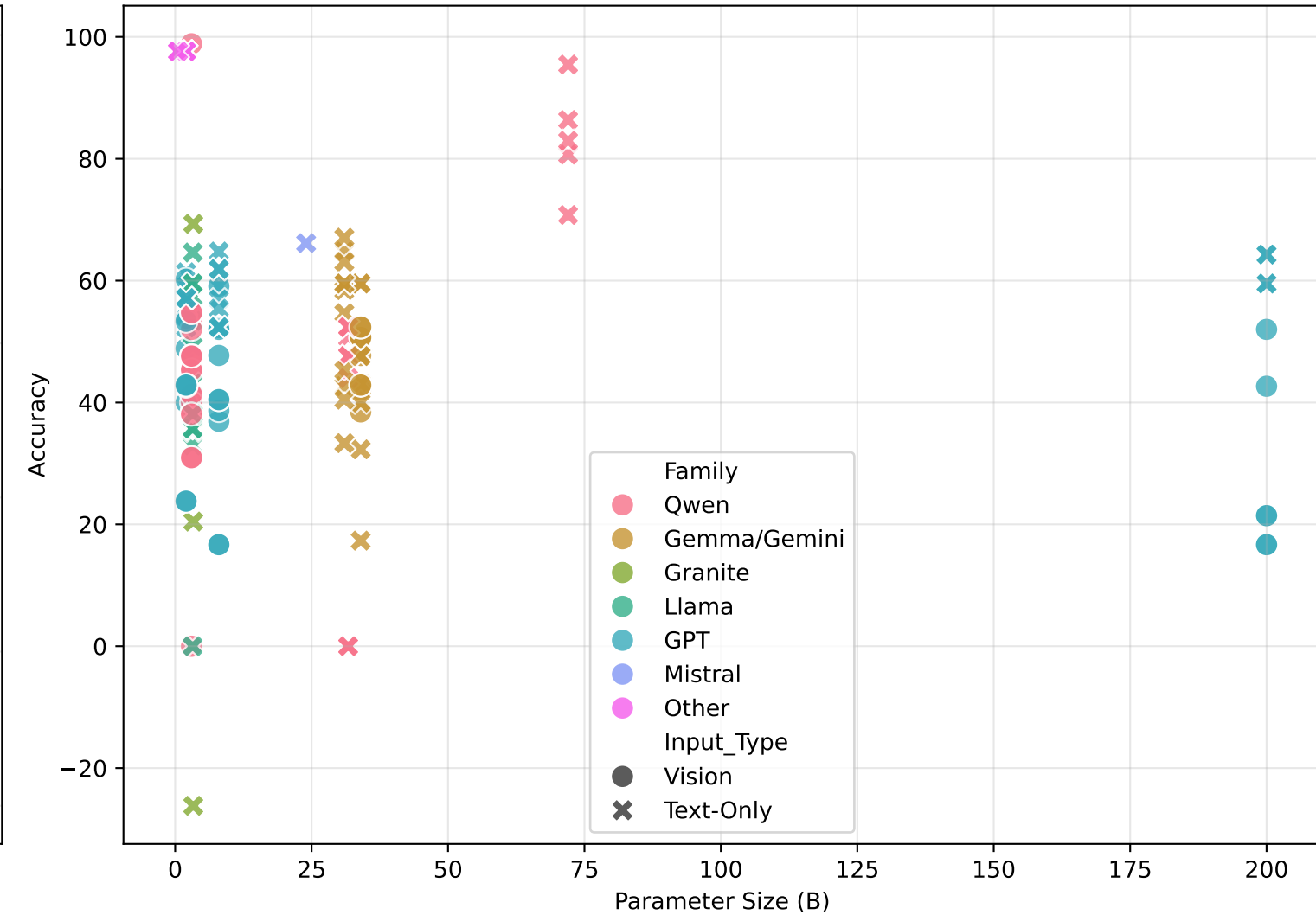
Average F1 Score by Model Family



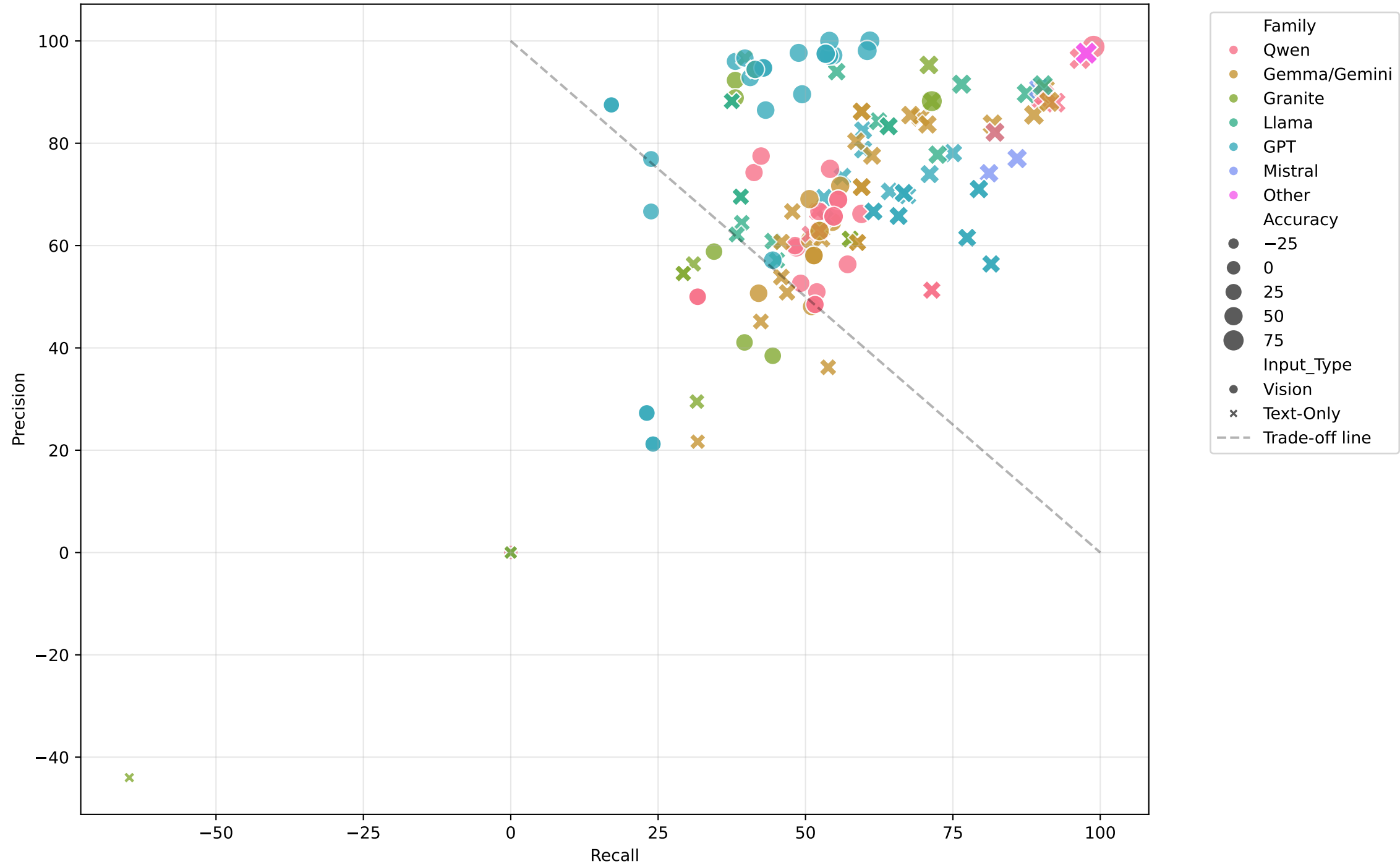
F1 Score vs Parameter Size by Family and Input Type

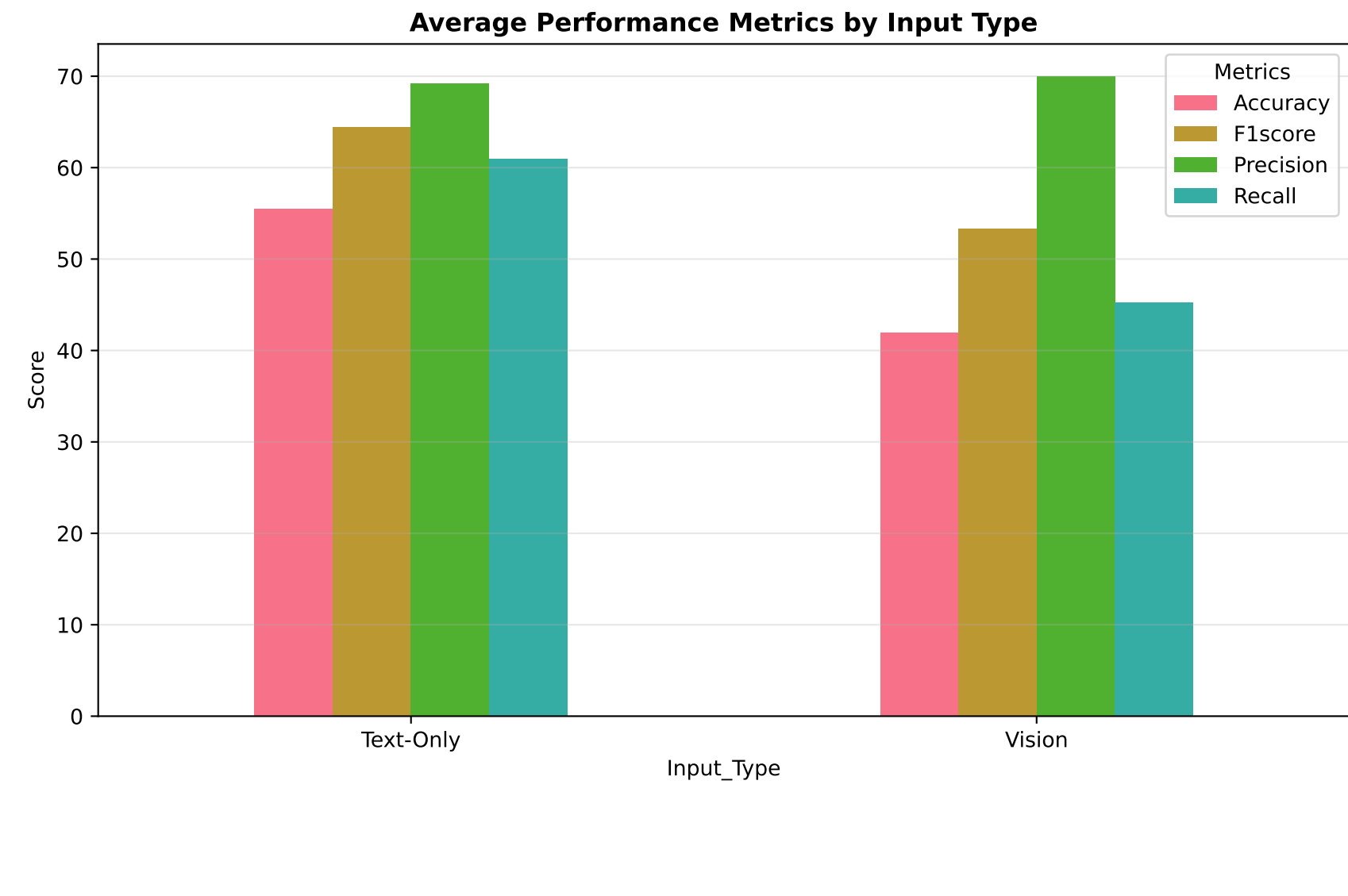
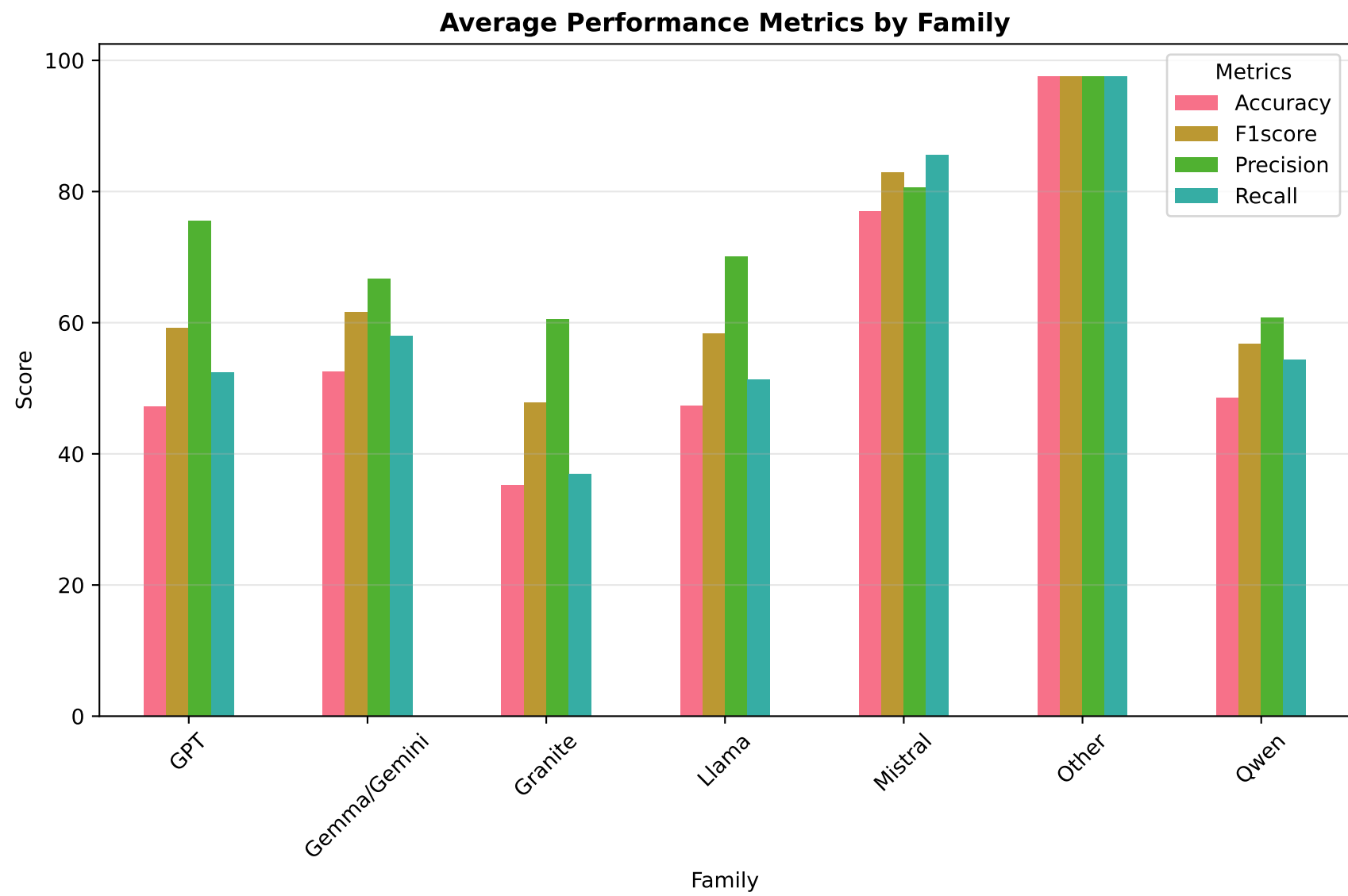
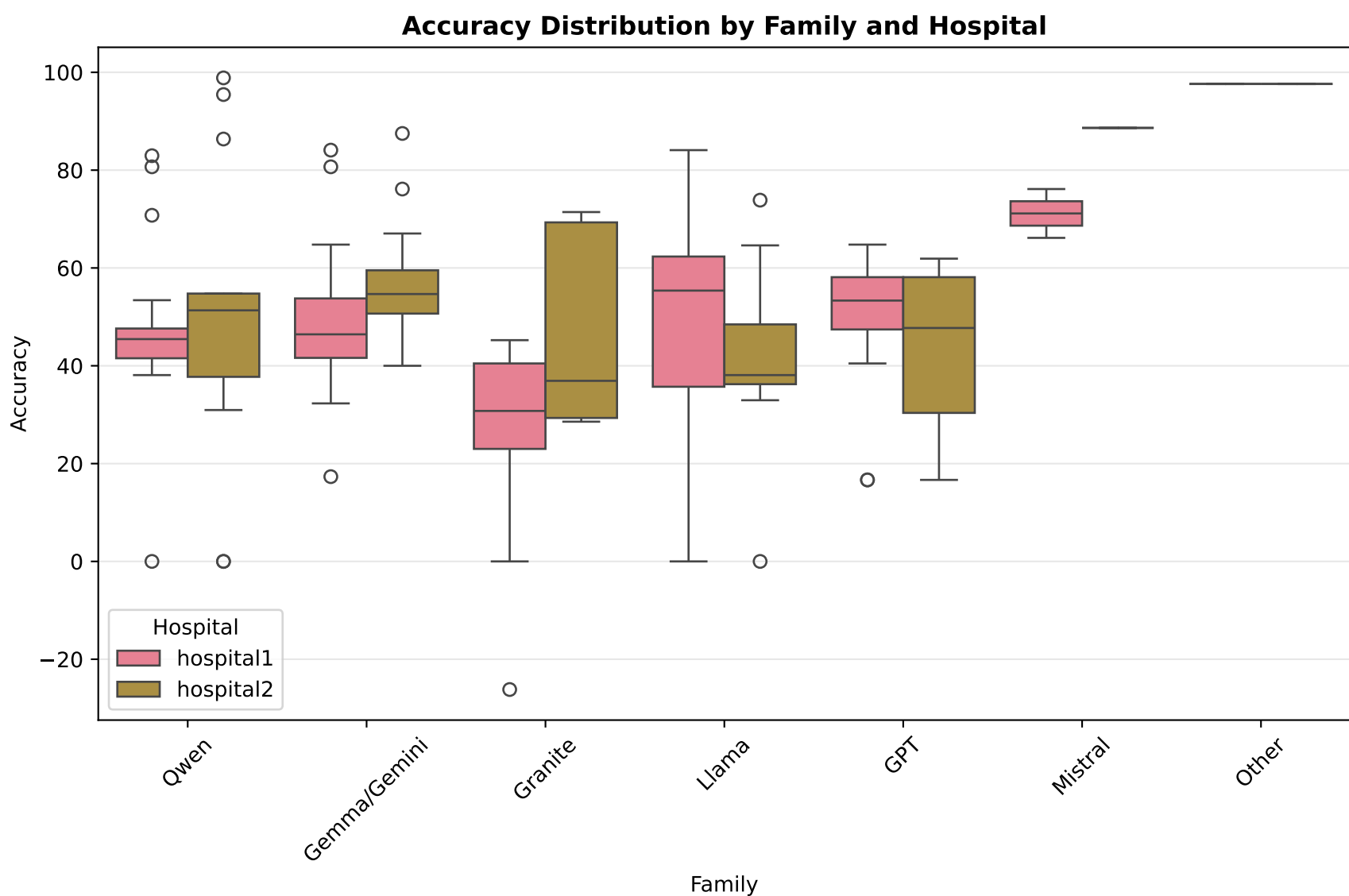
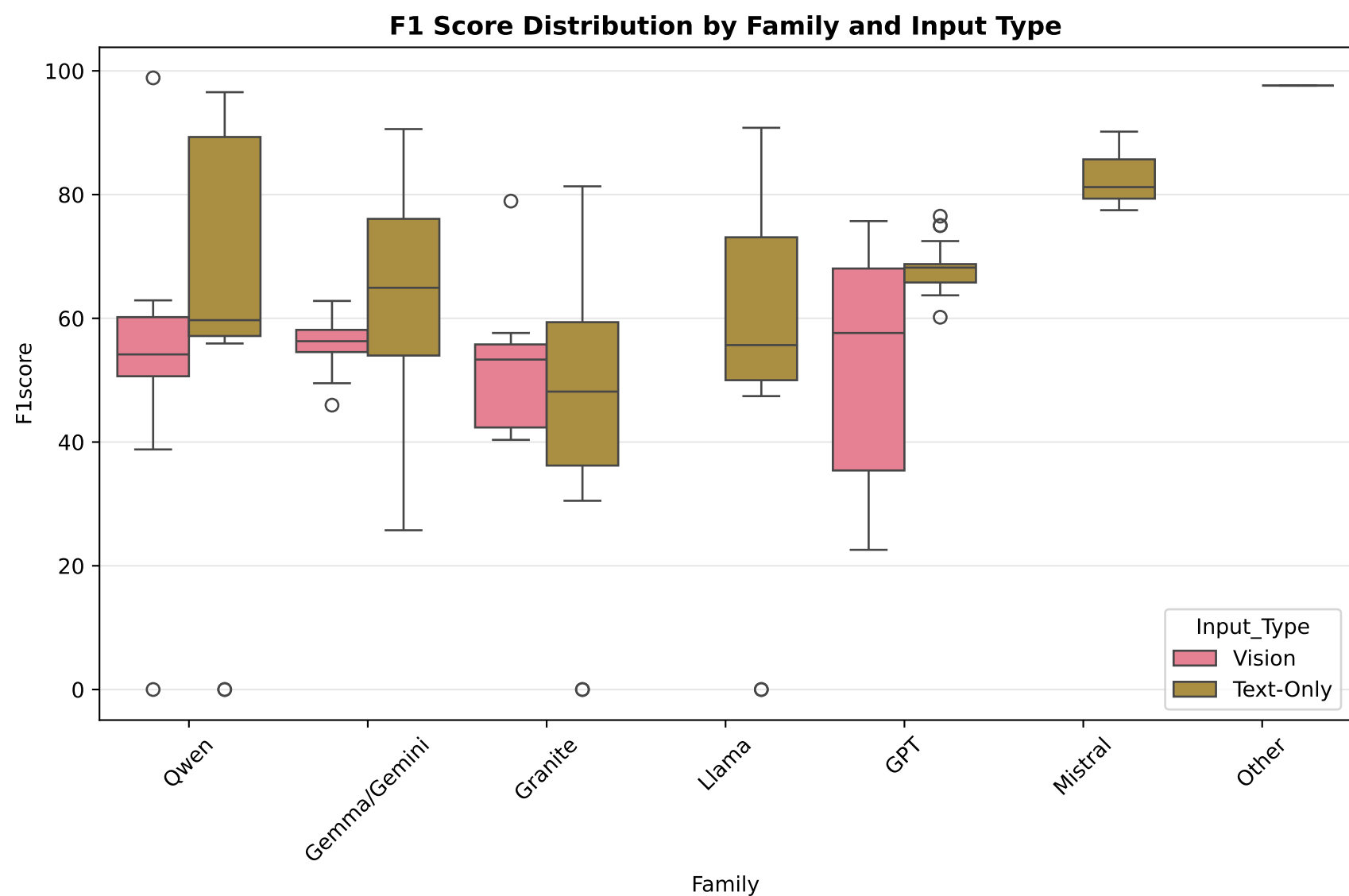


Accuracy vs Parameter Size by Family and Input Type

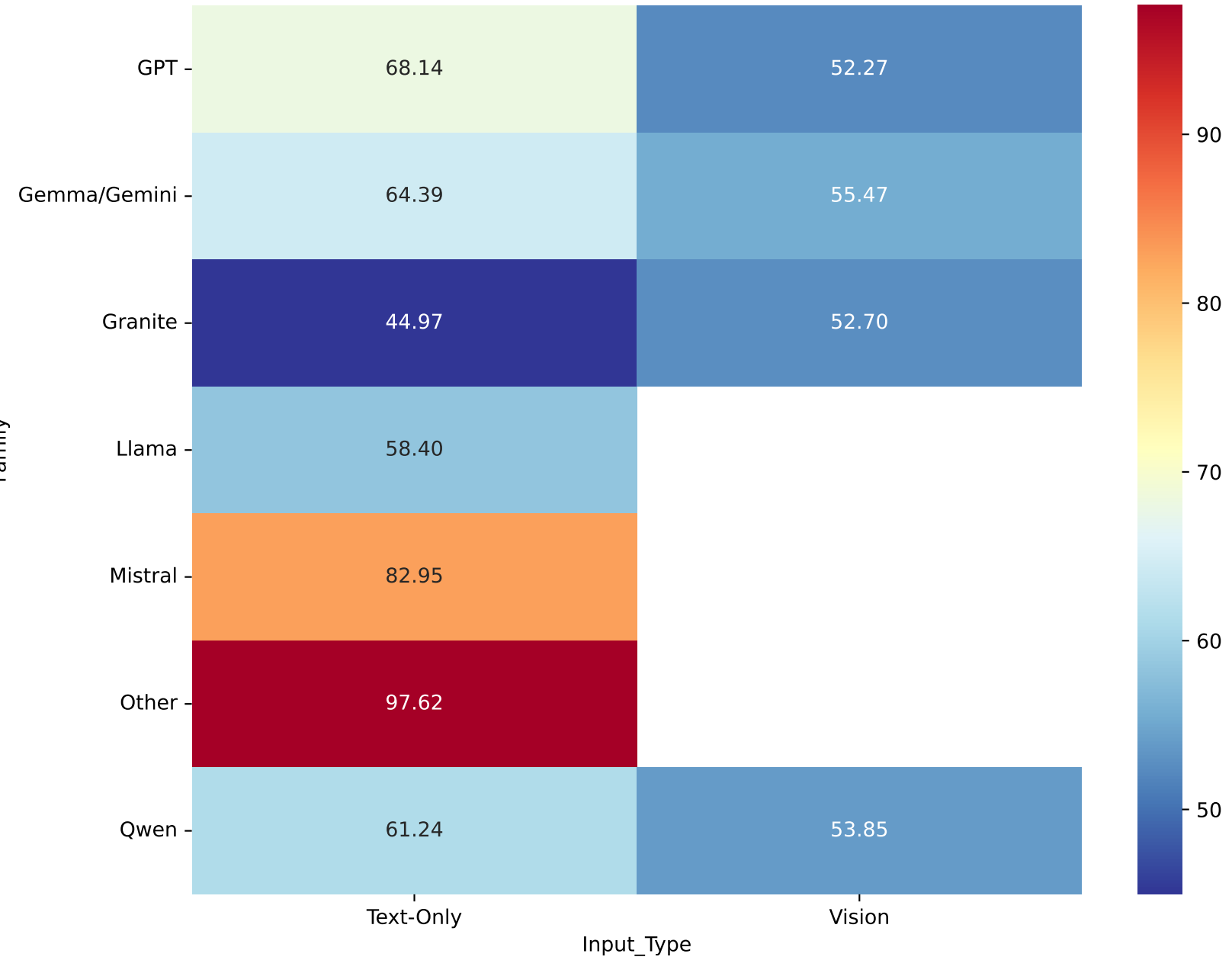


Precision vs Recall by Family and Input Type
(Size = Accuracy, Color = Family)

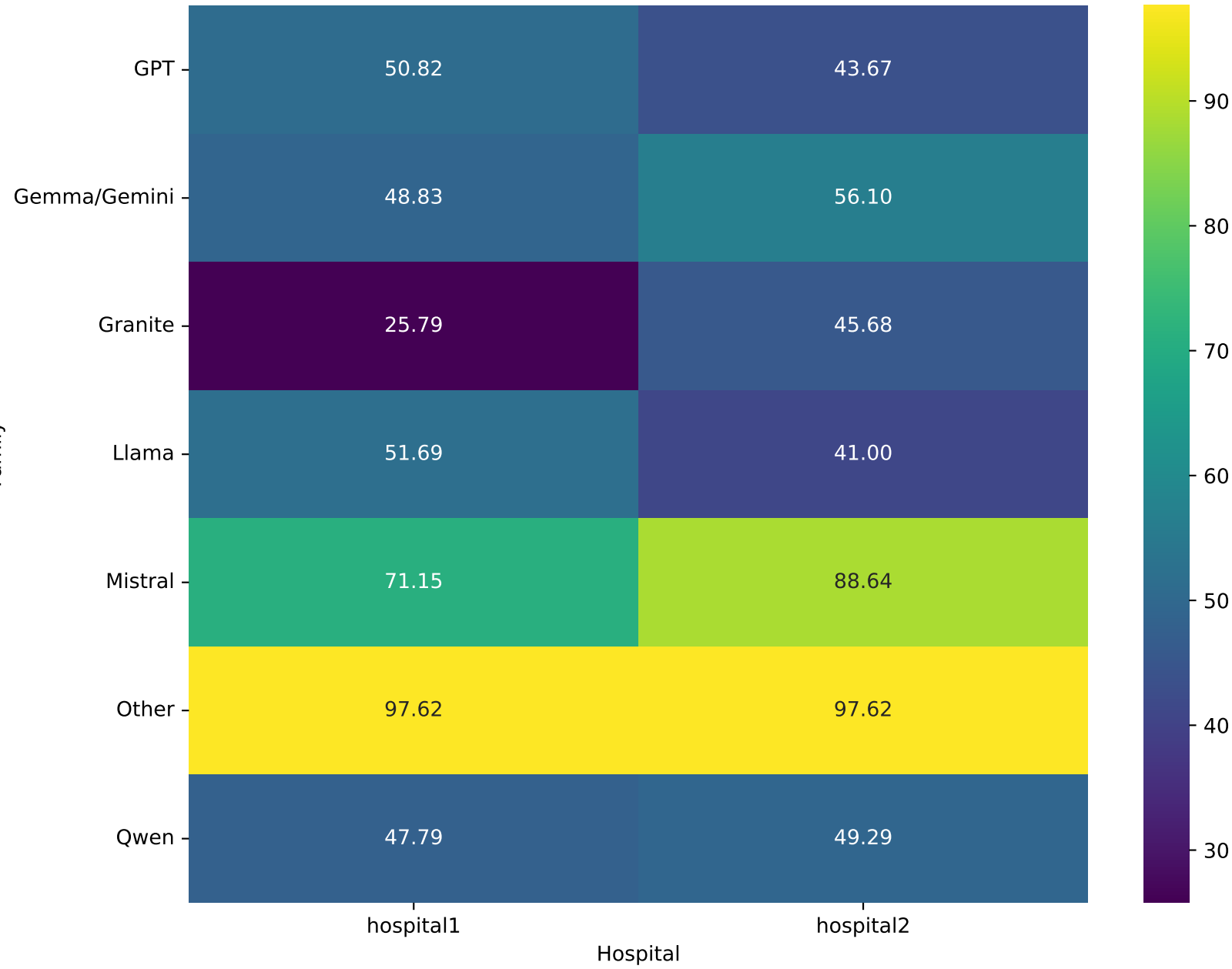


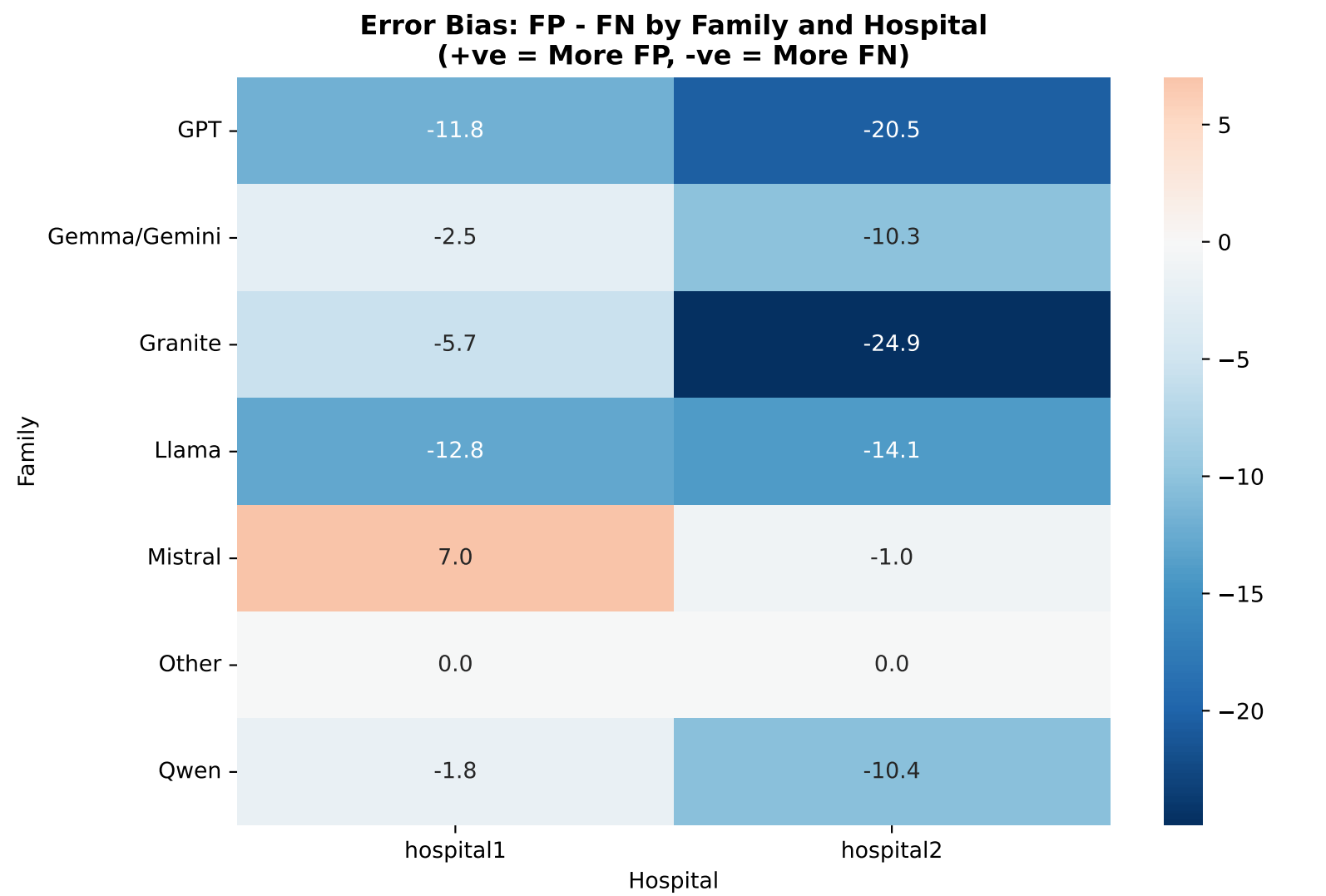
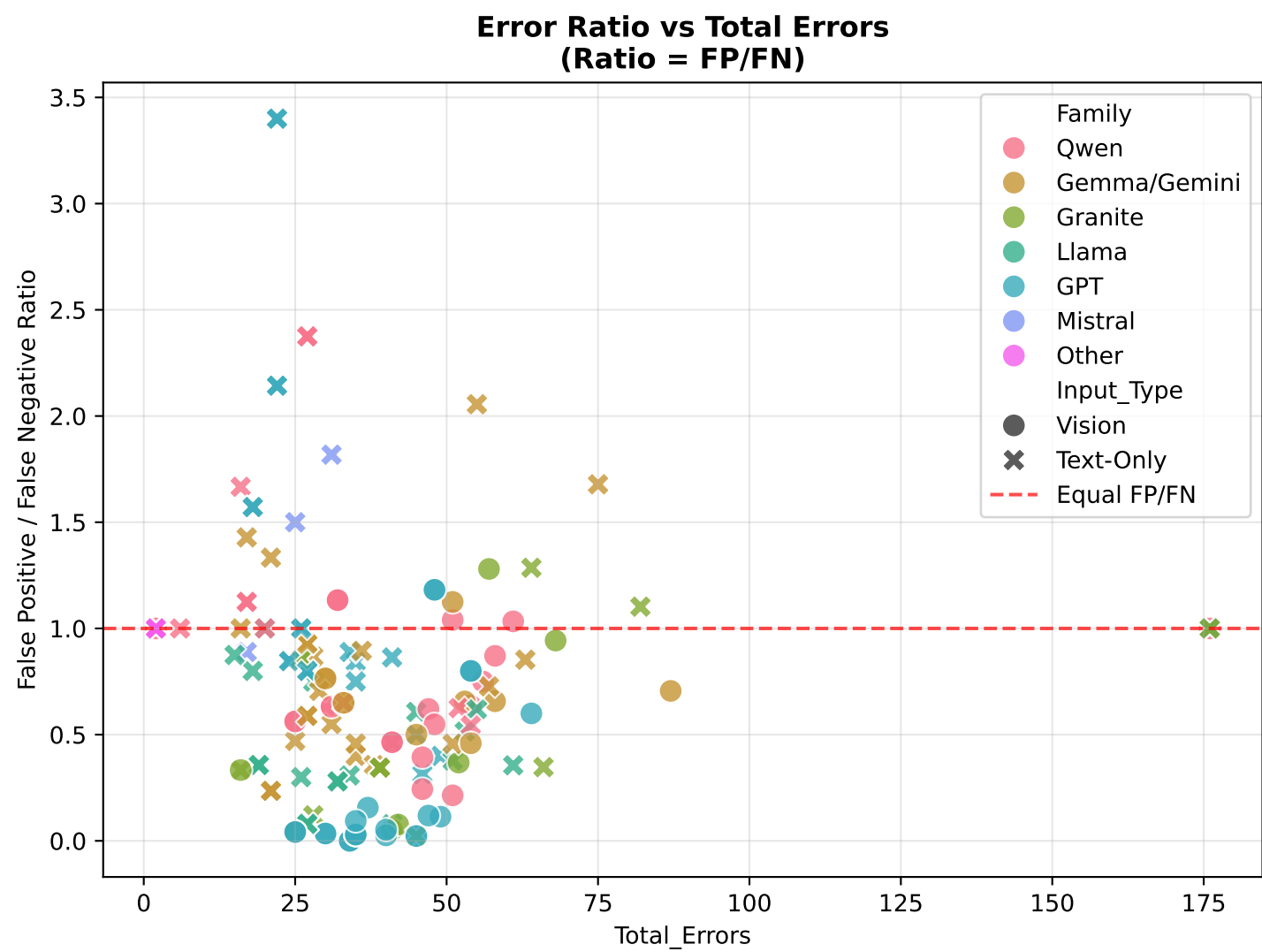
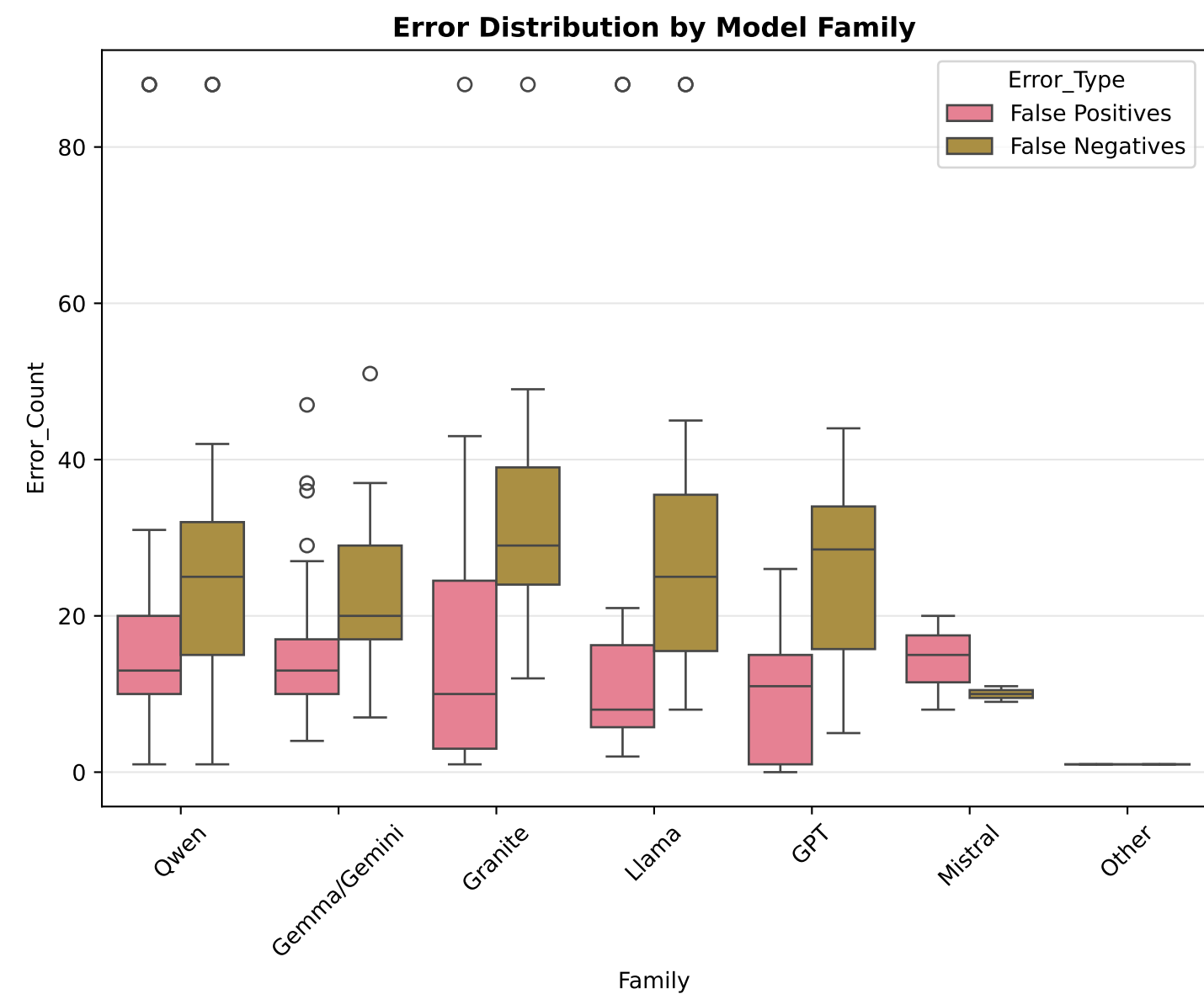
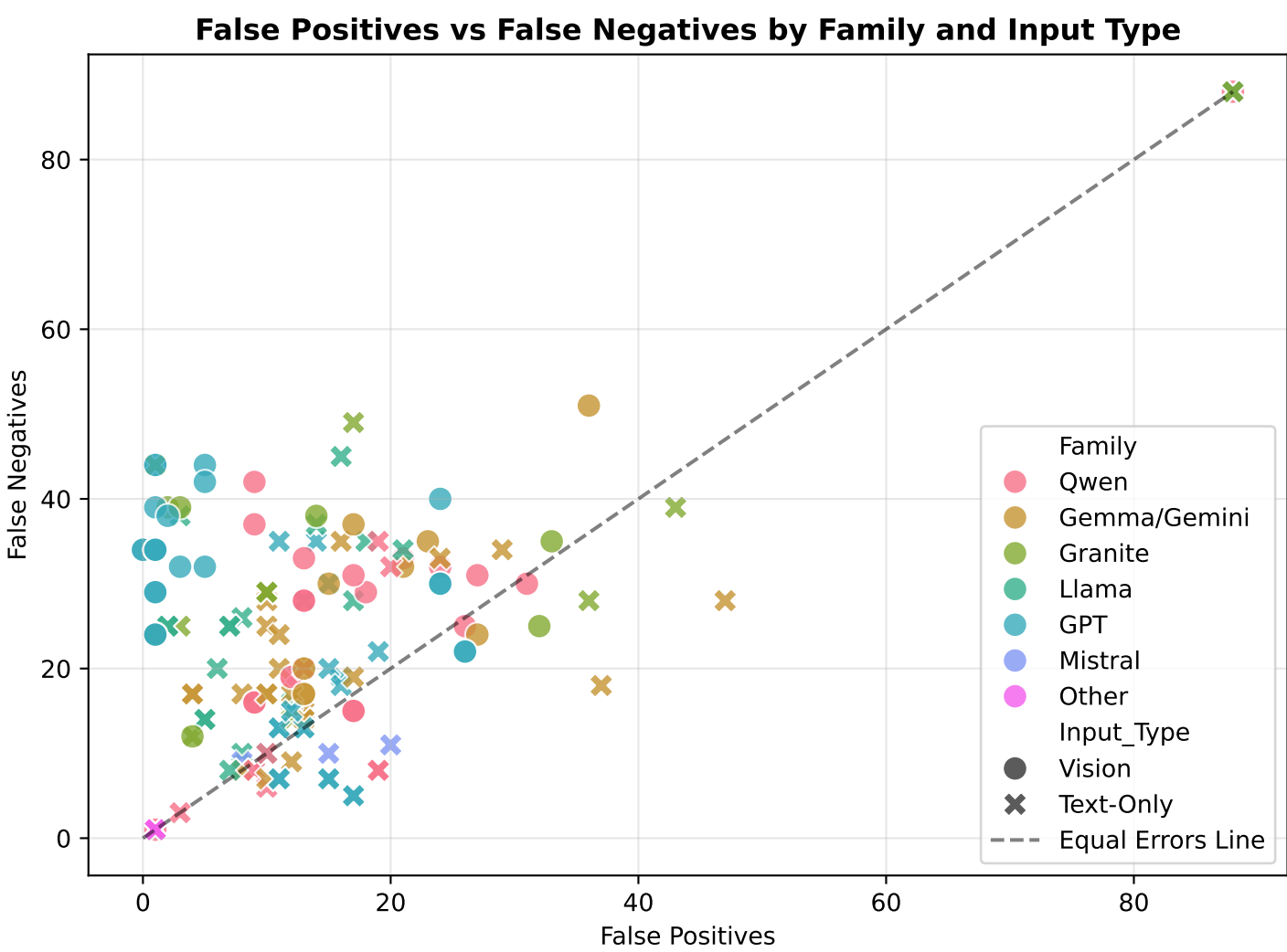


Average F1 Score: Family vs Input Type

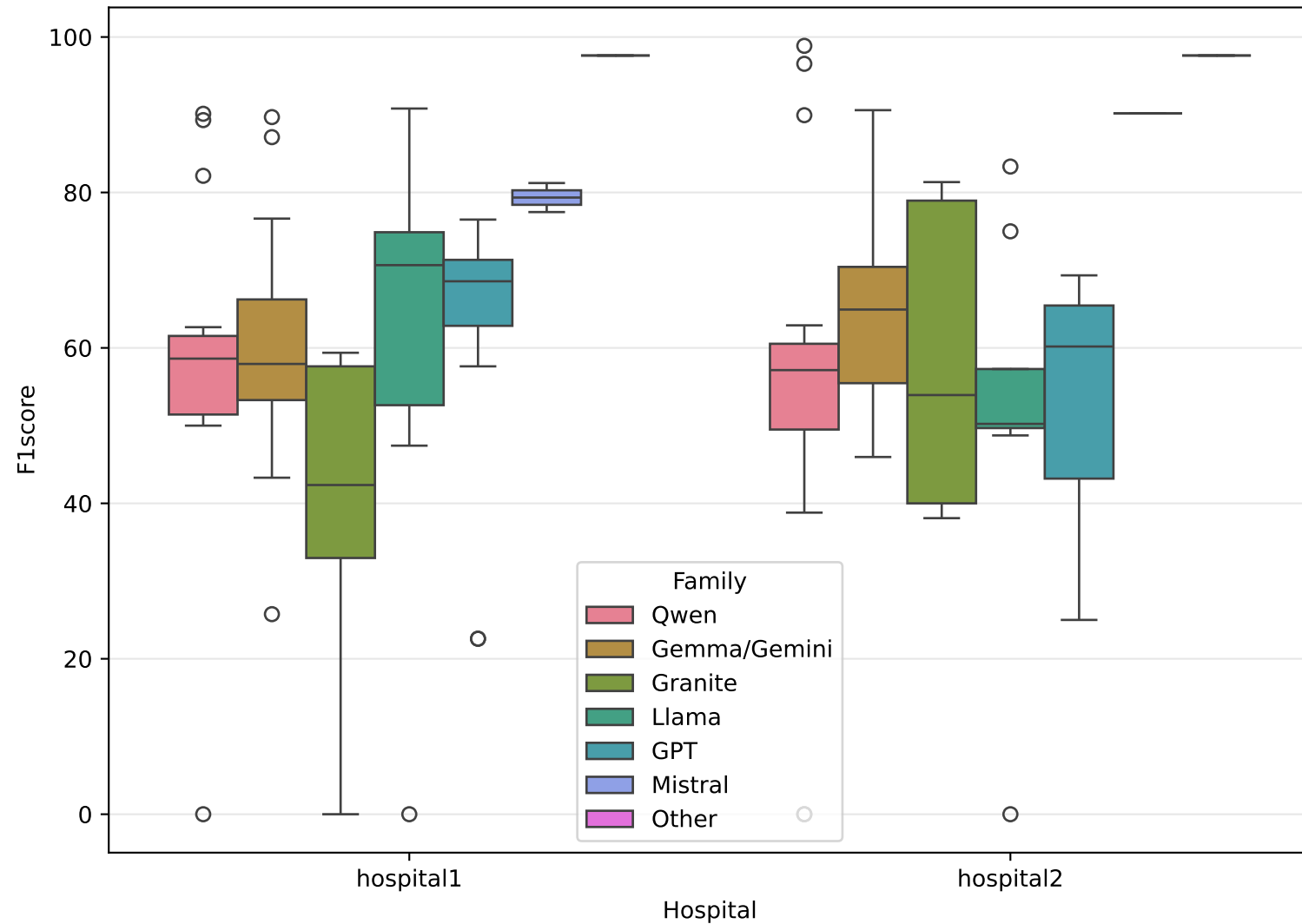


Average Accuracy: Family vs Hospital

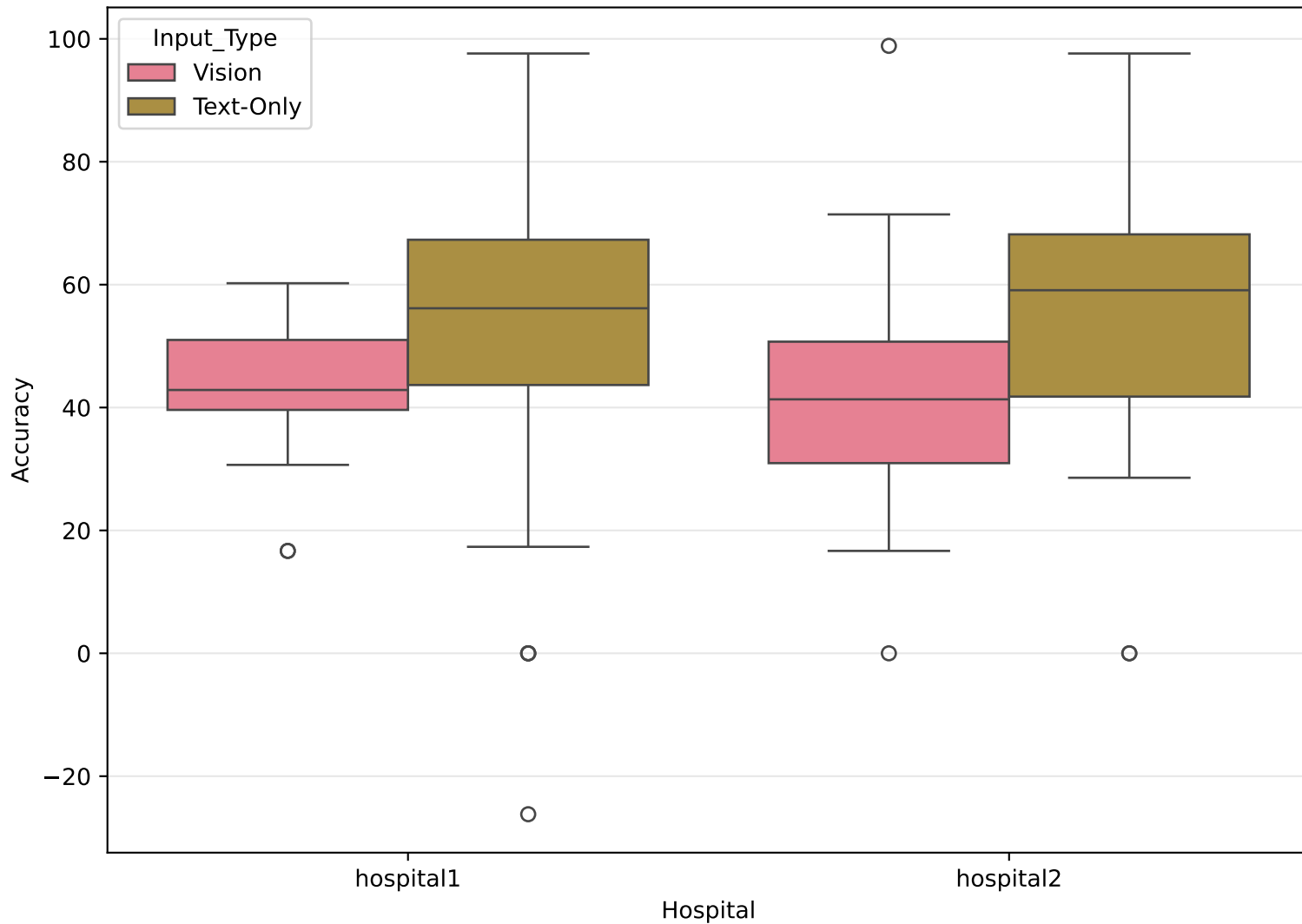




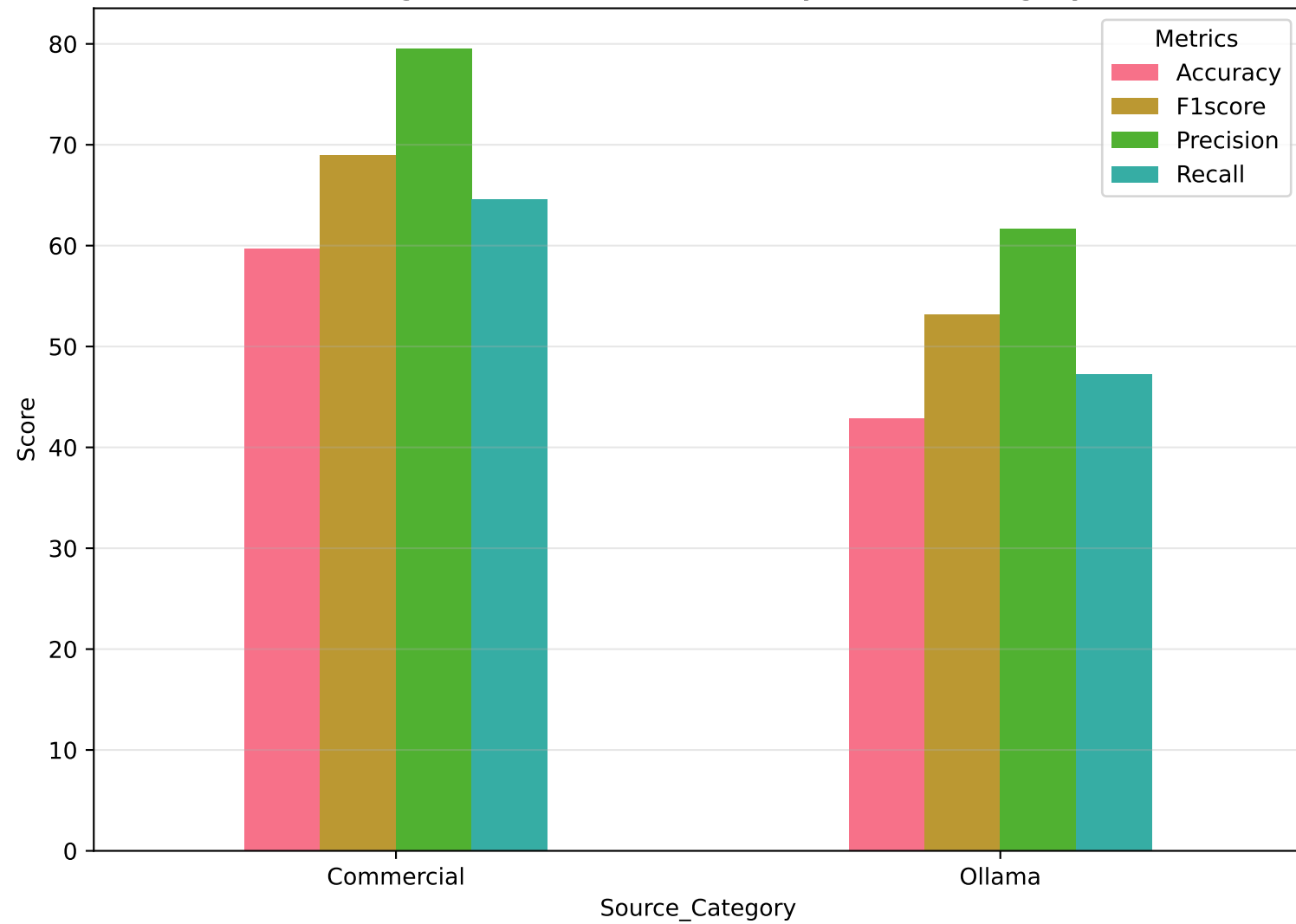
F1 Score Distribution by Hospital and Family



Accuracy Distribution by Hospital and Input Type



Average Performance Metrics by Source Category



F1 Score Distribution by Source and Input Type

