

COMPREHENSIVE LLM BENCHMARKING ANALYSIS - SUMMARY REPORT

DATASET OVERVIEW:

- Total Records: 121
- Model Families: Qwen, Gemma/Gemini, Granite, Llama, GPT, Mistral
- Hospitals: hospital1, hospital2
- Vision-enabled Models: 49
- Text-only Models: 72
- Ollama Models: 70
- Commercial Models: 51

TOP PERFORMERS:

- qwen2.5vl3b*ImageInput* (Qwen, 3.0B, Vision)
F1: 98.86, Accuracy: 98.86
- qwen2.5:72b (Qwen, 72.0B, Text-only)
F1: 96.55, Accuracy: 95.45
- llama-4-scout (Llama, Unknown, Text-only)
F1: 90.80, Accuracy: 84.09
- gemini-2.0 (Gemma/Gemini, Unknown, Text-only)
F1: 90.59, Accuracy: 87.50
- devstral-small (Mistral, Unknown, Text-only)
F1: 90.17, Accuracy: 88.64

FAMILY PERFORMANCE RANKING (by F1 Score):

- Mistral: 82.95 (n=3.0)
- Gemma/Gemini: 62.31 (n=26.0)
- GPT: 61.30 (n=34.0)
- Llama: 58.41 (n=17.0)
- Qwen: 57.27 (n=27.0)
- Granite: 48.18 (n=14.0)

KEY INSIGHTS:

- Vision Models Avg F1: 54.54
- Text-only Models Avg F1: 62.43
- Hospital 1 Avg F1: 60.39
- Hospital 2 Avg F1: 57.98
- Ollama Models Avg F1: 52.73
- Commercial Models Avg F1: 68.16

Grouped Model F1 Score Statistics:

- Unique Base Models: 16
- Total Test Instances: 121
- Best Performing Model: qwen2.5:72b (F1: 89.61)
- Worst Performing Model: qwen31.7b (F1: 38.54)
- Overall Average F1: 63.15
- Models with Vision: 6

Top 5 Performers:

- llama-4-scout (Llama, Text-only): F1 = 68.98 ± 38.73
- mistral-3.1-24b (Mistral, Text-only): F1 = 77.48 ± nan
- 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

Bottom 5 Performers:

- qwen31.7b (Qwen, Text-only): F1 = 38.54 ± 29.88
- llama3.21b (Llama, Text-only): F1 = 44.97 ± 23.45
- granite3.2 (Granite, with Vision): F1 = 46.89 ± 18.80
- gpt-4o (GPT, with Vision): F1 = 51.23 ± 22.82
- gemma34b (Gemma/Gemini, with Vision): F1 = 52.50 ± 10.01

Error Analysis Summary:

- Average False Positives: 16.64
- Average False Negatives: 28.50
- Models with more FP than FN: 20
- Models with more FN than FP: 89

SOURCE CATEGORY DETAILED STATISTICS:

COMMERCIAL MODELS:

GPT:

Accuracy: 48.8±13.2 (n=34)
F1 Score: 61.3±14.0
Precision: 78.9±19.6
False Positives: 8.8
False Negatives: 28.0

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

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: 48.1±12.0 (n=22)
F1 Score: 57.7±12.5
Precision: 63.2±17.0
False Positives: 18.7
False Negatives: 26.5

Granite:

Accuracy: 36.4±18.0 (n=14)
F1 Score: 48.2±20.2
Precision: 64.0±29.9
False Positives: 18.8
False Negatives: 35.7

Llama:

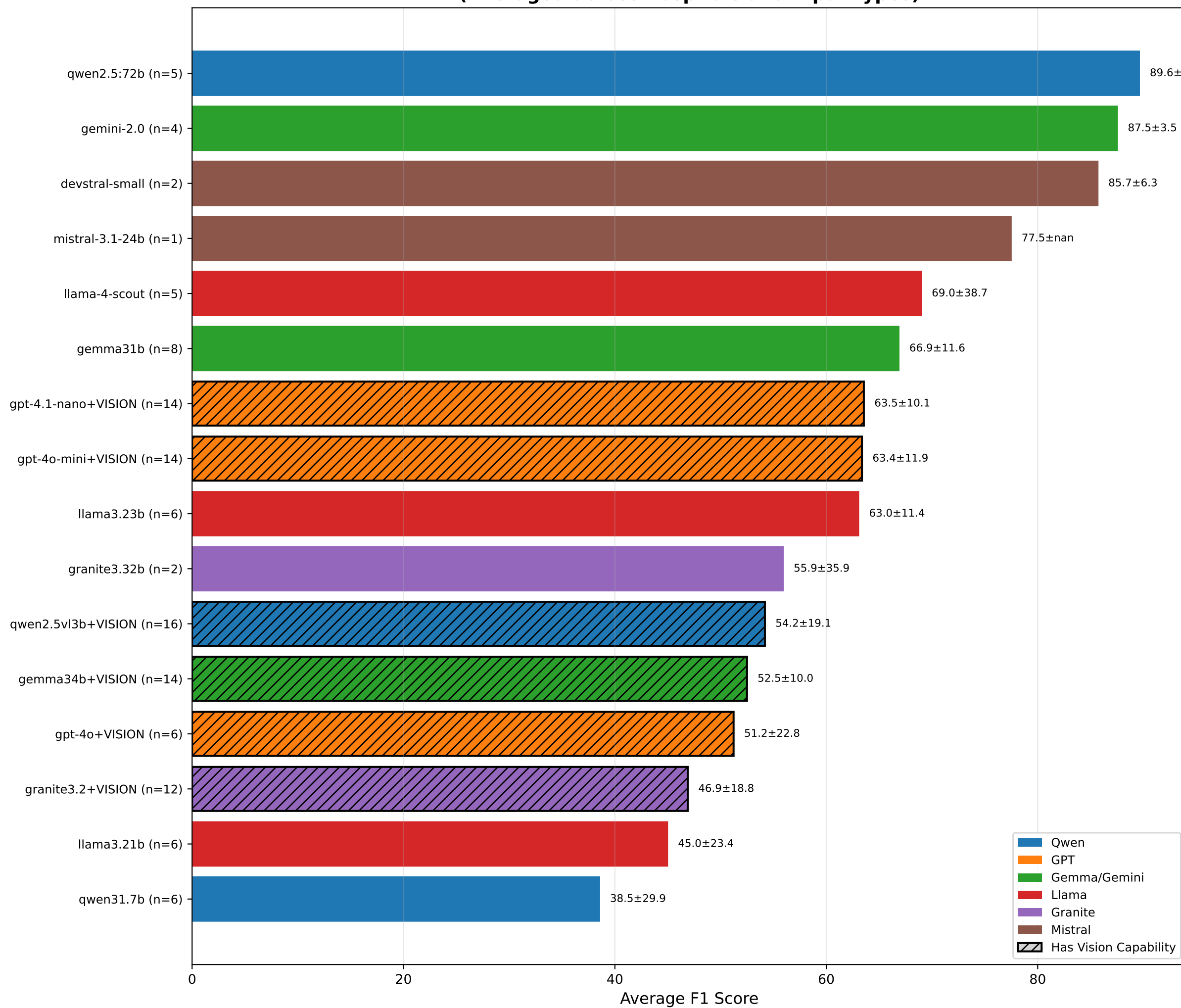
Accuracy: 42.2±17.1 (n=12)
F1 Score: 54.0±20.0
Precision: 67.3±24.4
False Positives: 17.6
False Negatives: 34.2

Qwen:

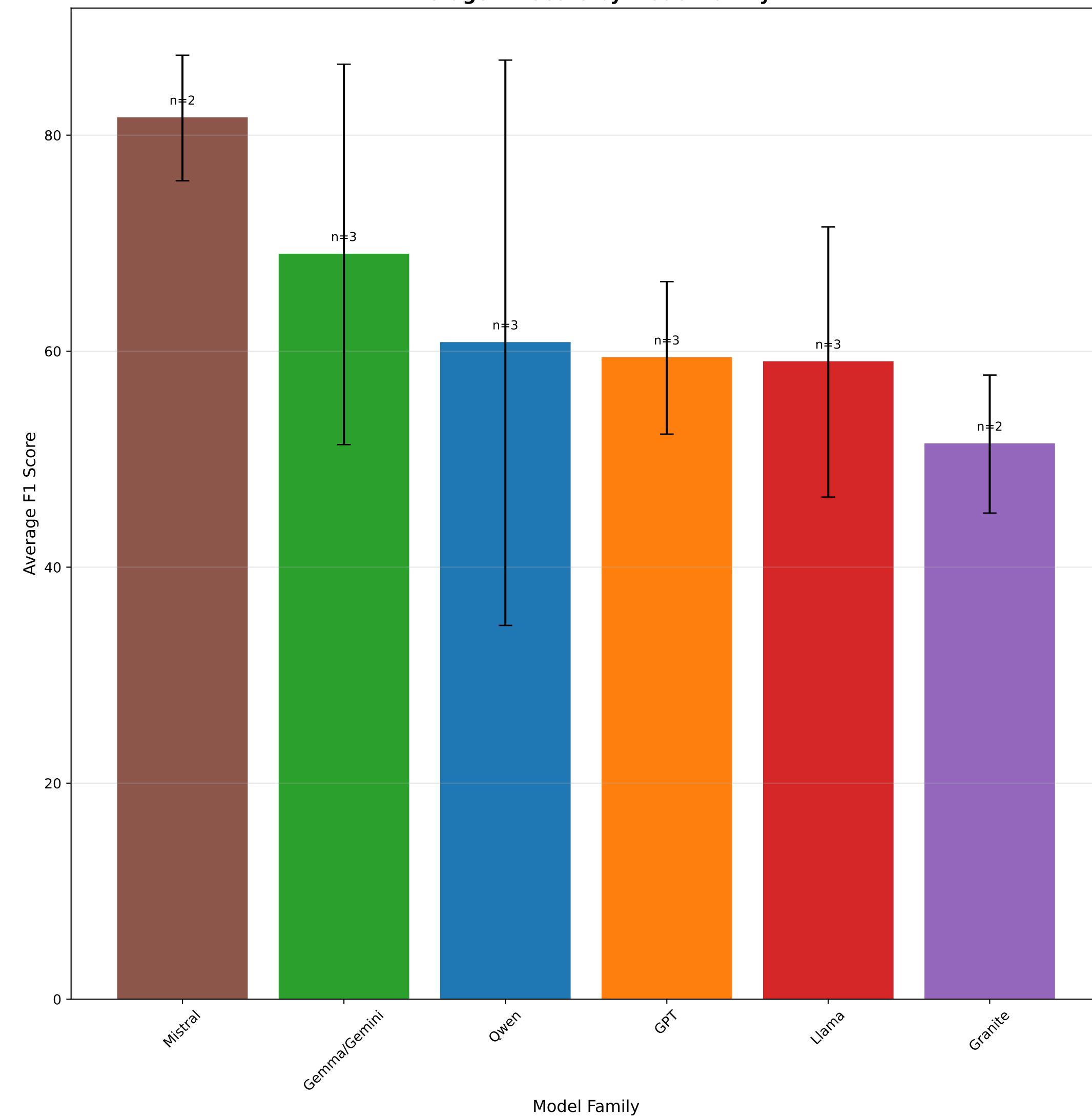
Accuracy: 41.5±21.1 (n=22)
F1 Score: 49.9±22.9
Precision: 55.2±25.1
False Positives: 26.5
False Negatives: 34.6

=====

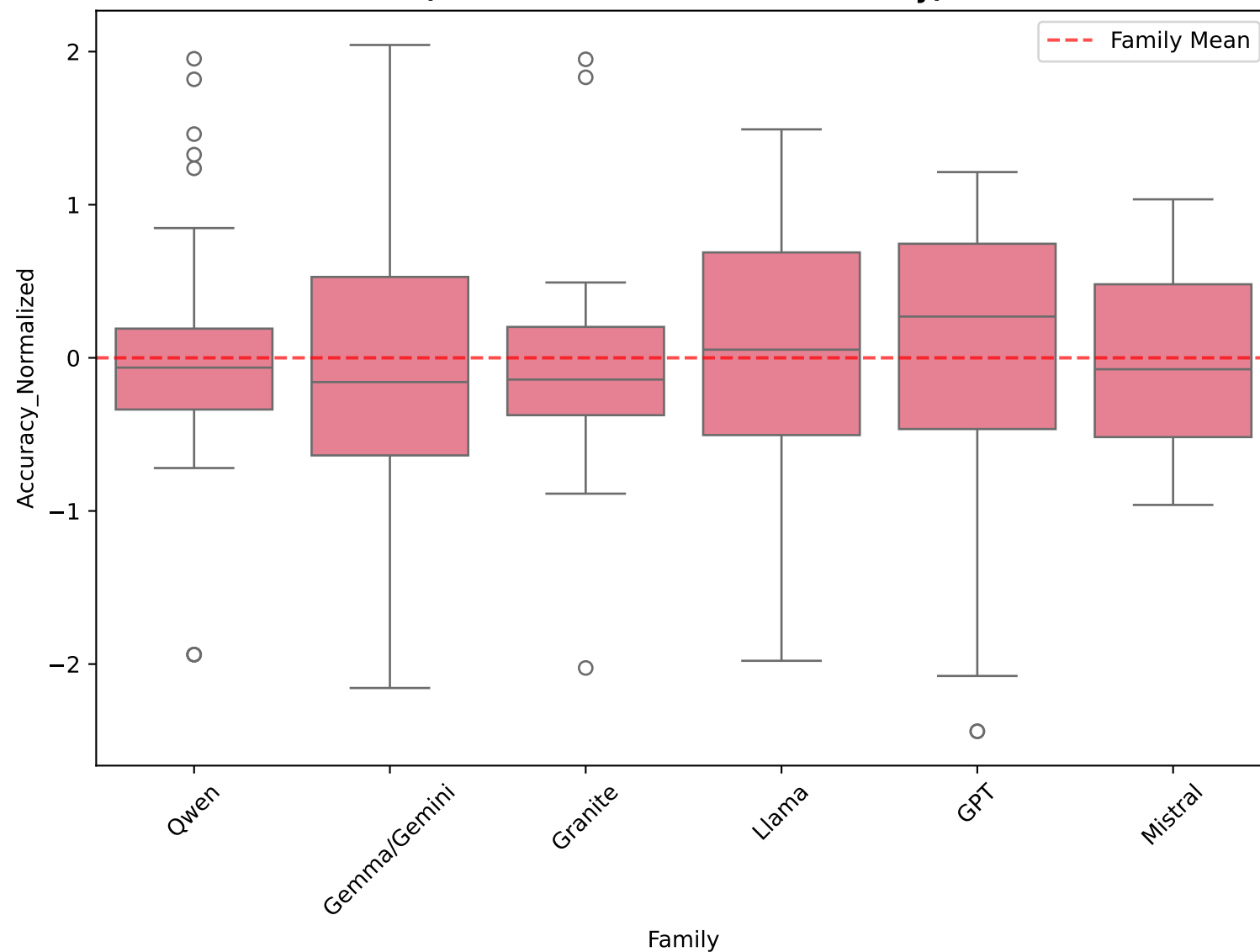
Overall F1 Score Performance - Models Grouped by Base Name
(Averaged across hospitals and input types)



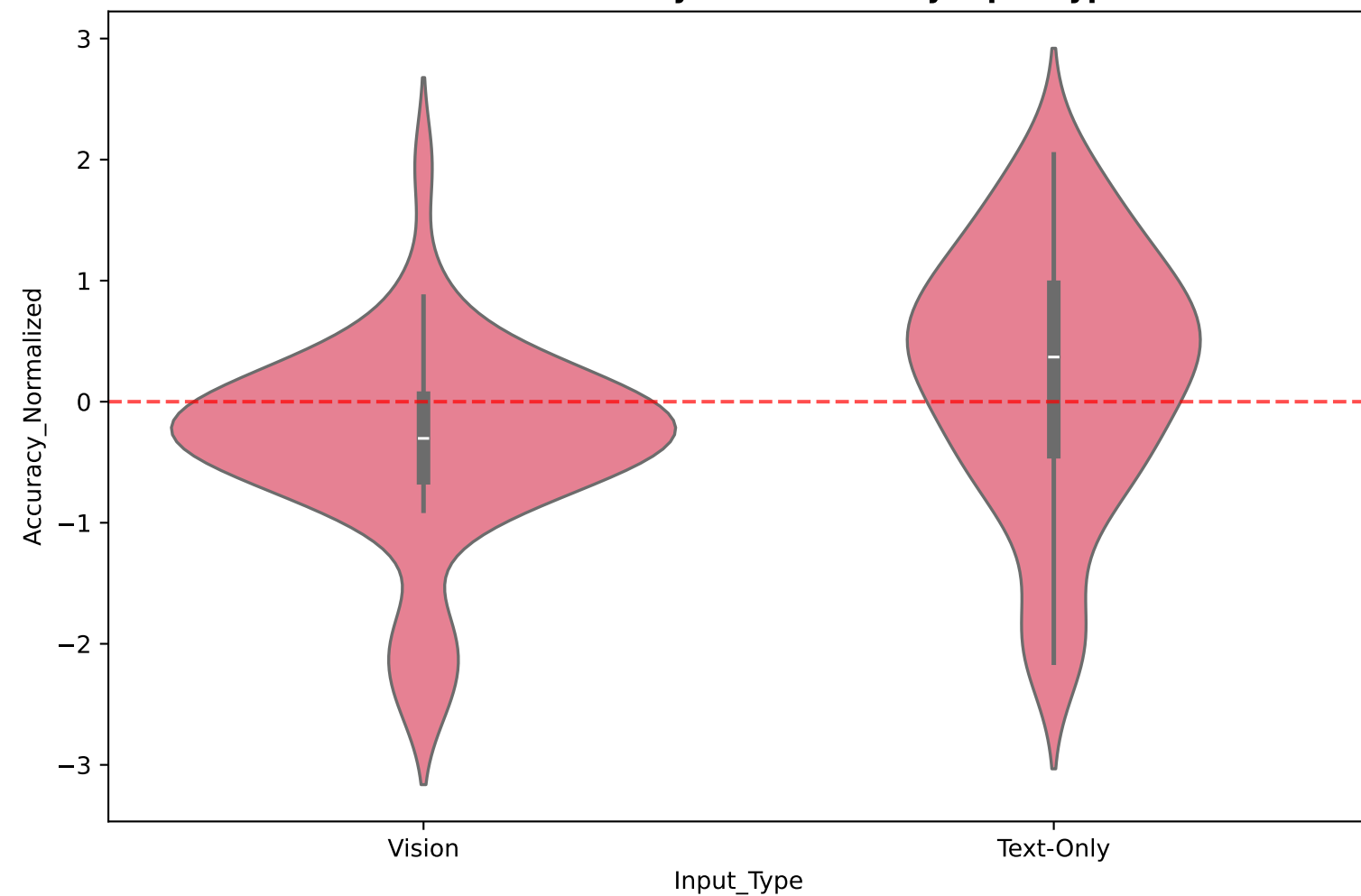
Average F1 Score by Model Family



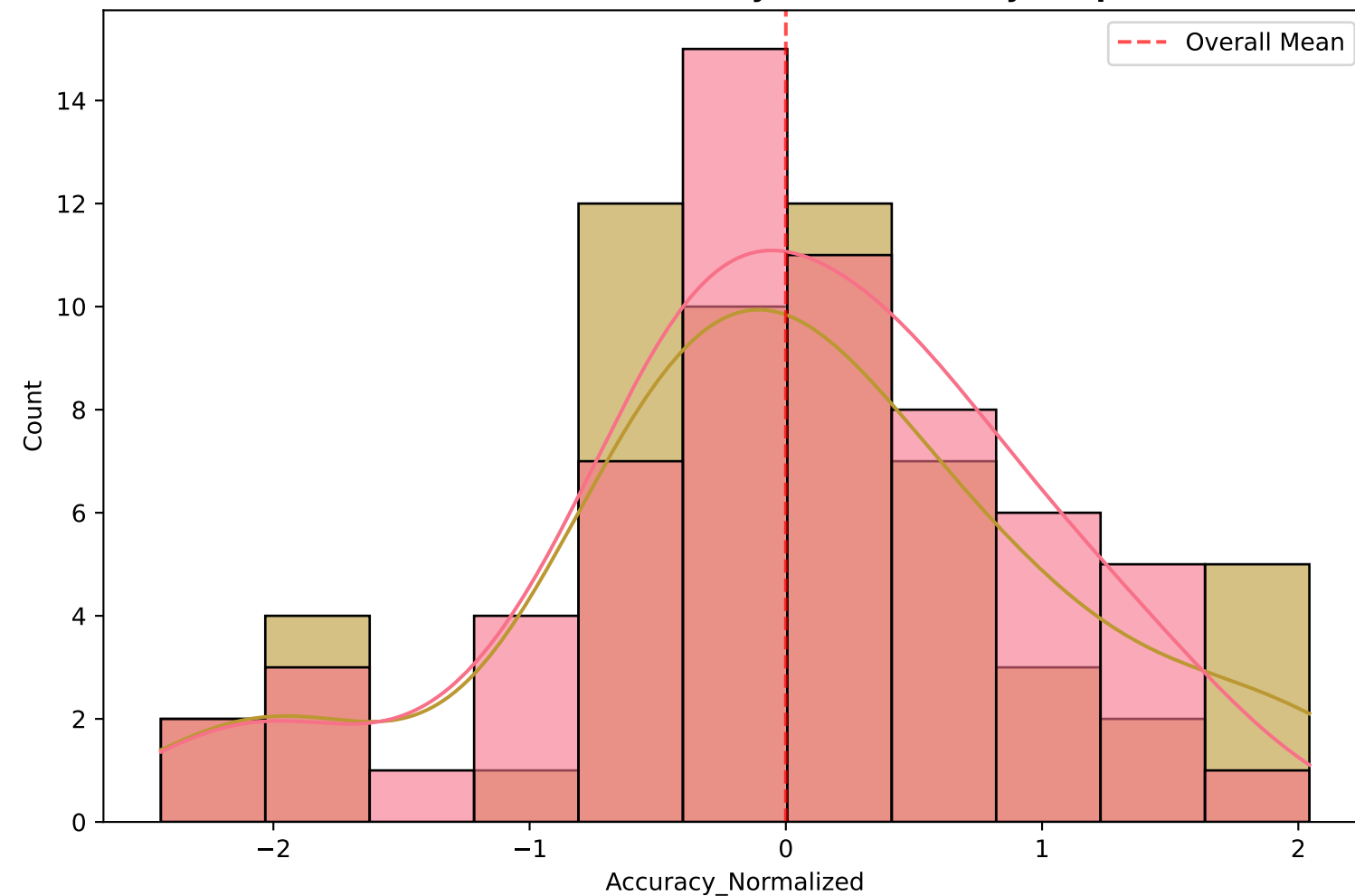
**Normalized Accuracy Distribution by Family
(Z-score normalized within family)**



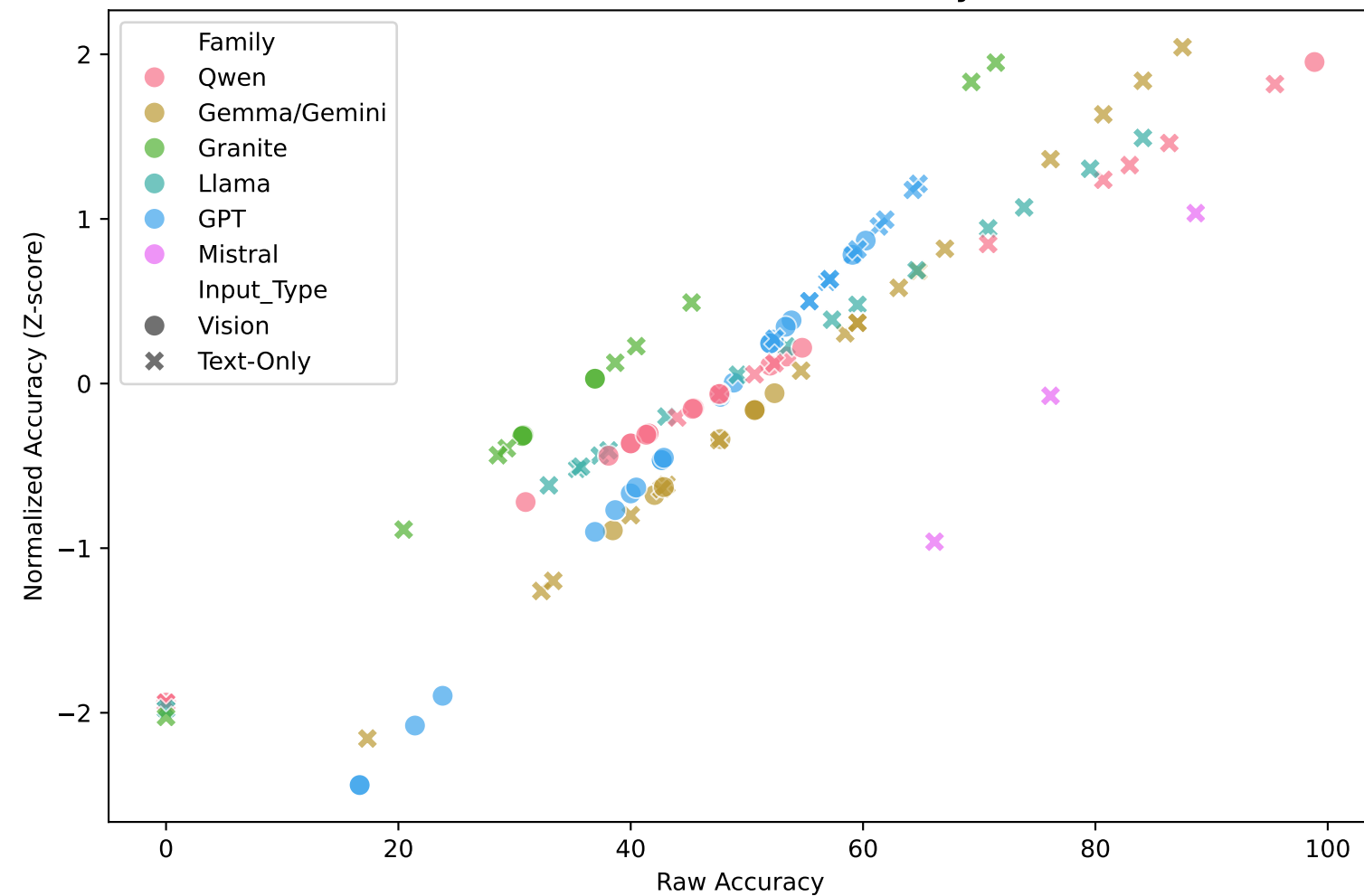
Normalized Accuracy Distribution by Input Type



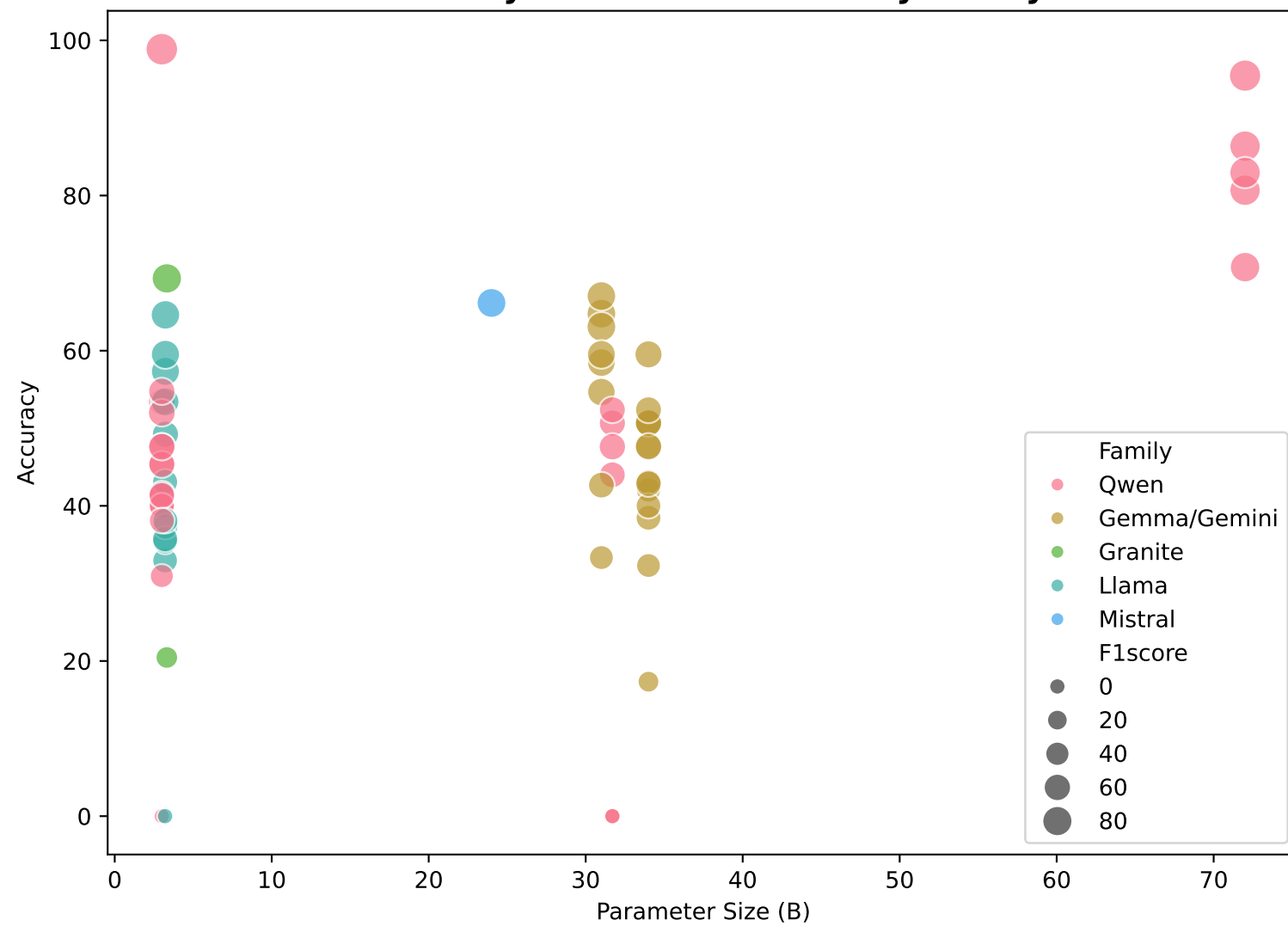
Overall Normalized Accuracy Distribution by Hospital



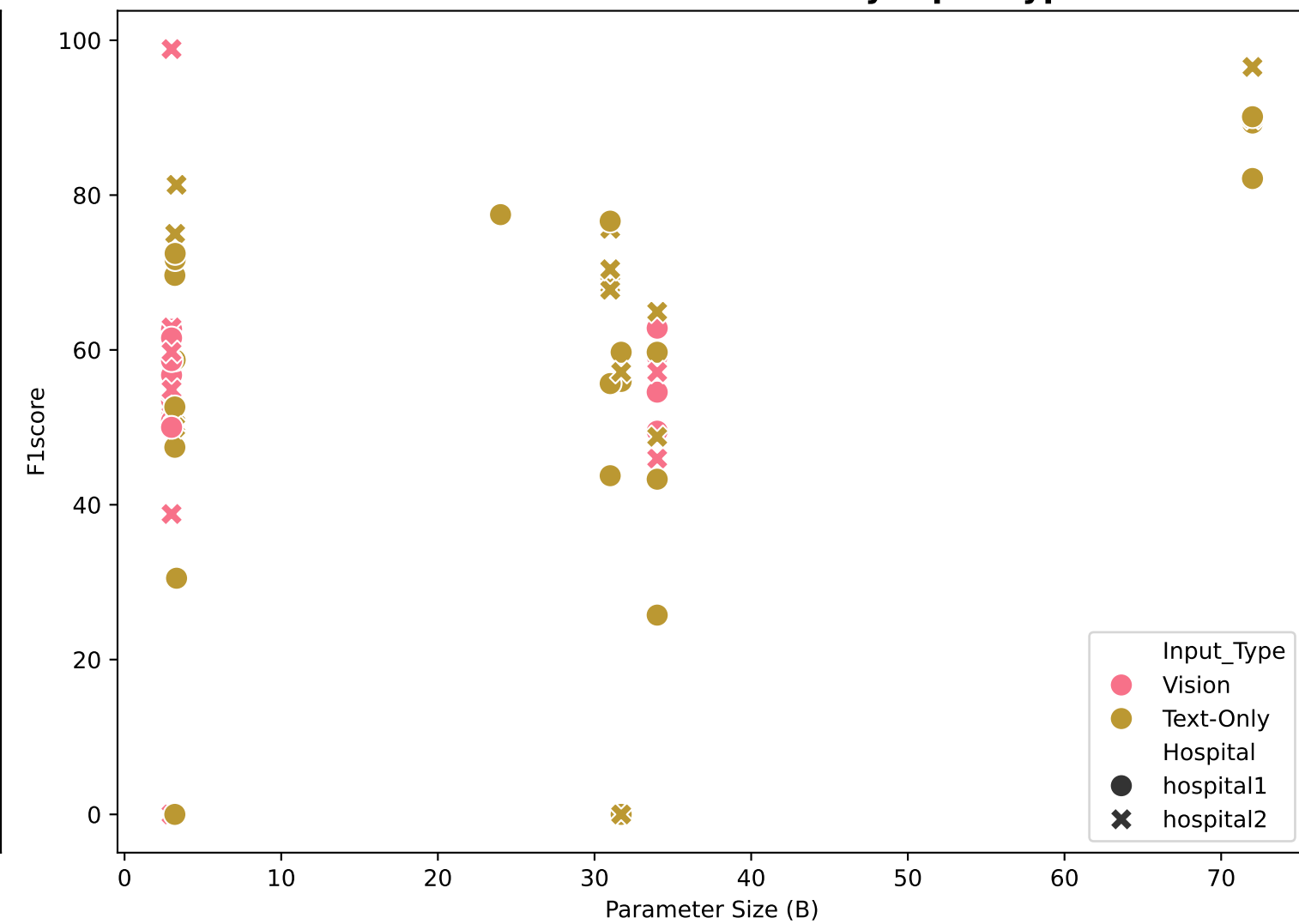
Raw vs Normalized Accuracy



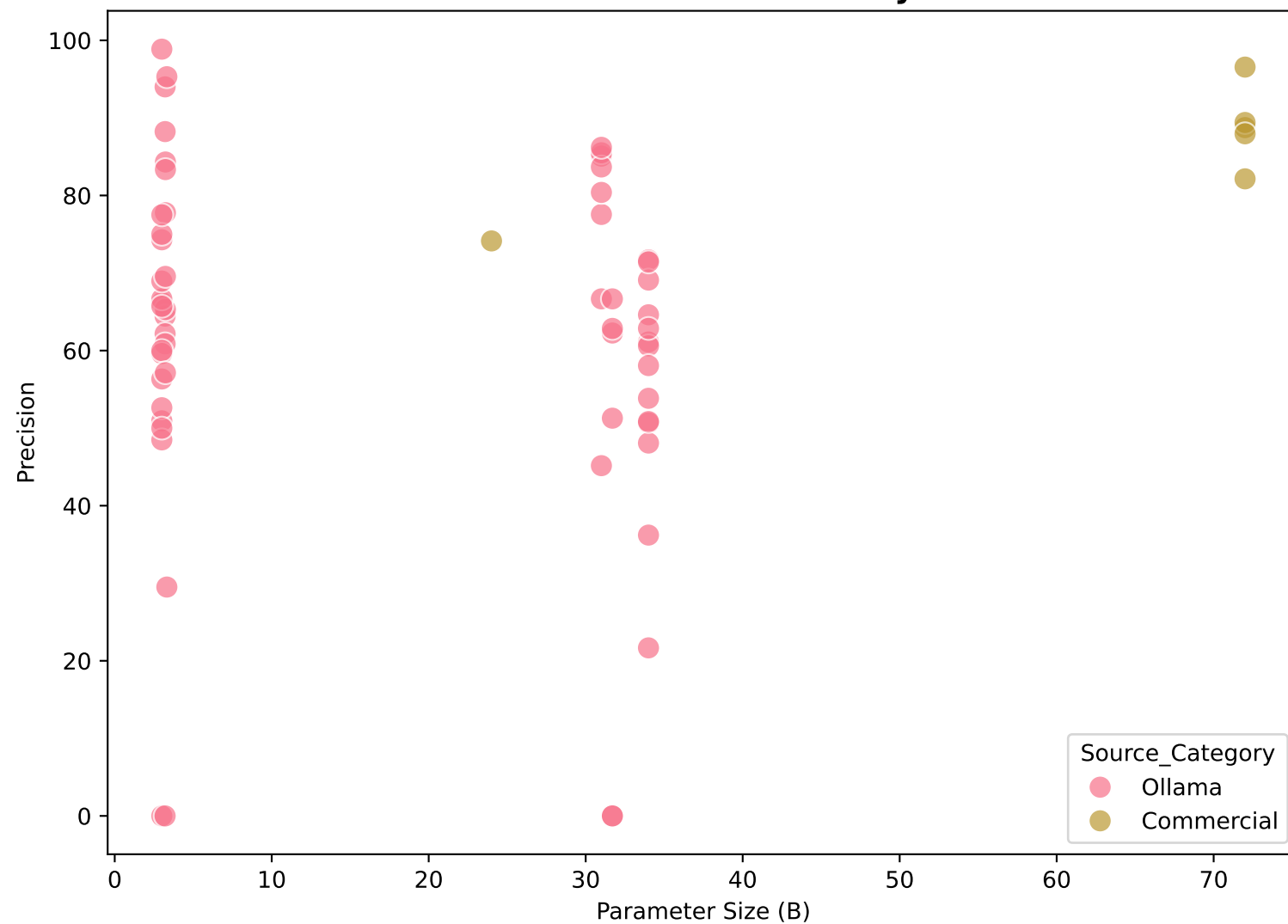
Accuracy vs Parameter Size by Family



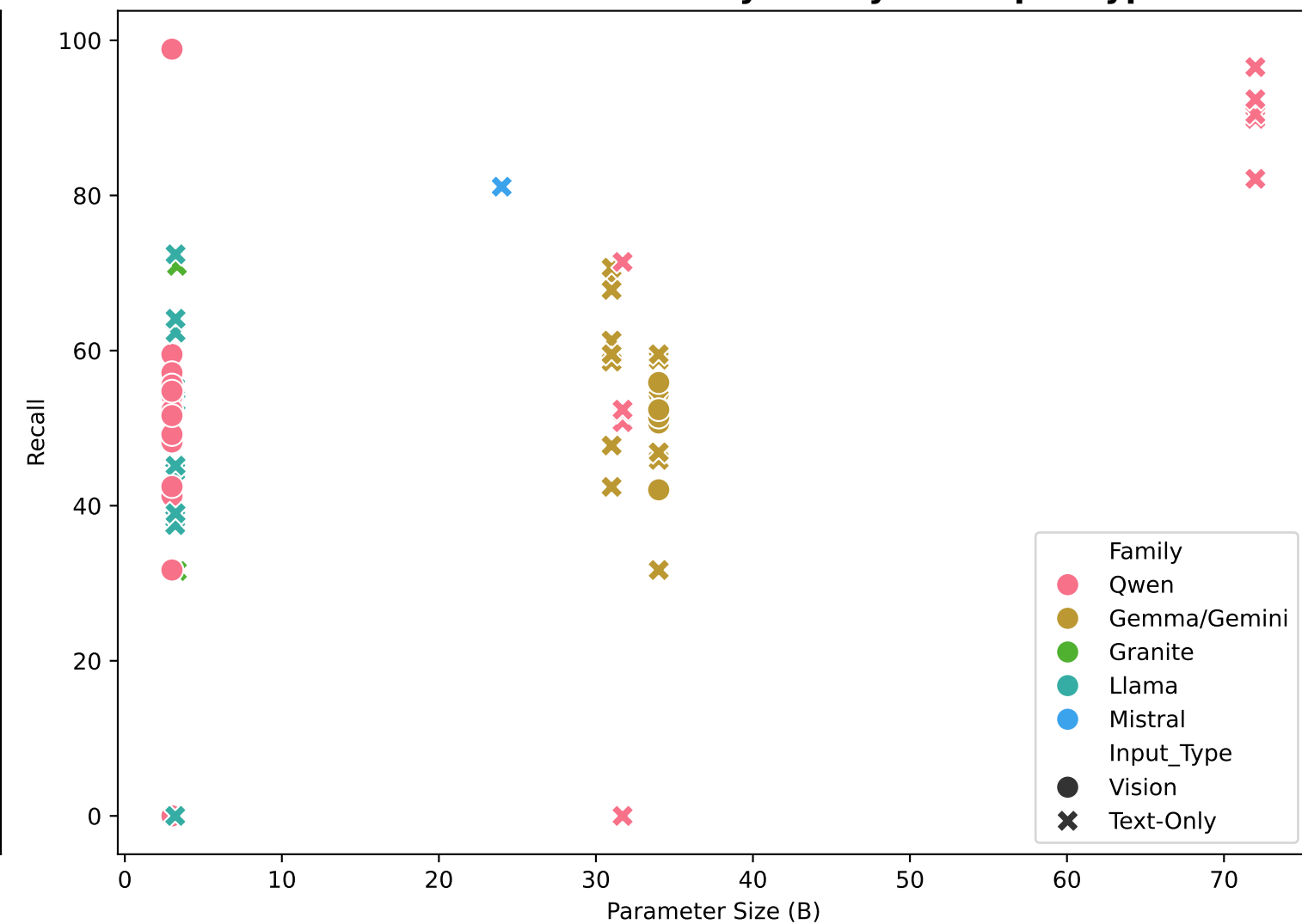
F1 Score vs Parameter Size by Input Type



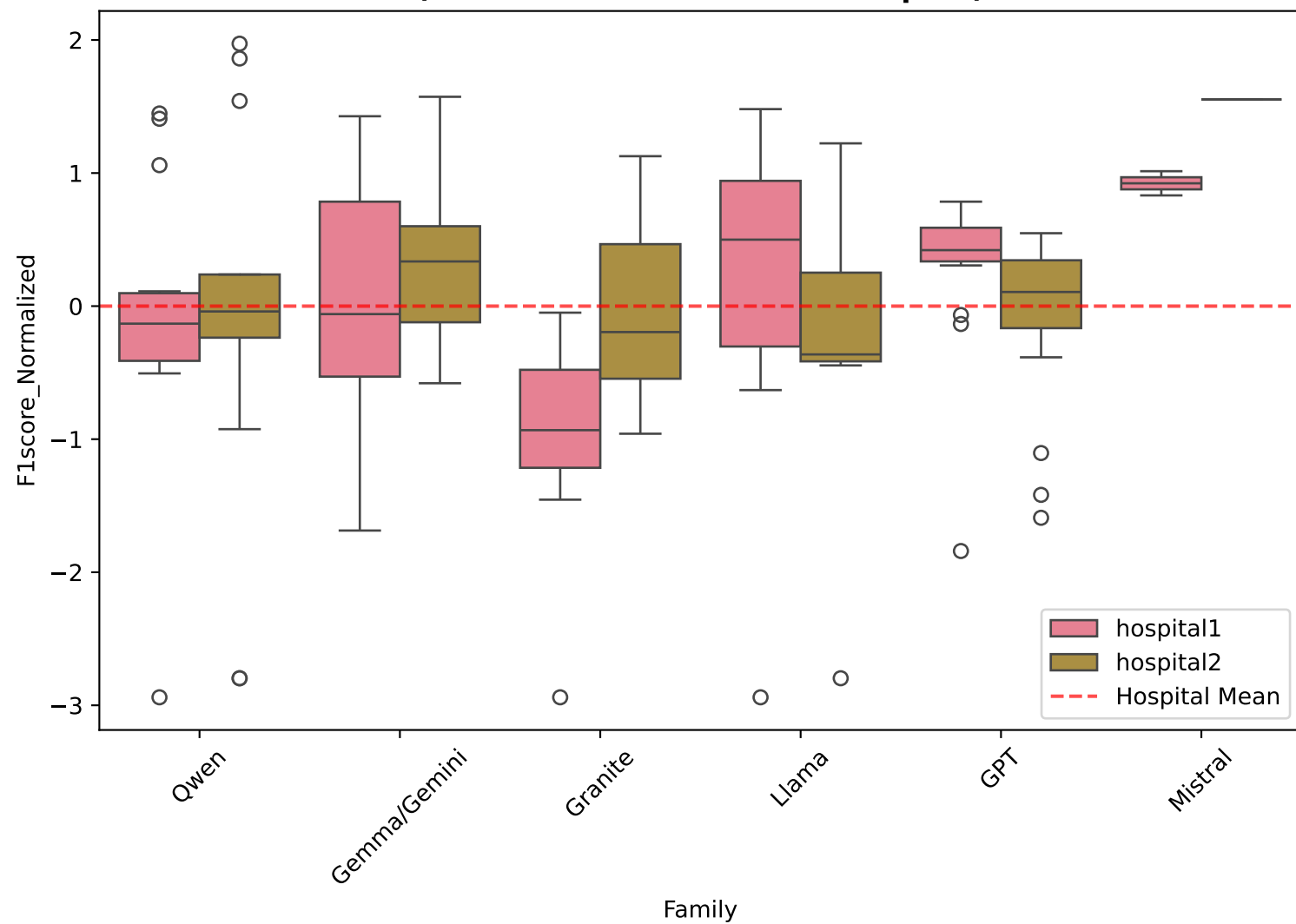
Precision vs Parameter Size by Source



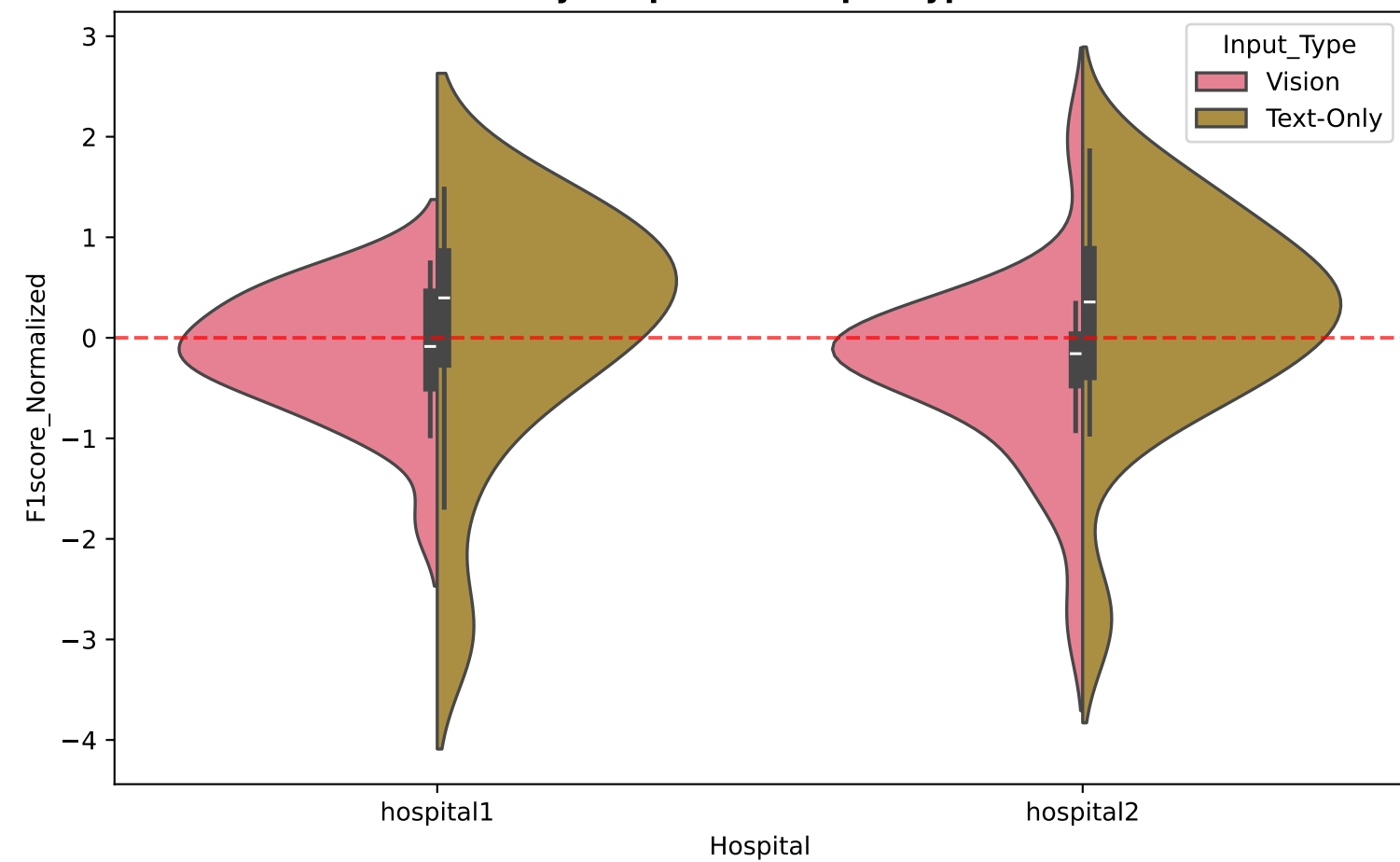
Recall vs Parameter Size by Family and Input Type



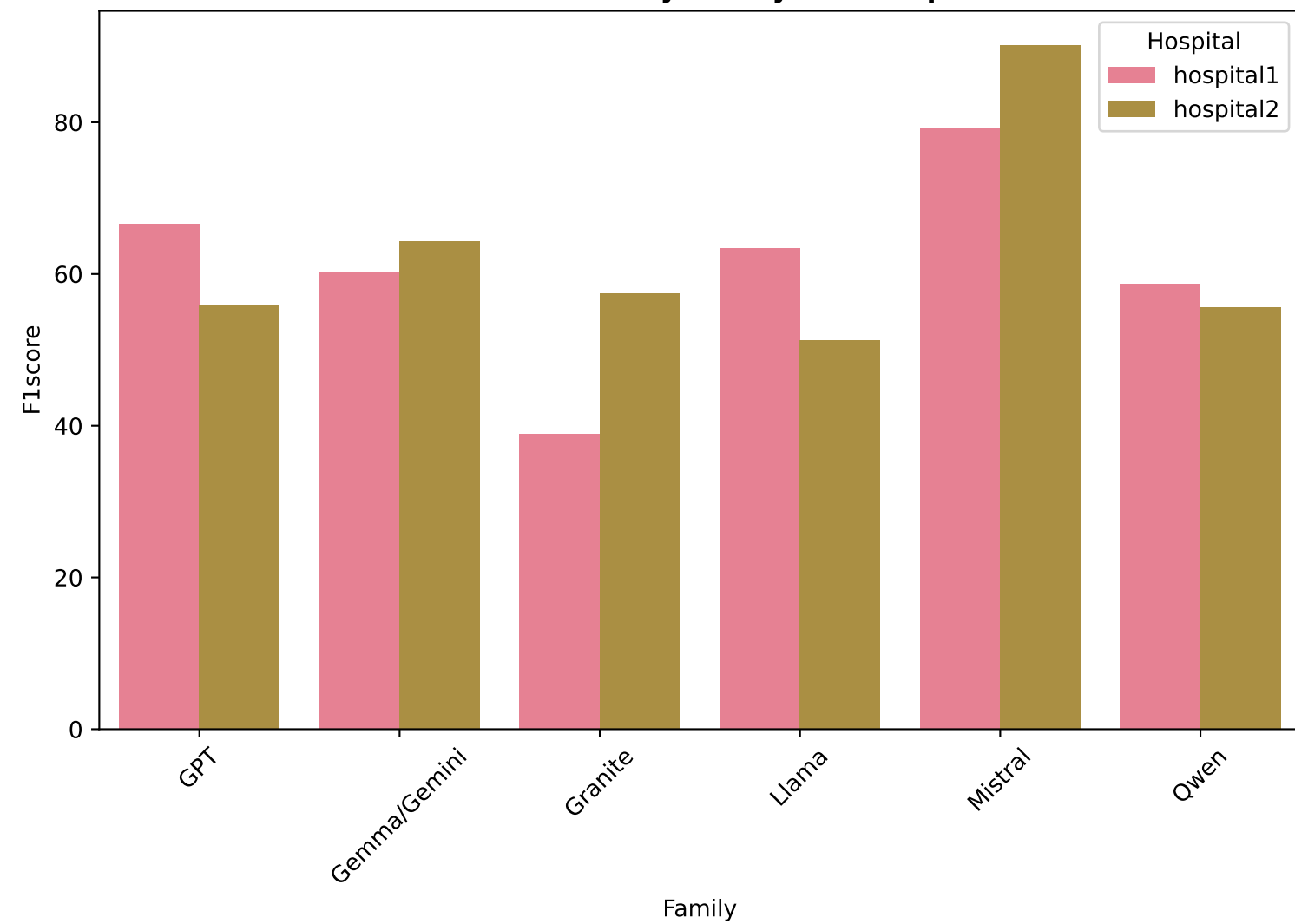
Normalized F1 Score by Family and Hospital
(Z-score normalized within hospital)



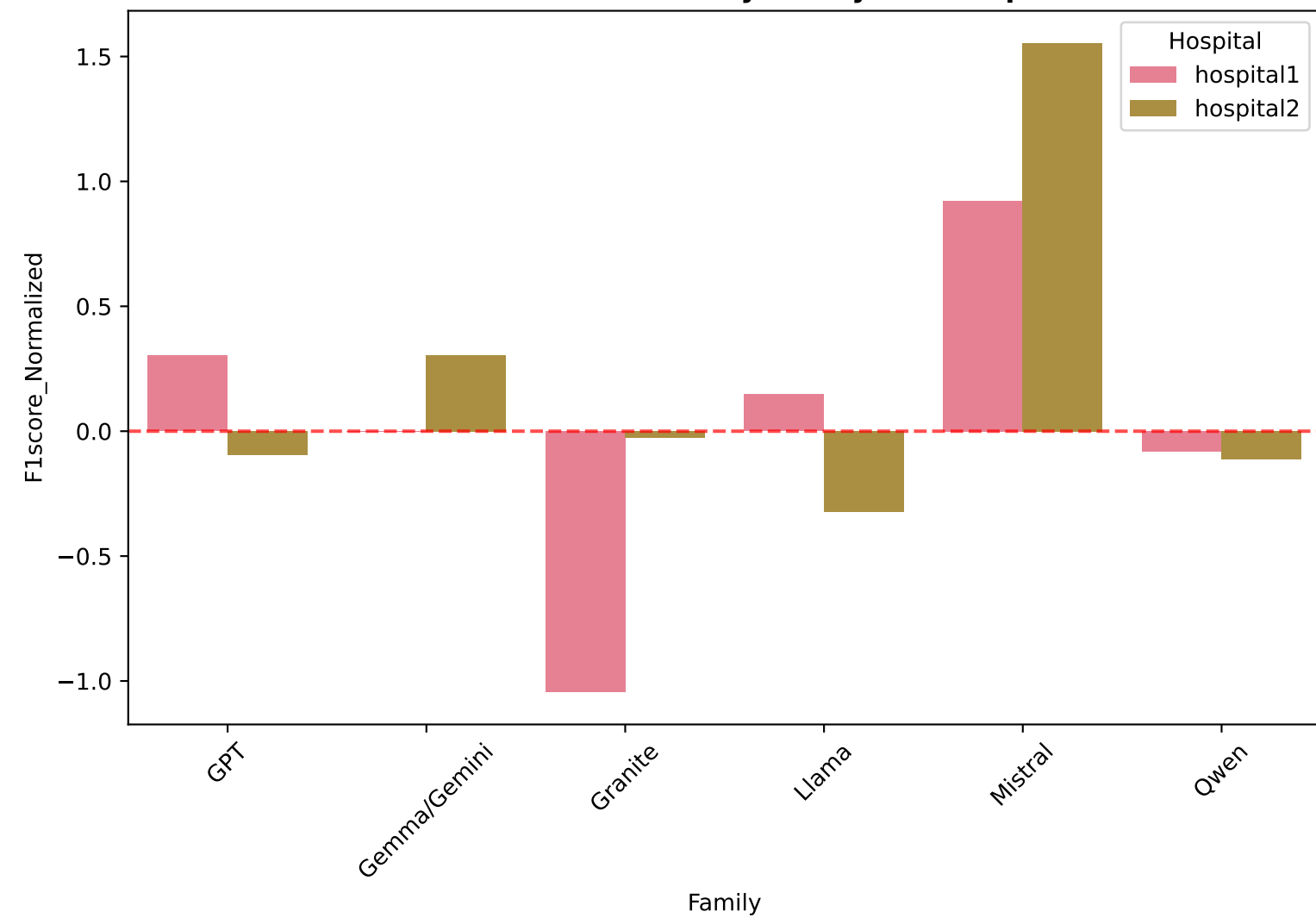
Normalized F1 Score Distribution
by Hospital and Input Type



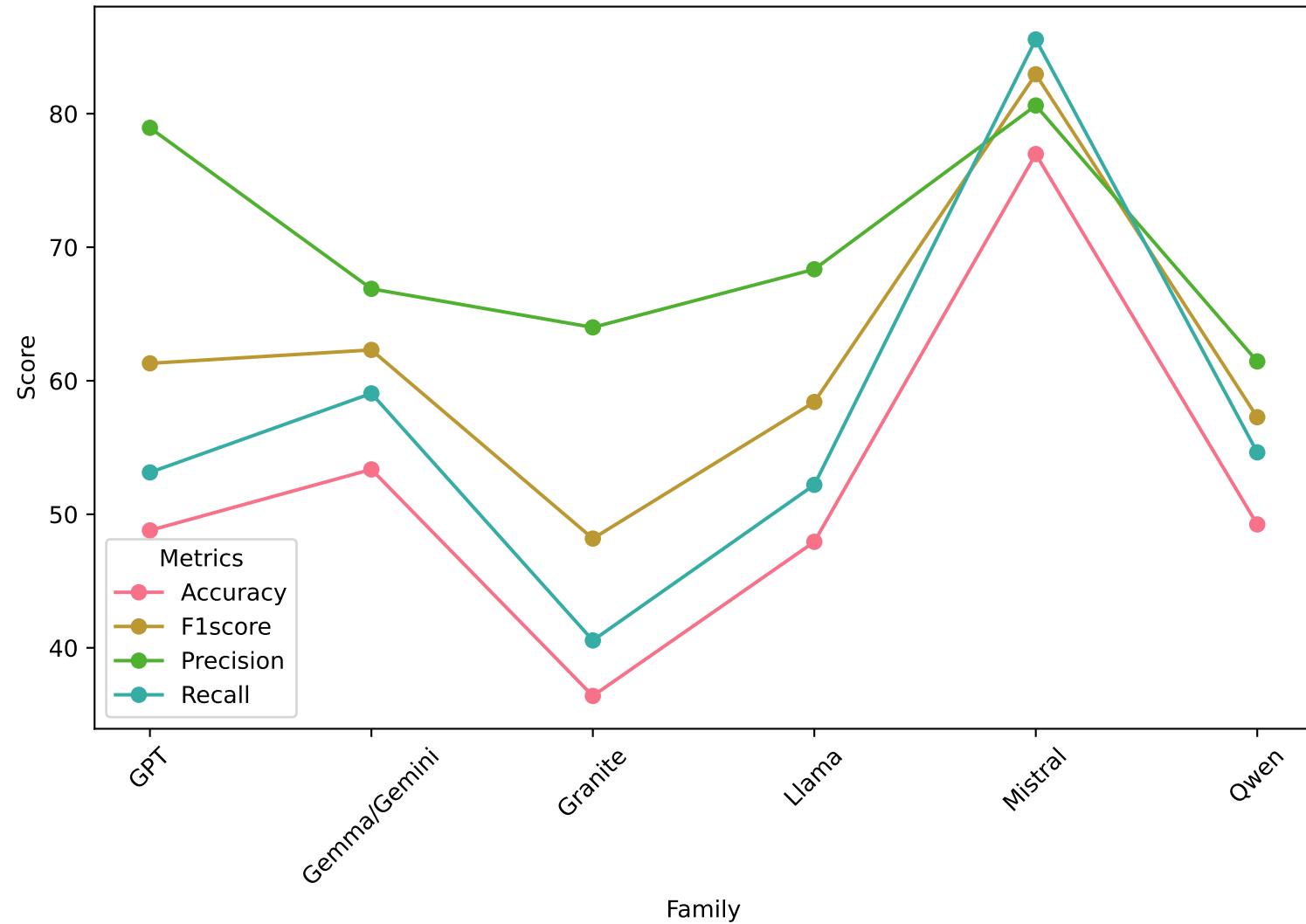
Raw F1 Scores by Family and Hospital



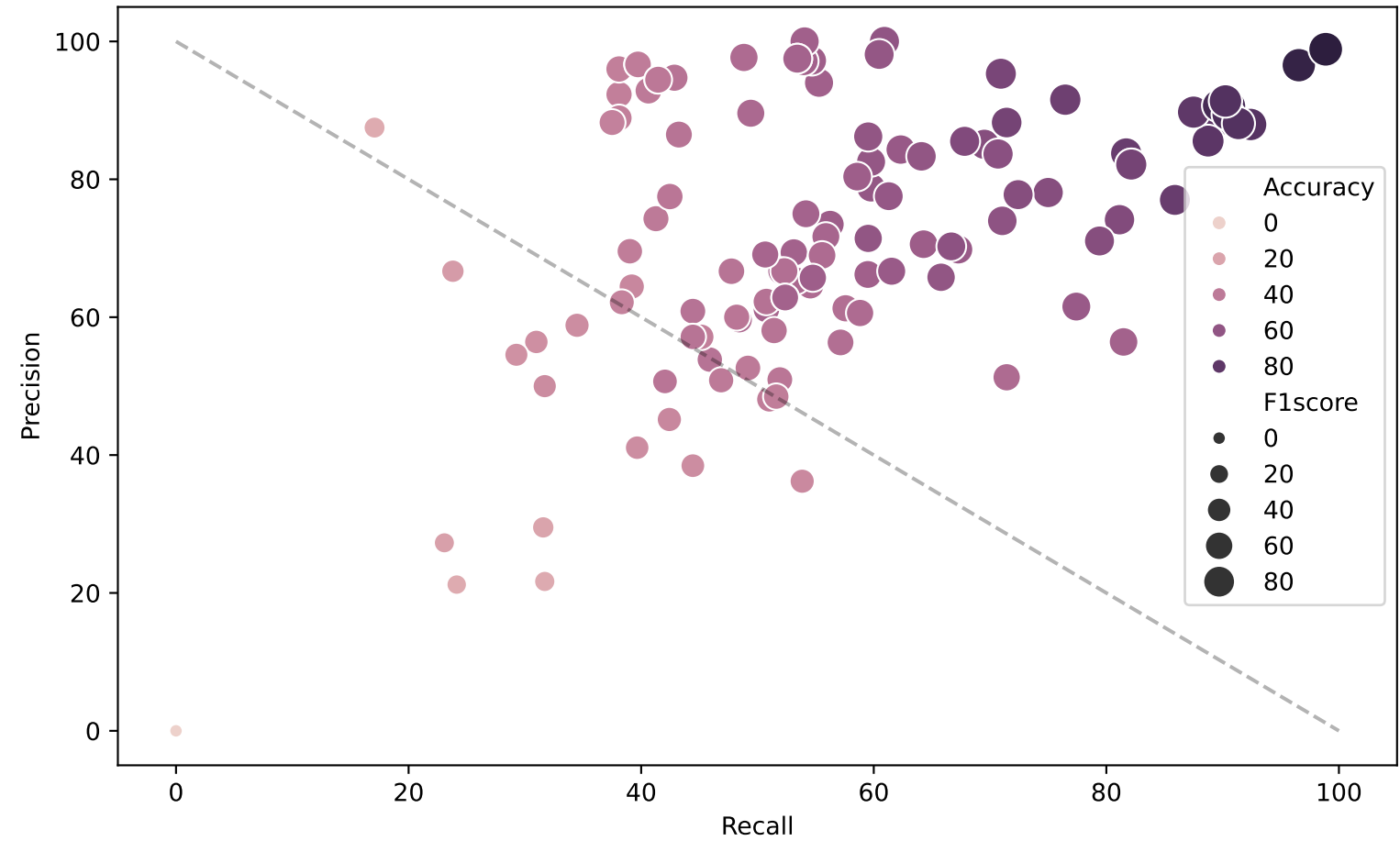
Normalized F1 Scores by Family and Hospital



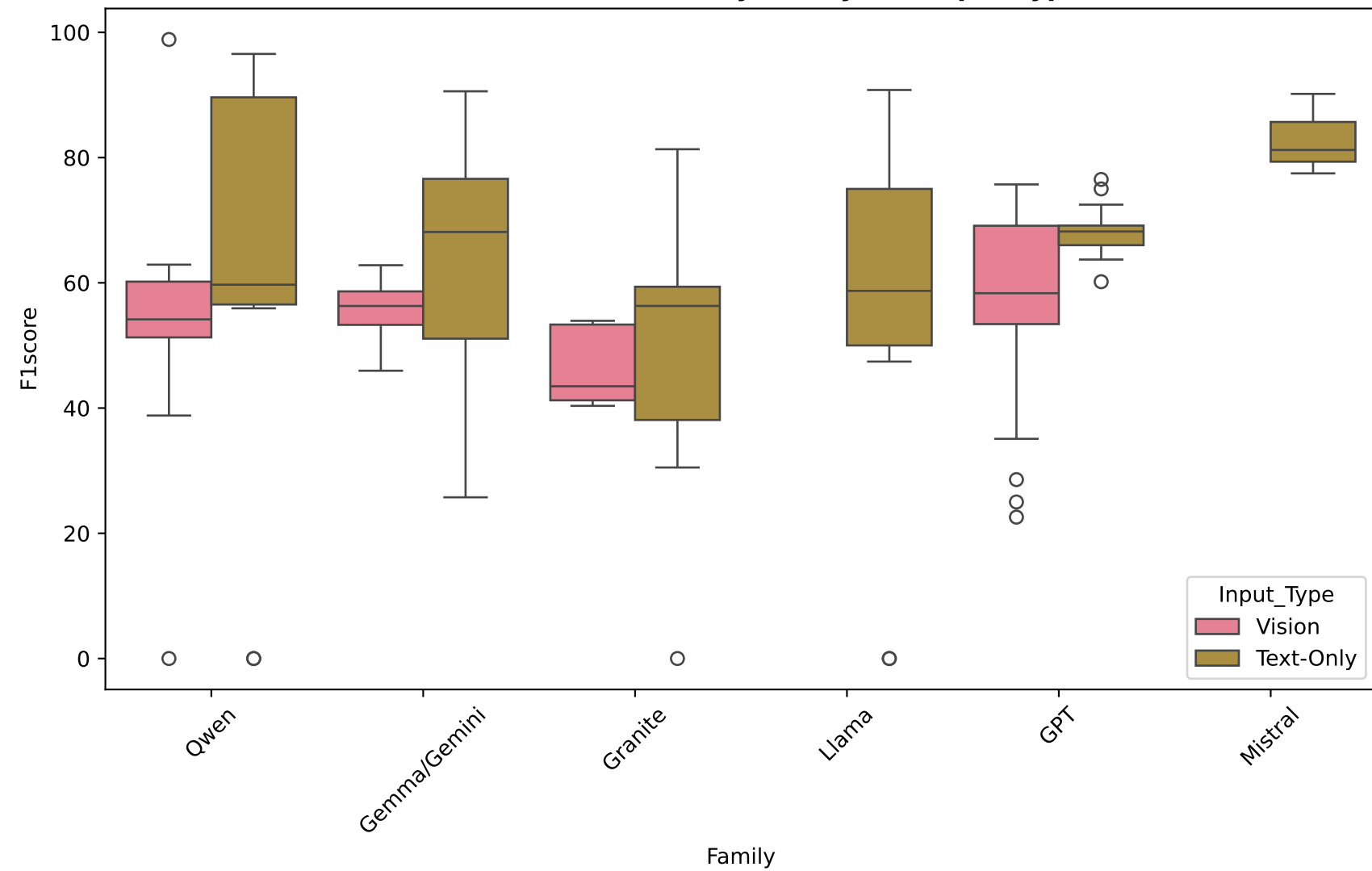
Average Performance Metrics by Model Family



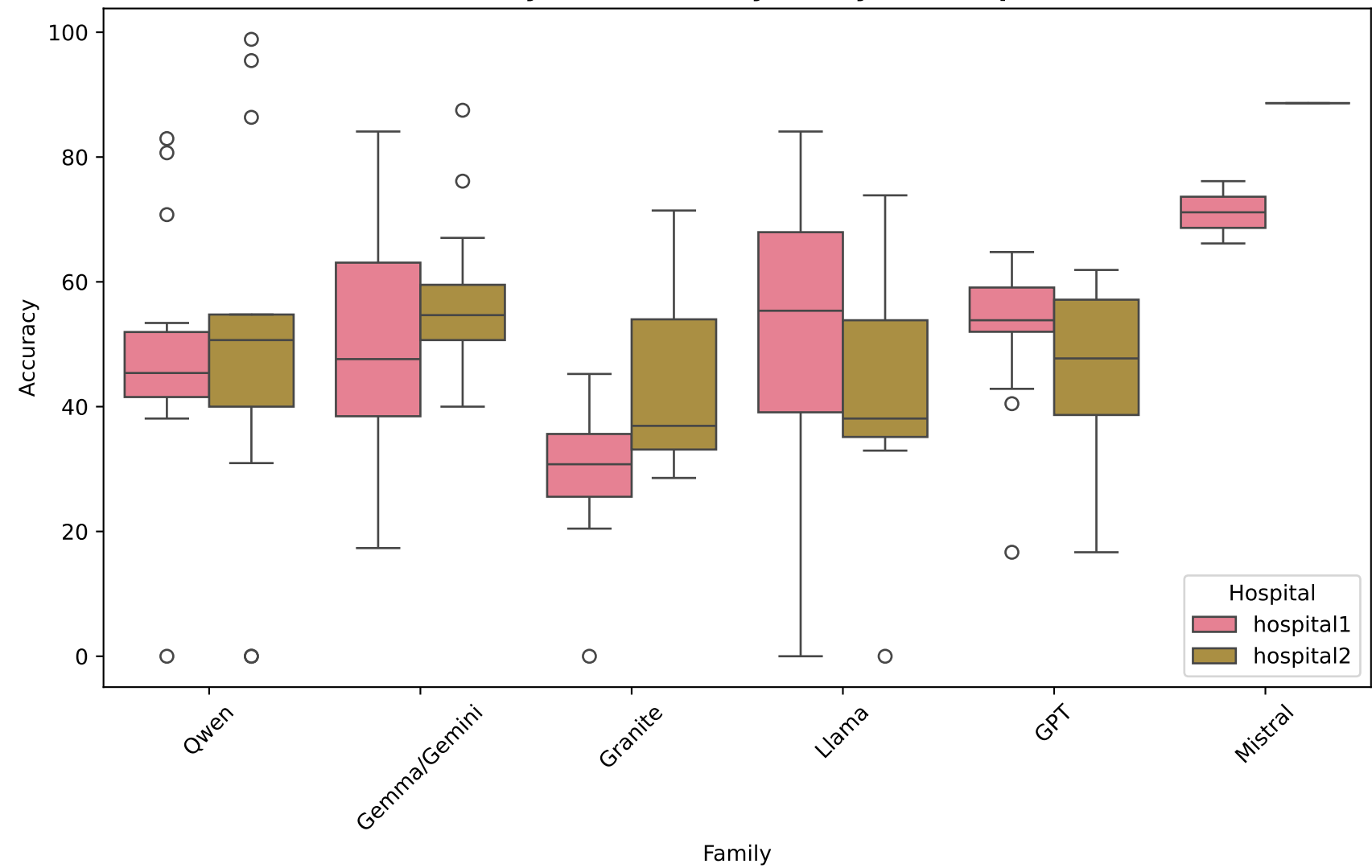
Precision vs Recall (colored by Accuracy)



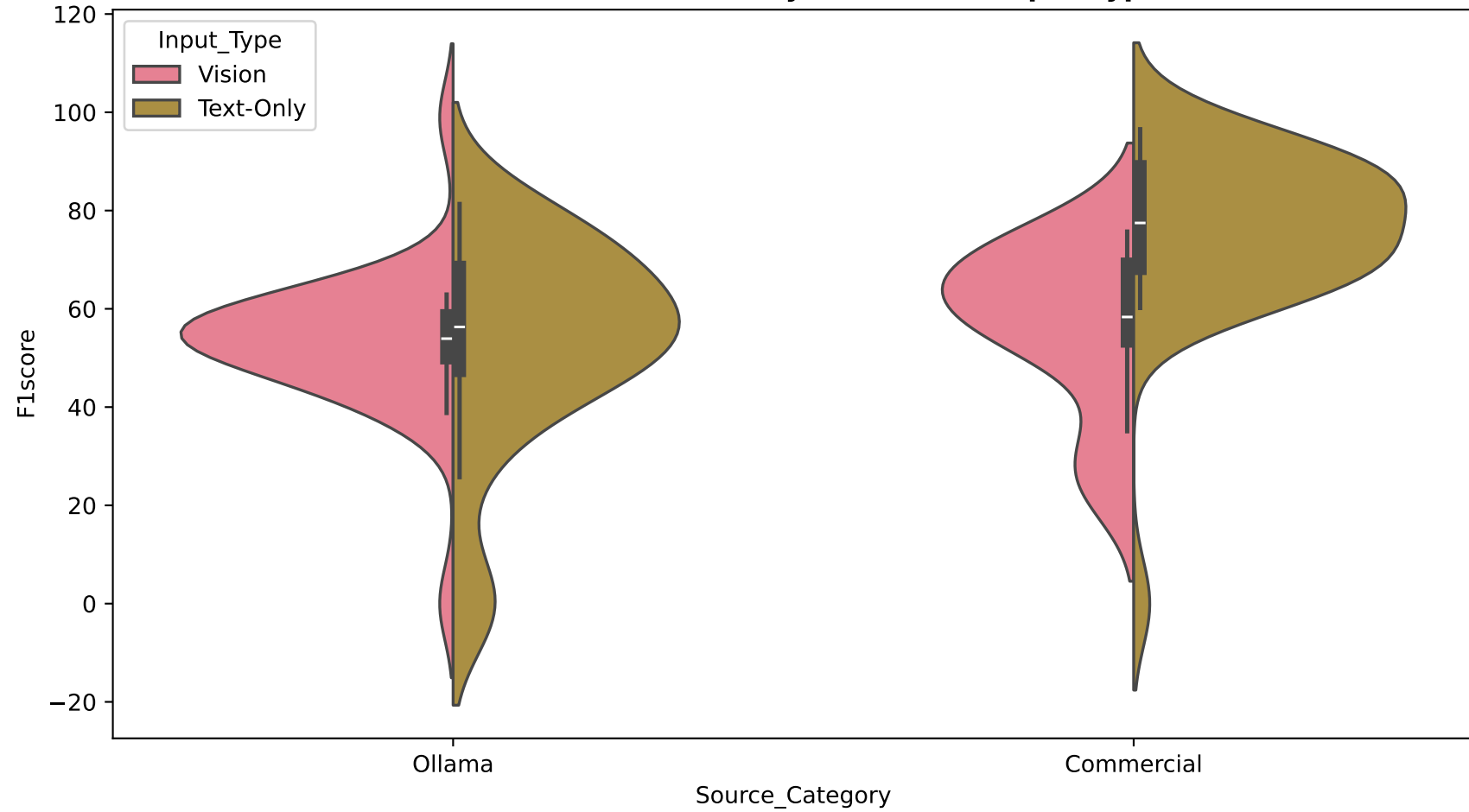
F1 Score Distribution by Family and Input Type



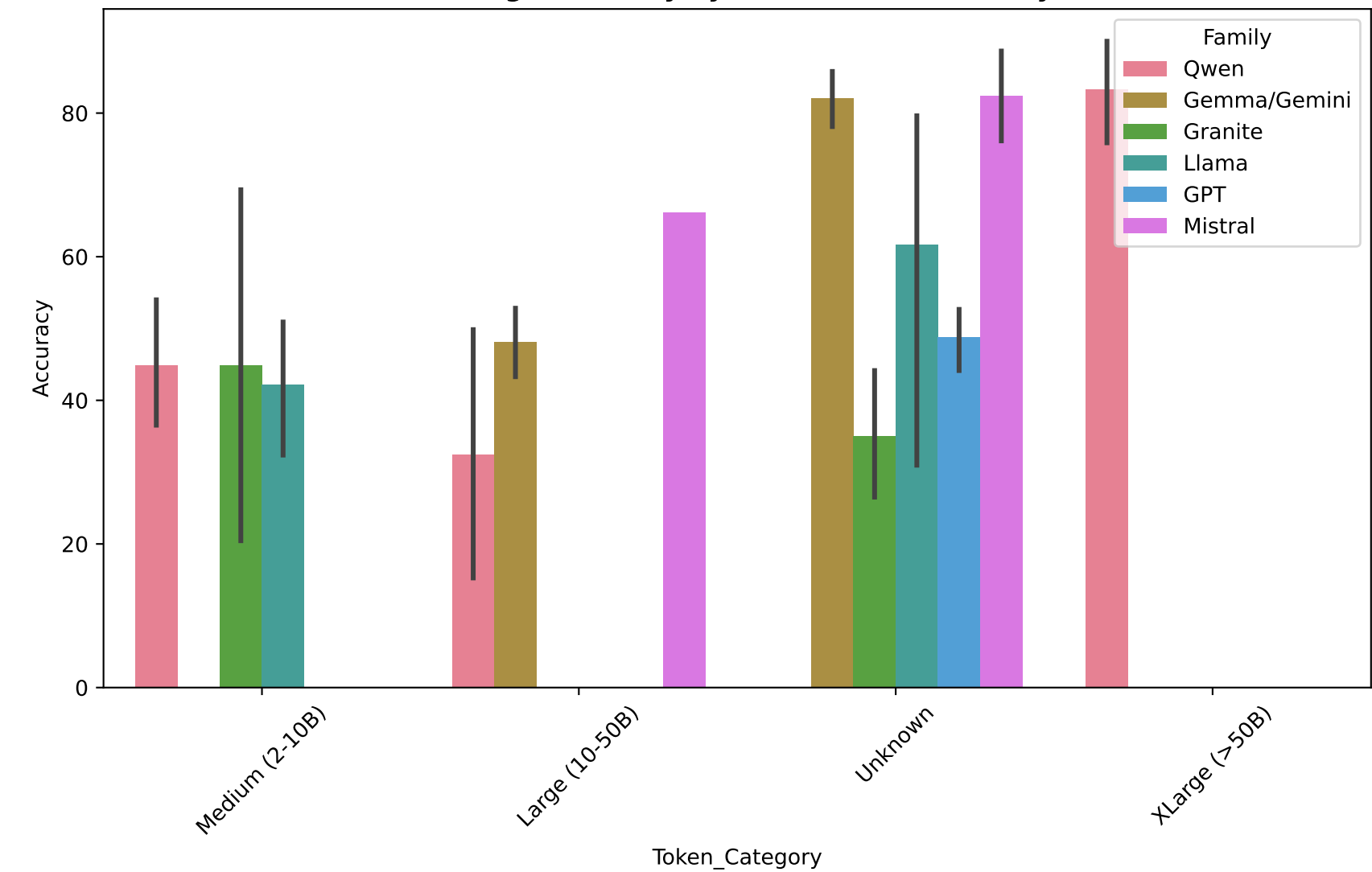
Accuracy Distribution by Family and Hospital



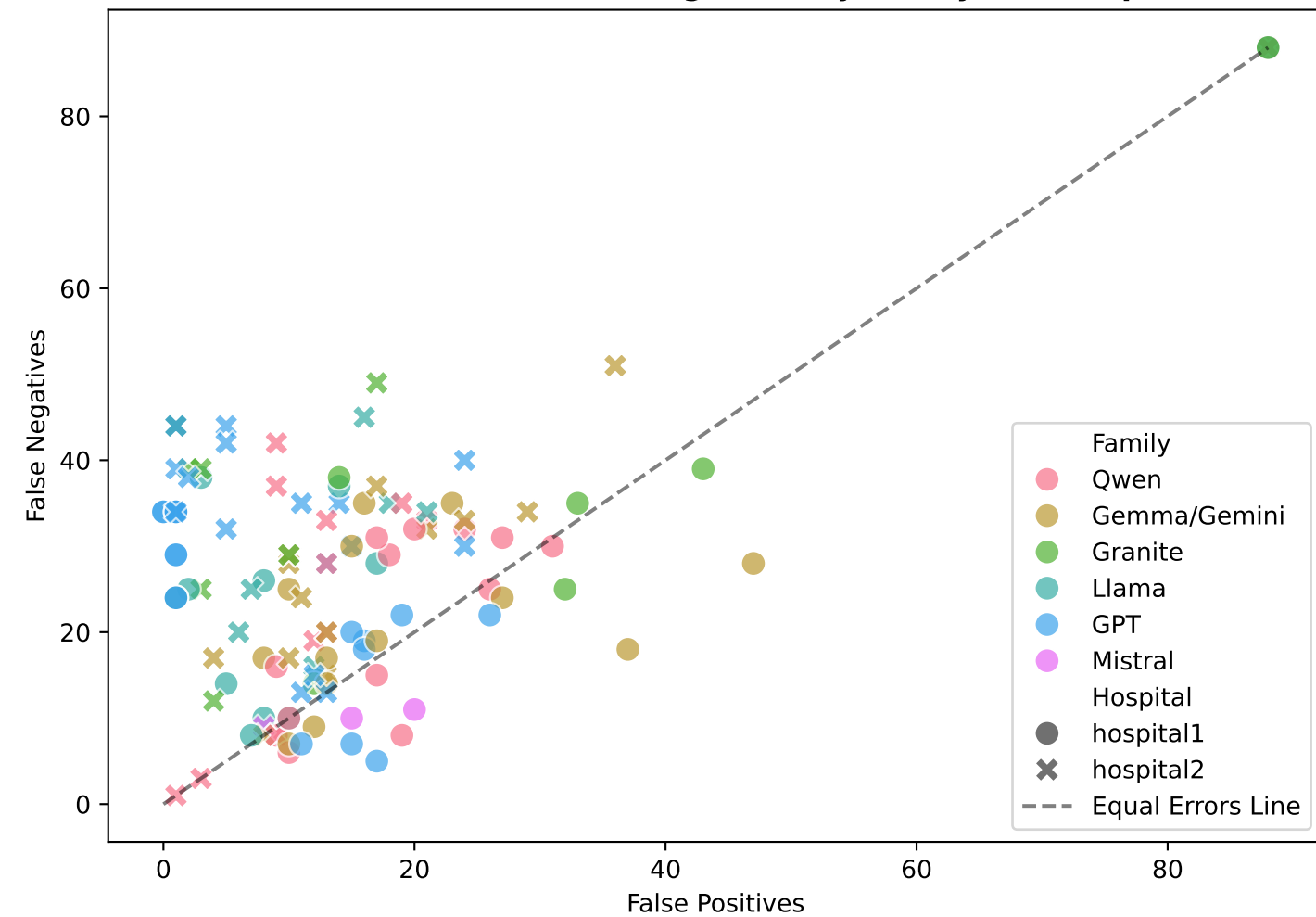
F1 Score Distribution by Source and Input Type



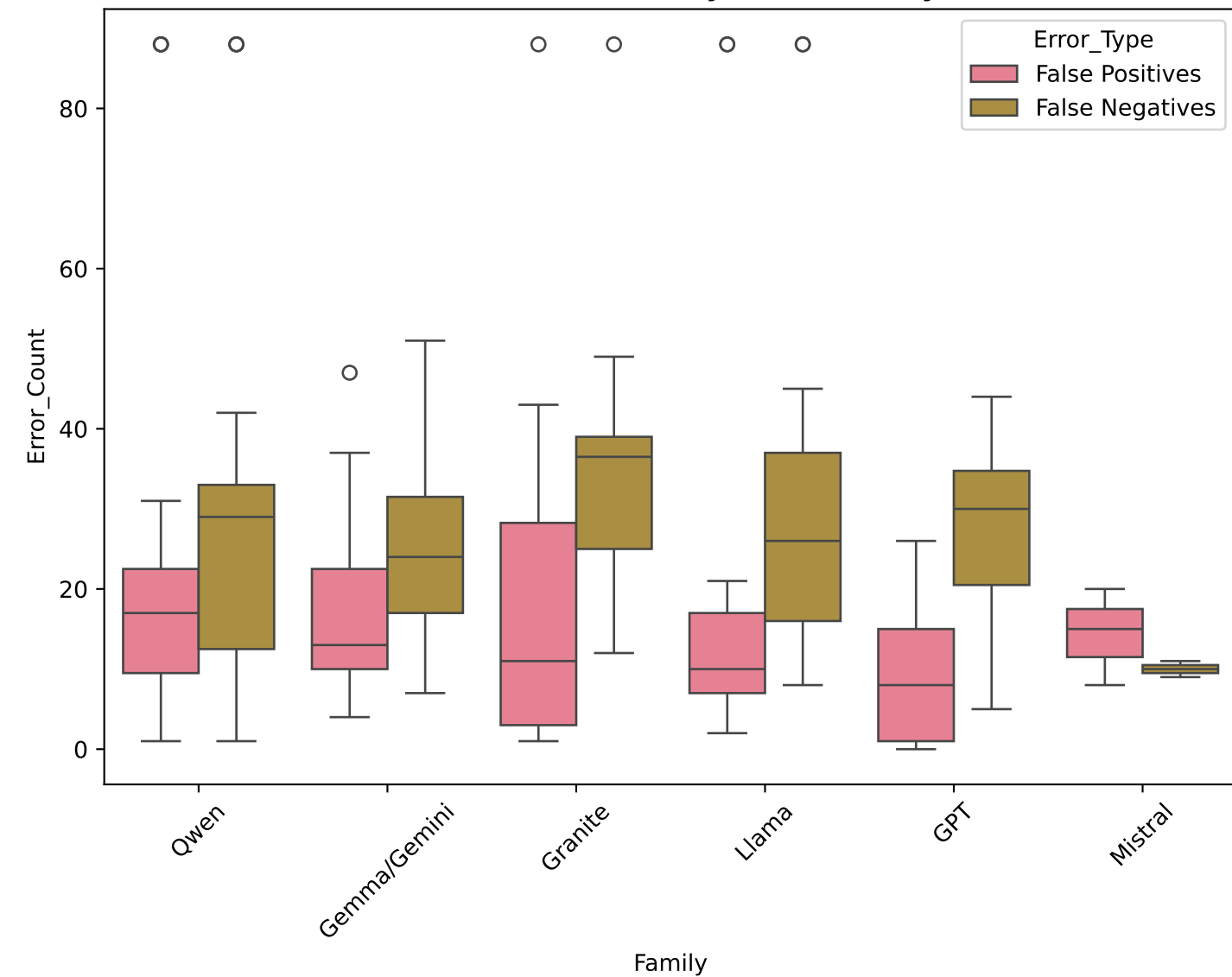
Average Accuracy by Token Size and Family



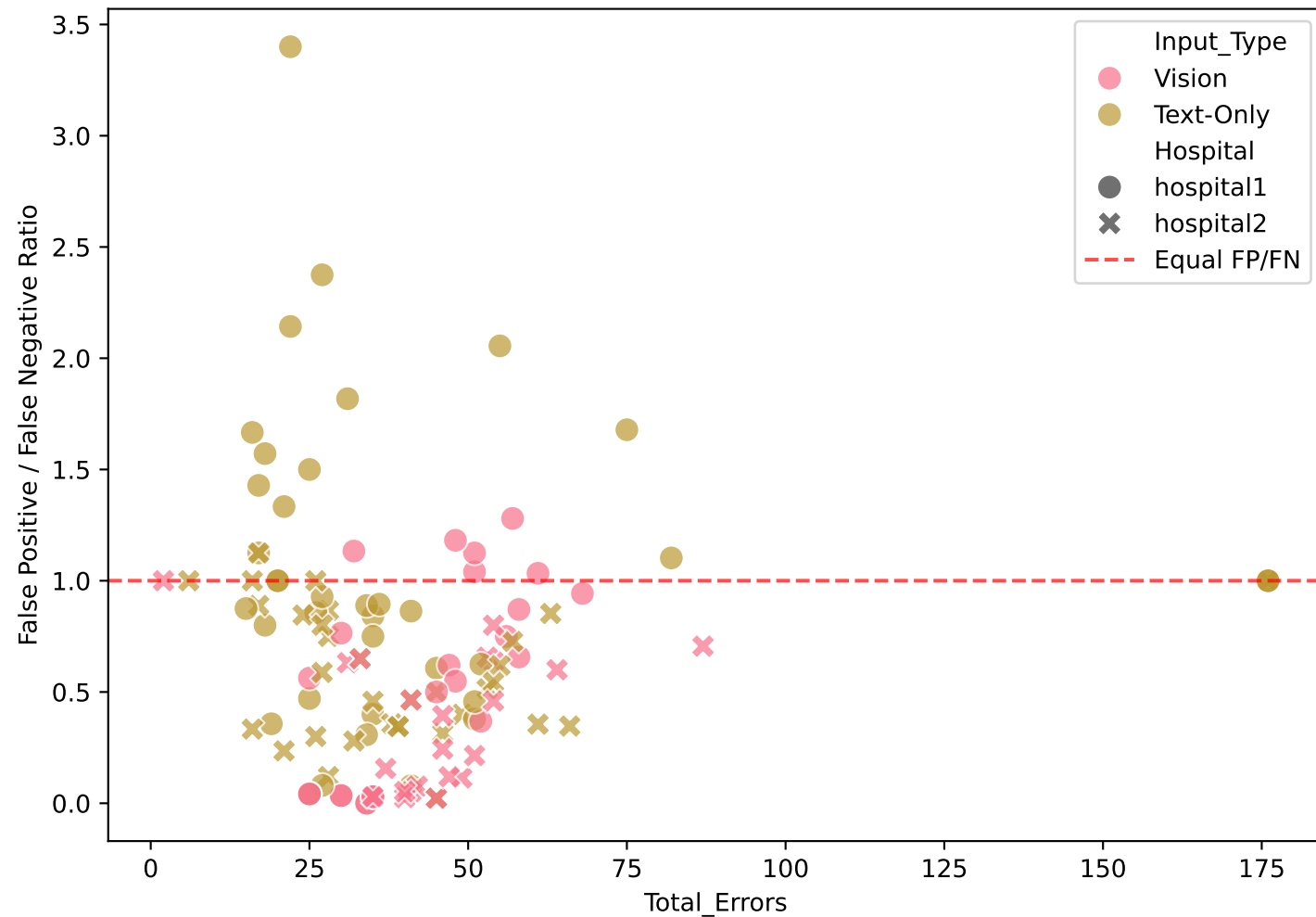
False Positives vs False Negatives by Family and Hospital



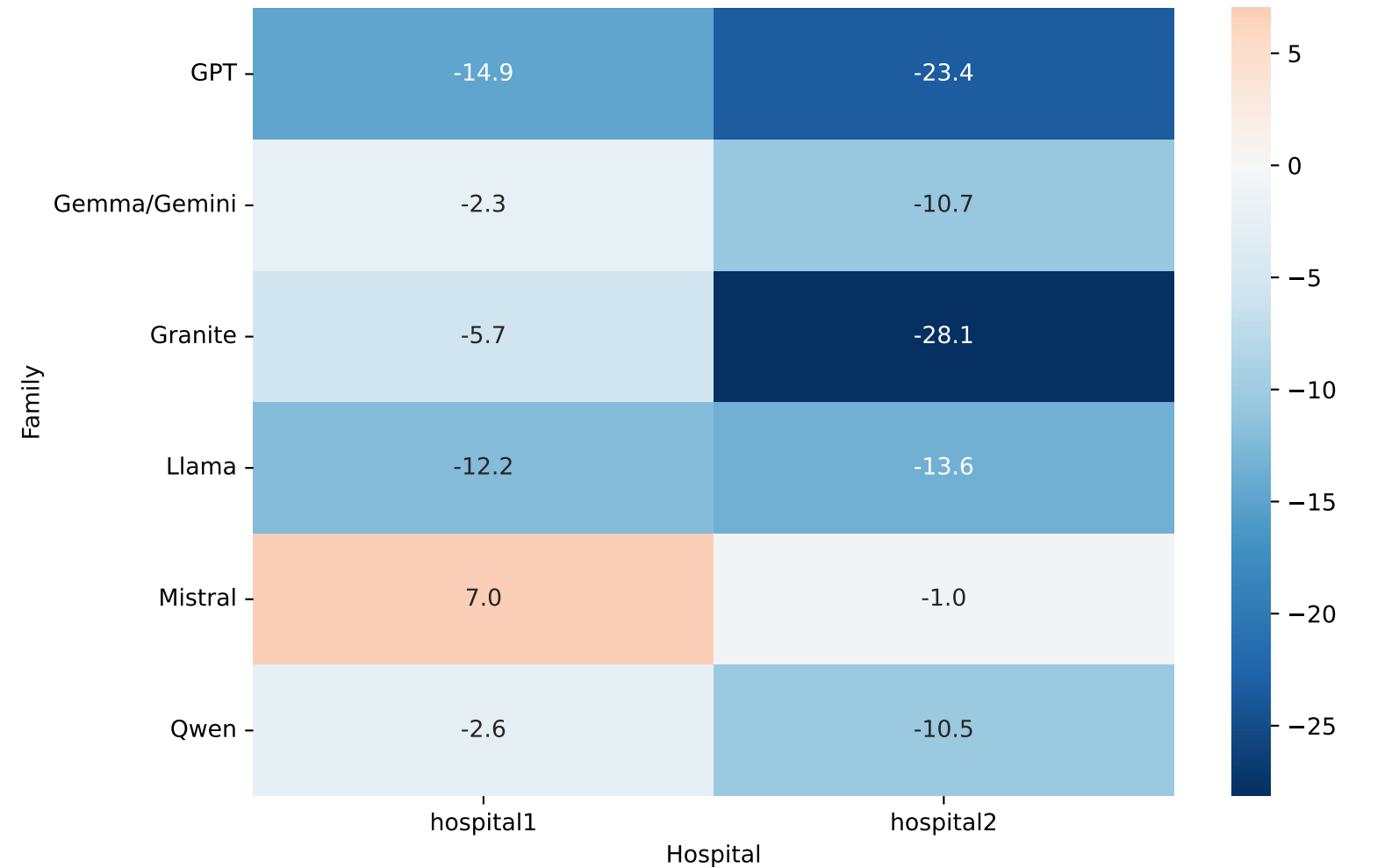
Error Distribution by Model Family



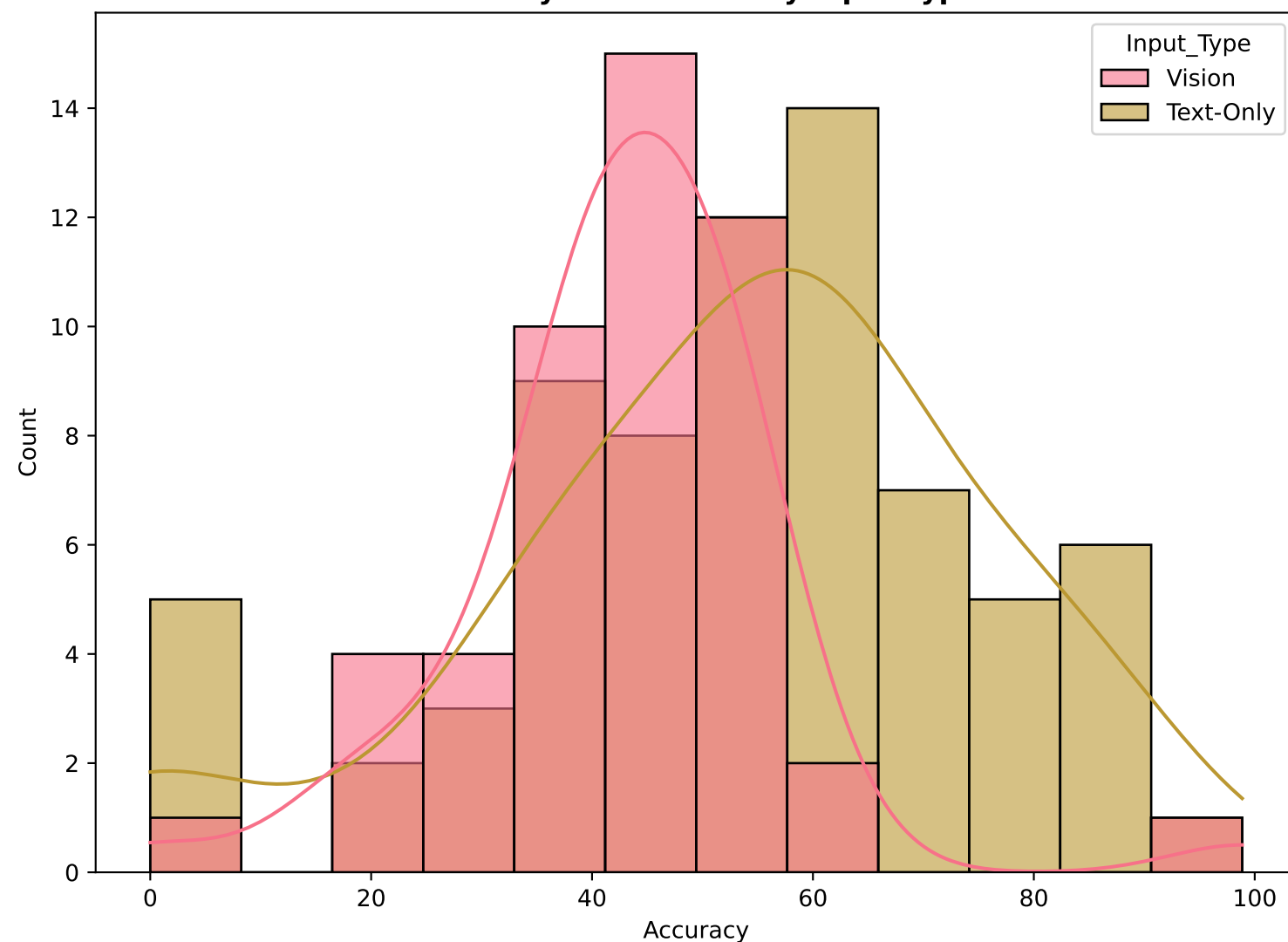
Error Ratio vs Total Errors
(Ratio = FP/FN)



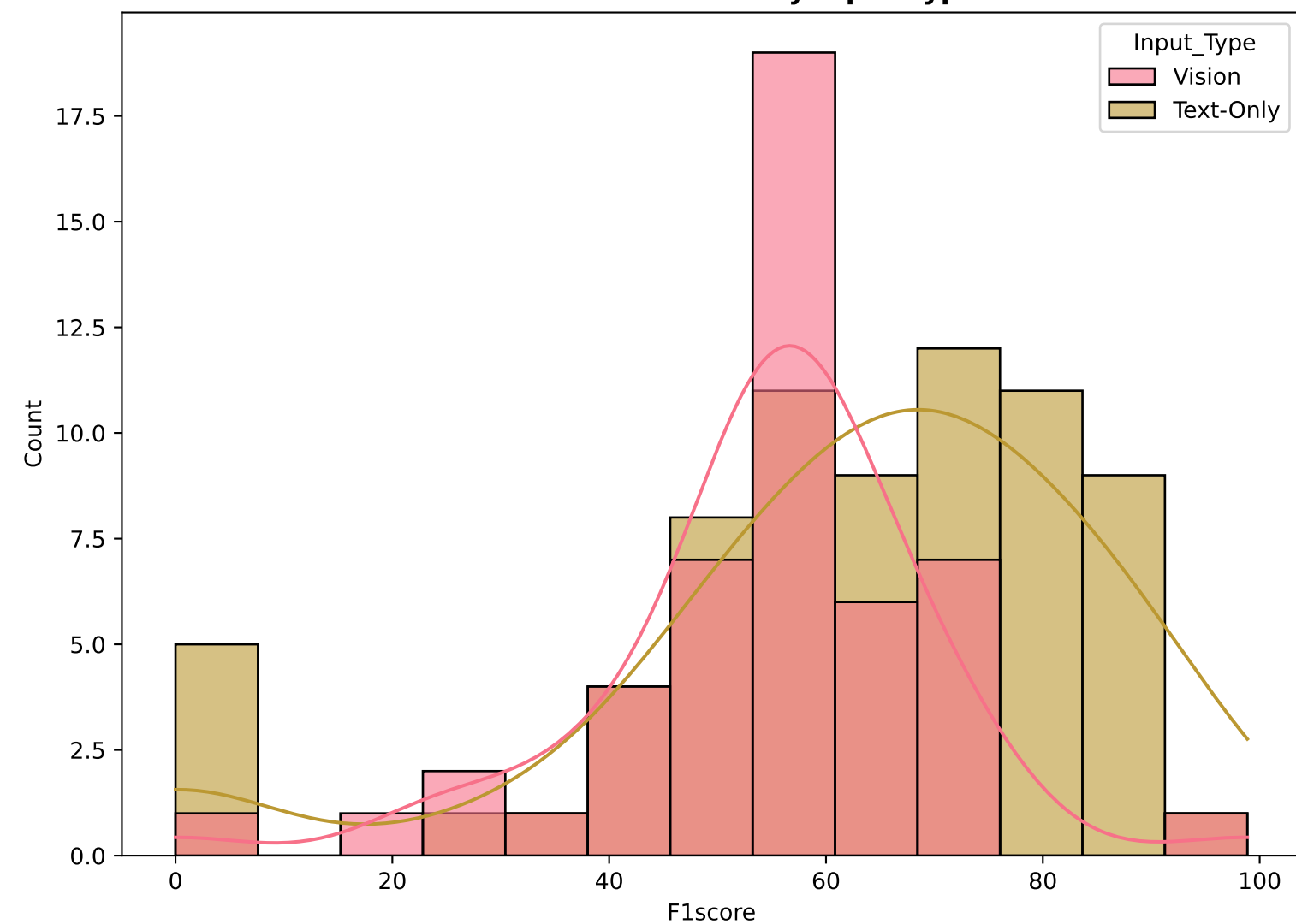
Error Bias: FP - FN by Family and Hospital
(+ve = More FP, -ve = More FN)



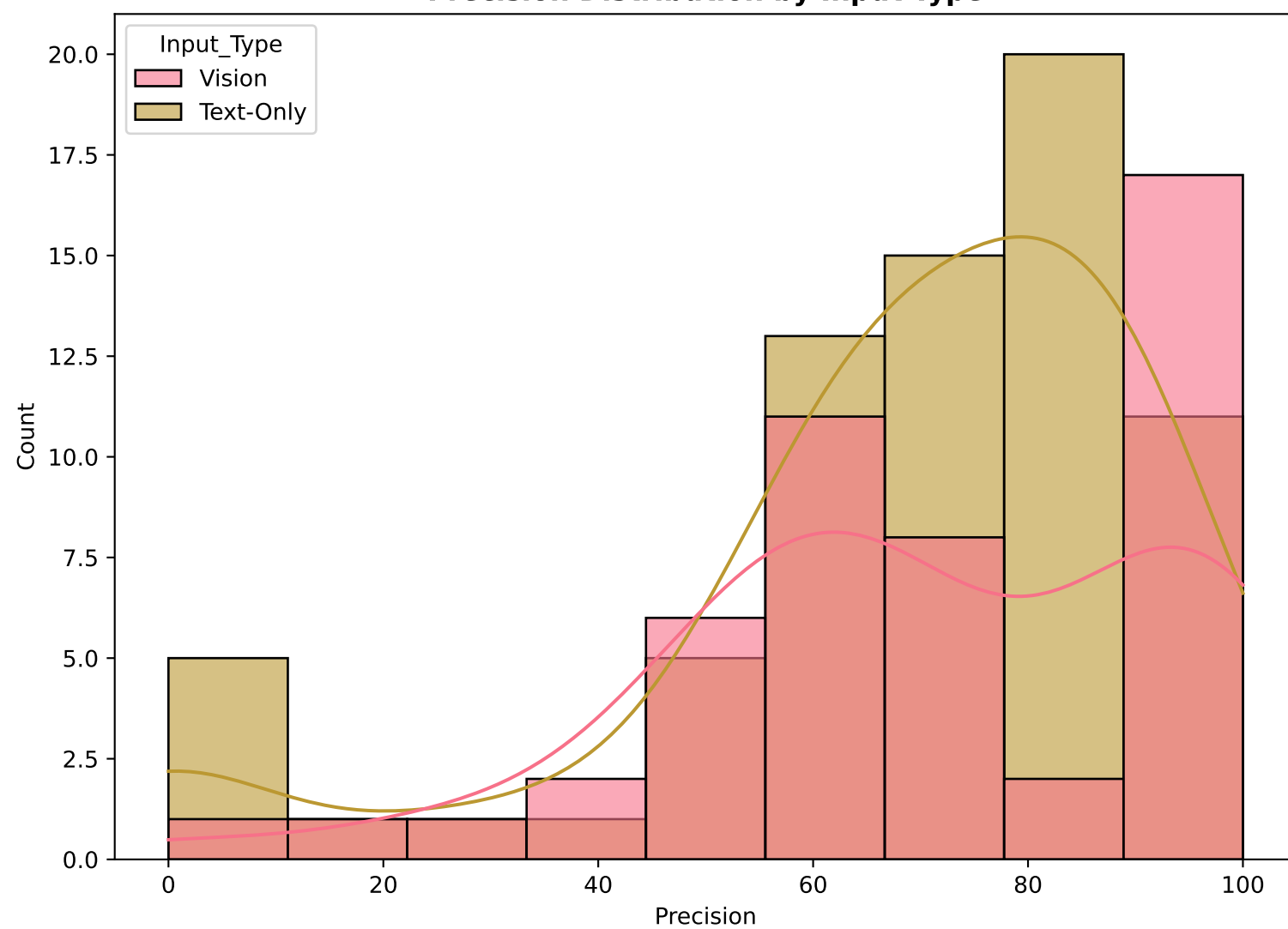
Accuracy Distribution by Input Type



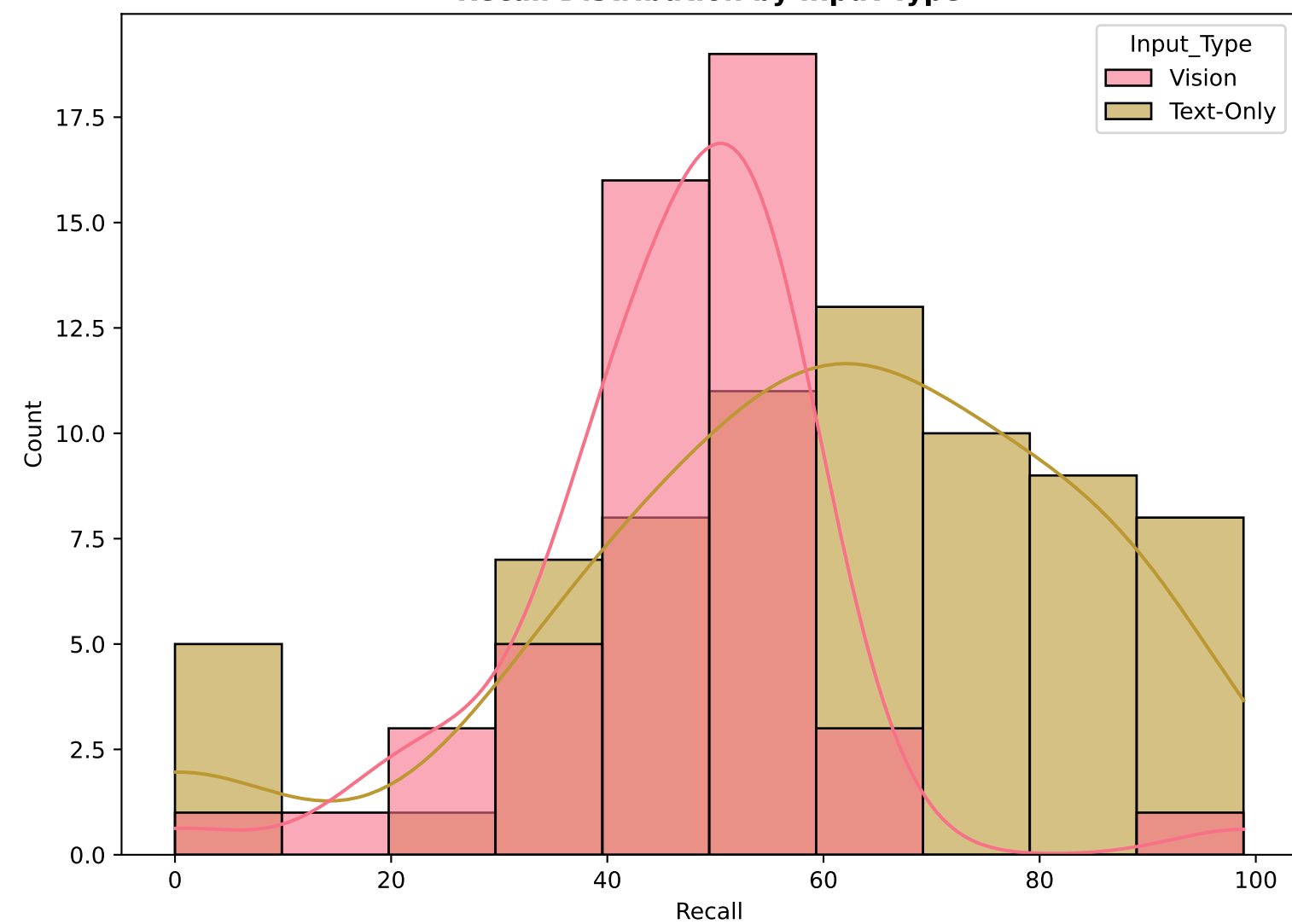
F1score Distribution by Input Type

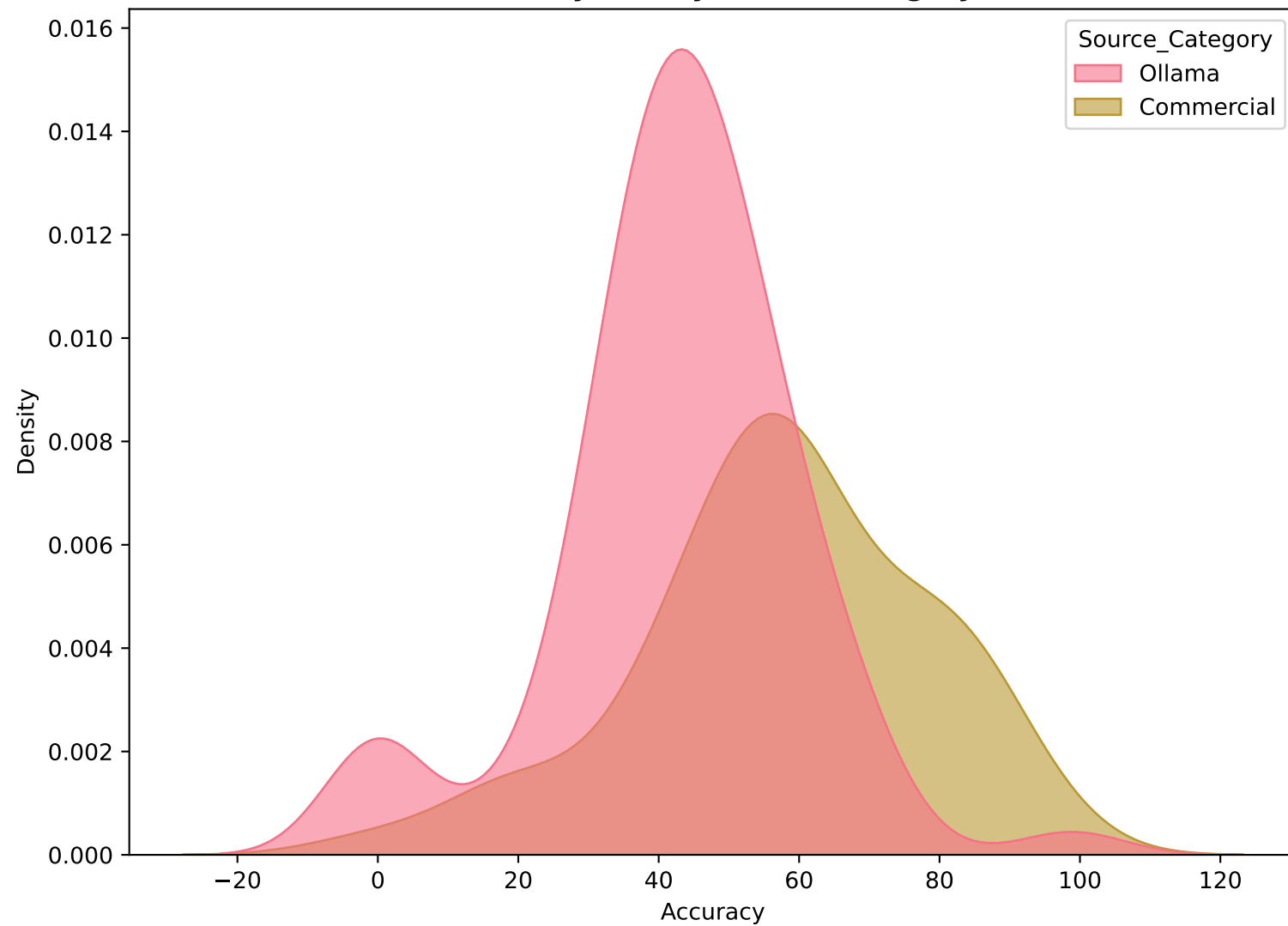
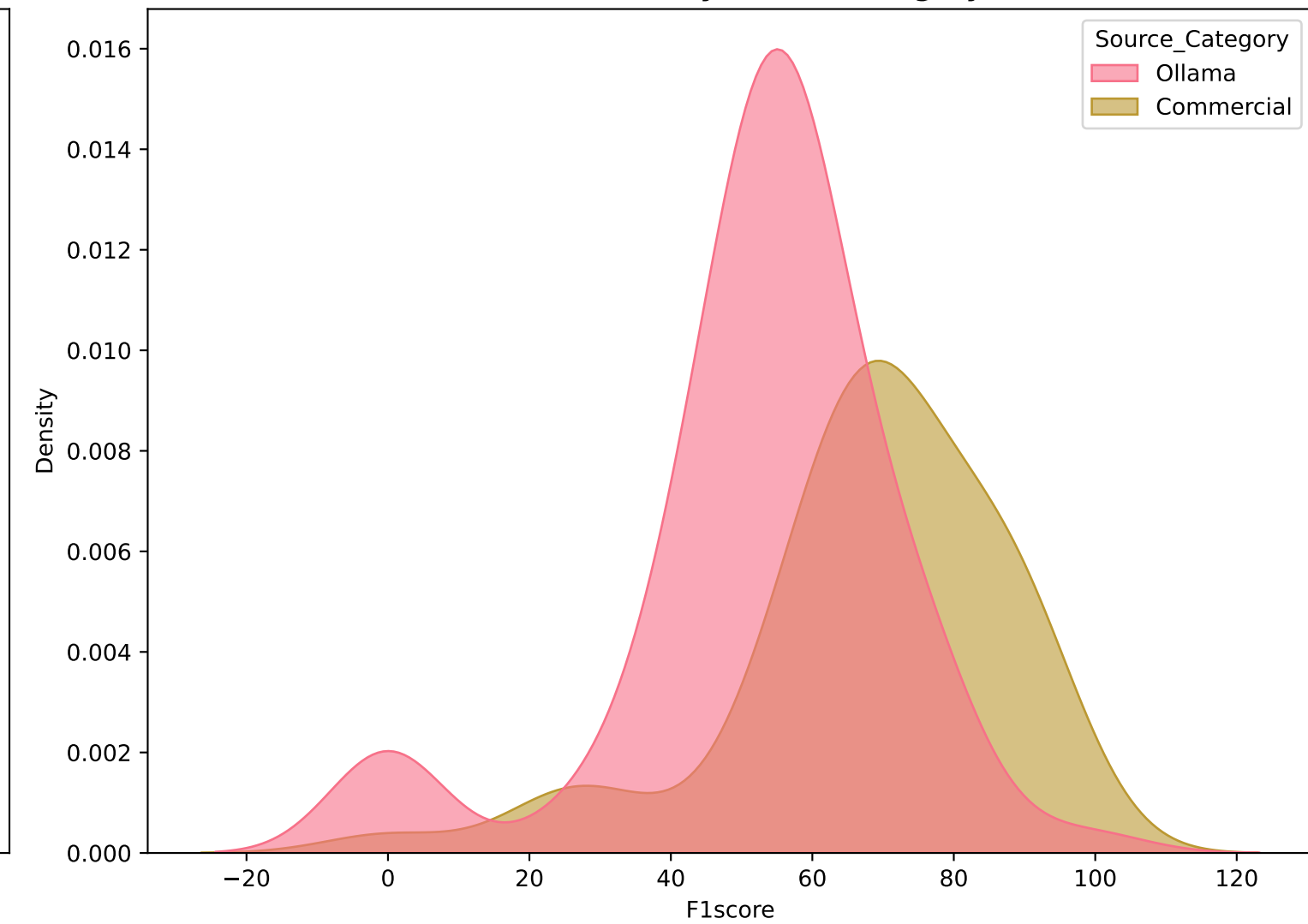
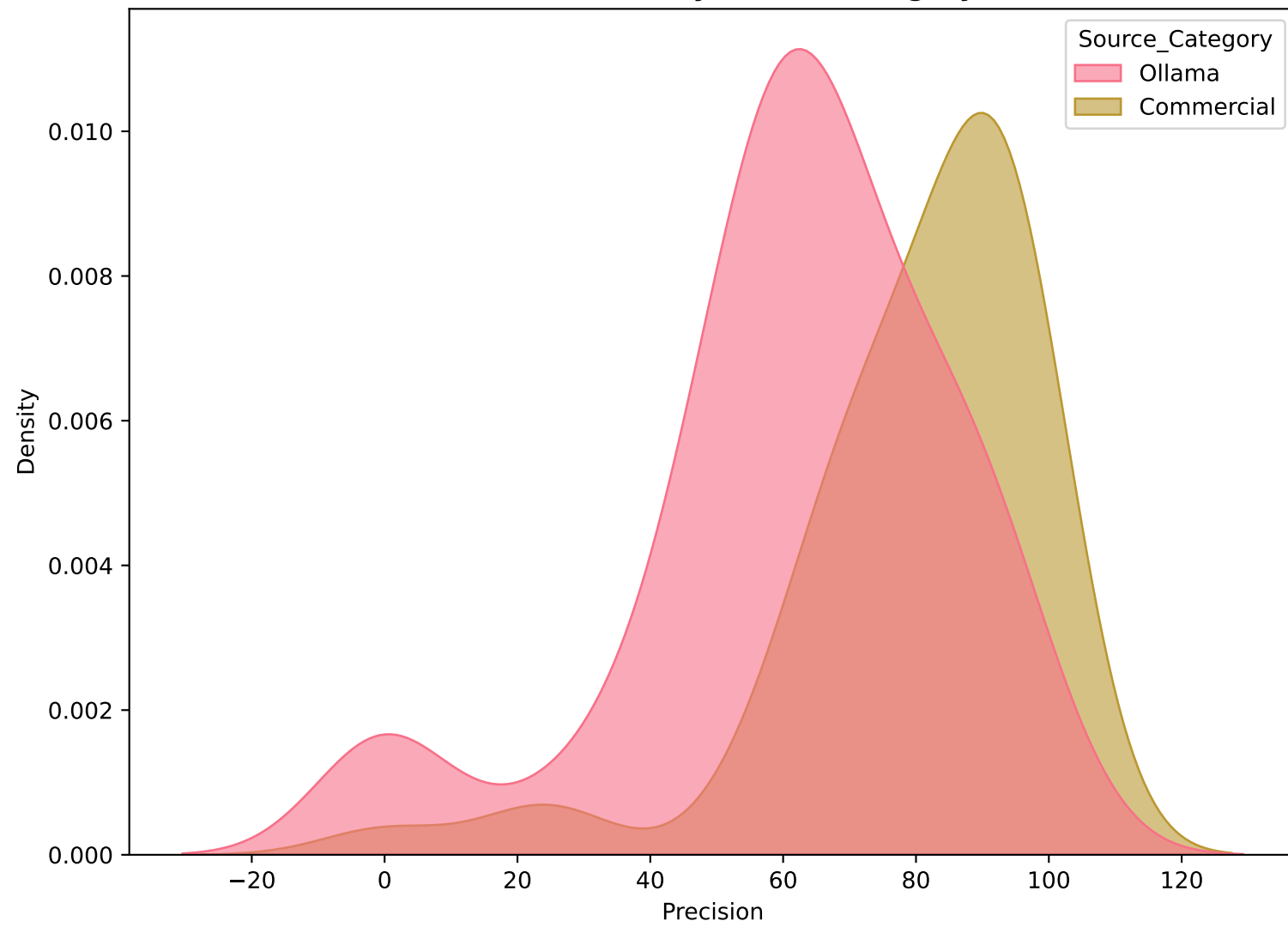
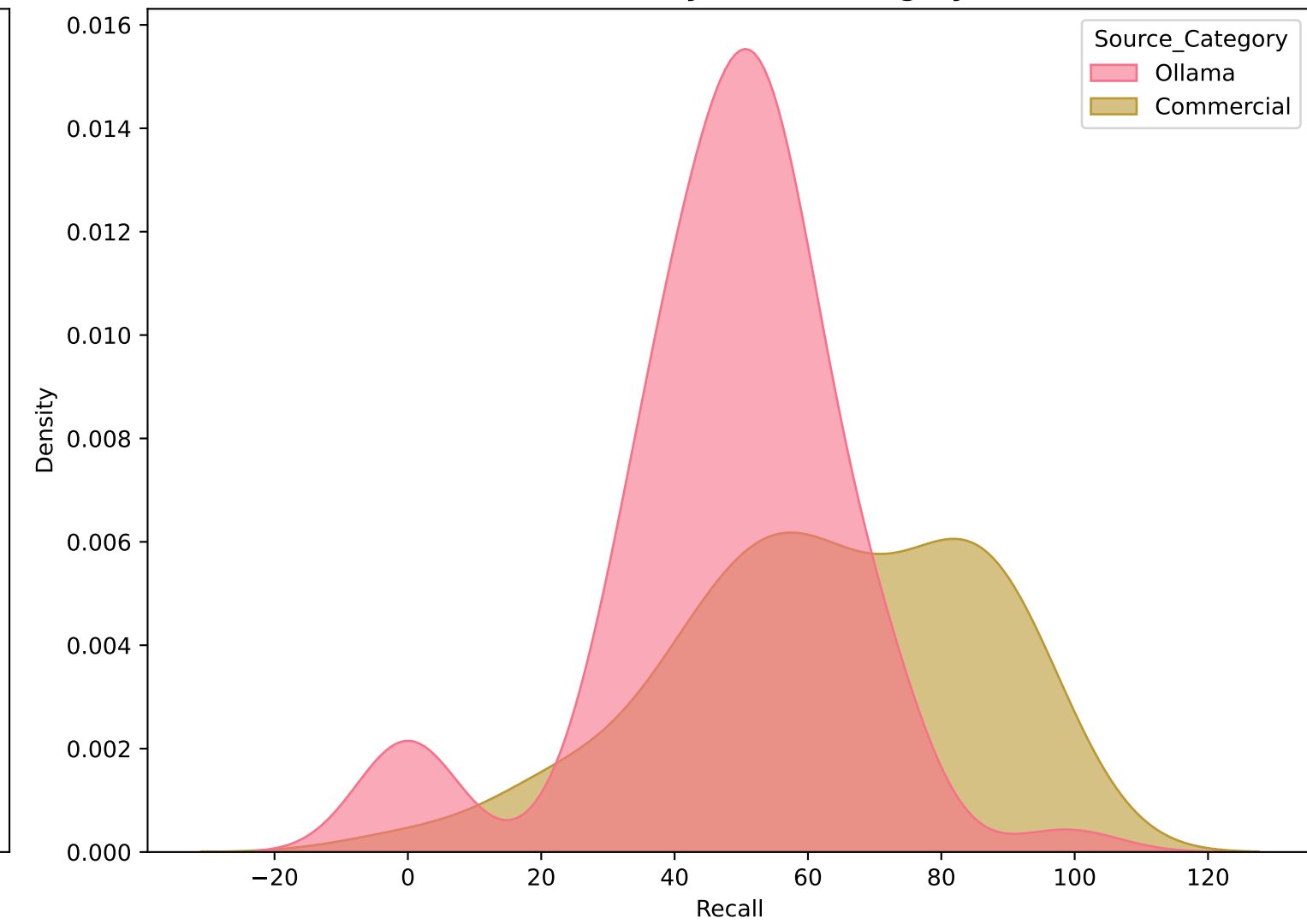


Precision Distribution by Input Type

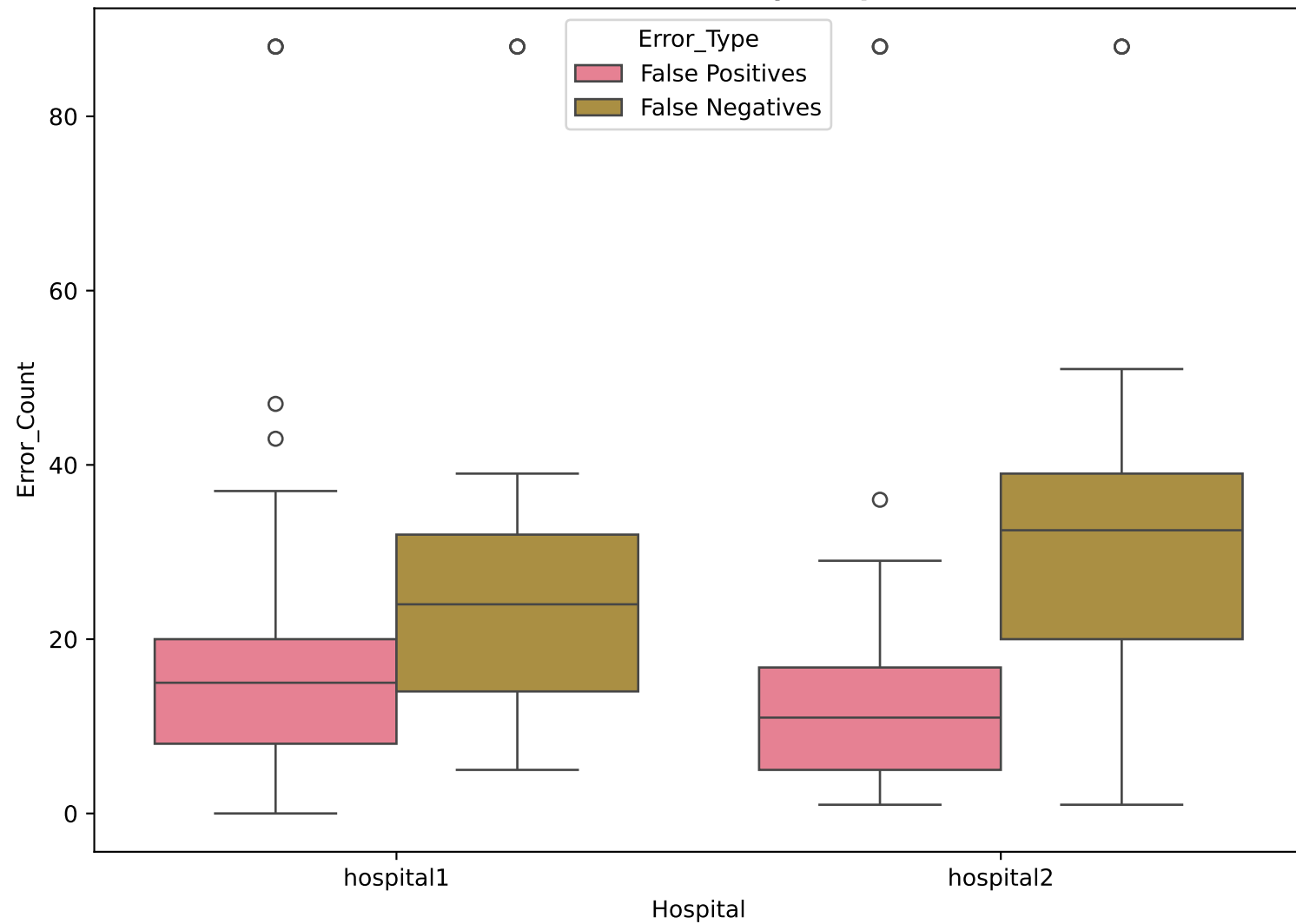


Recall Distribution by Input Type

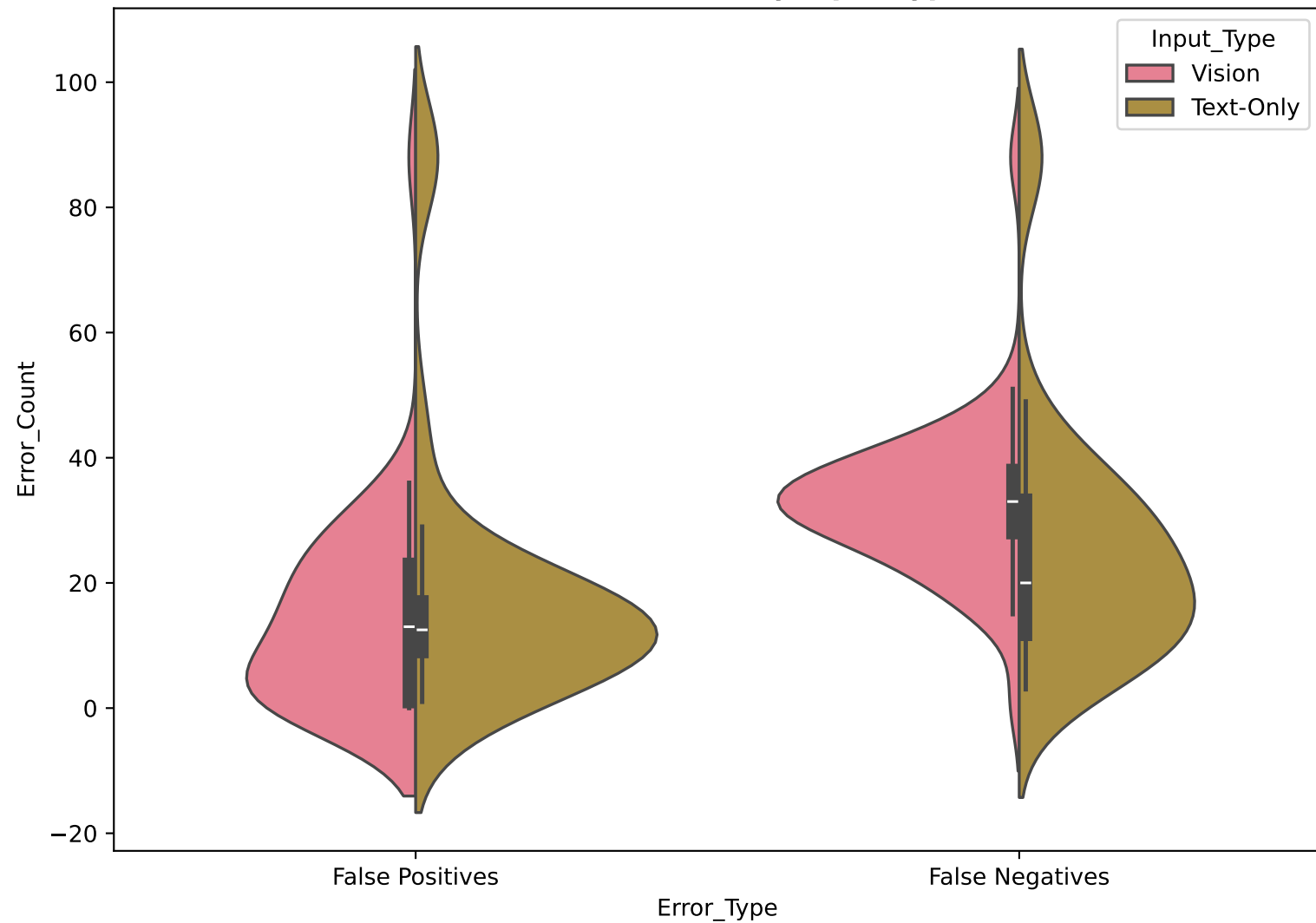


Accuracy KDE by Source Category**F1score KDE by Source Category****Precision KDE by Source Category****Recall KDE by Source Category**

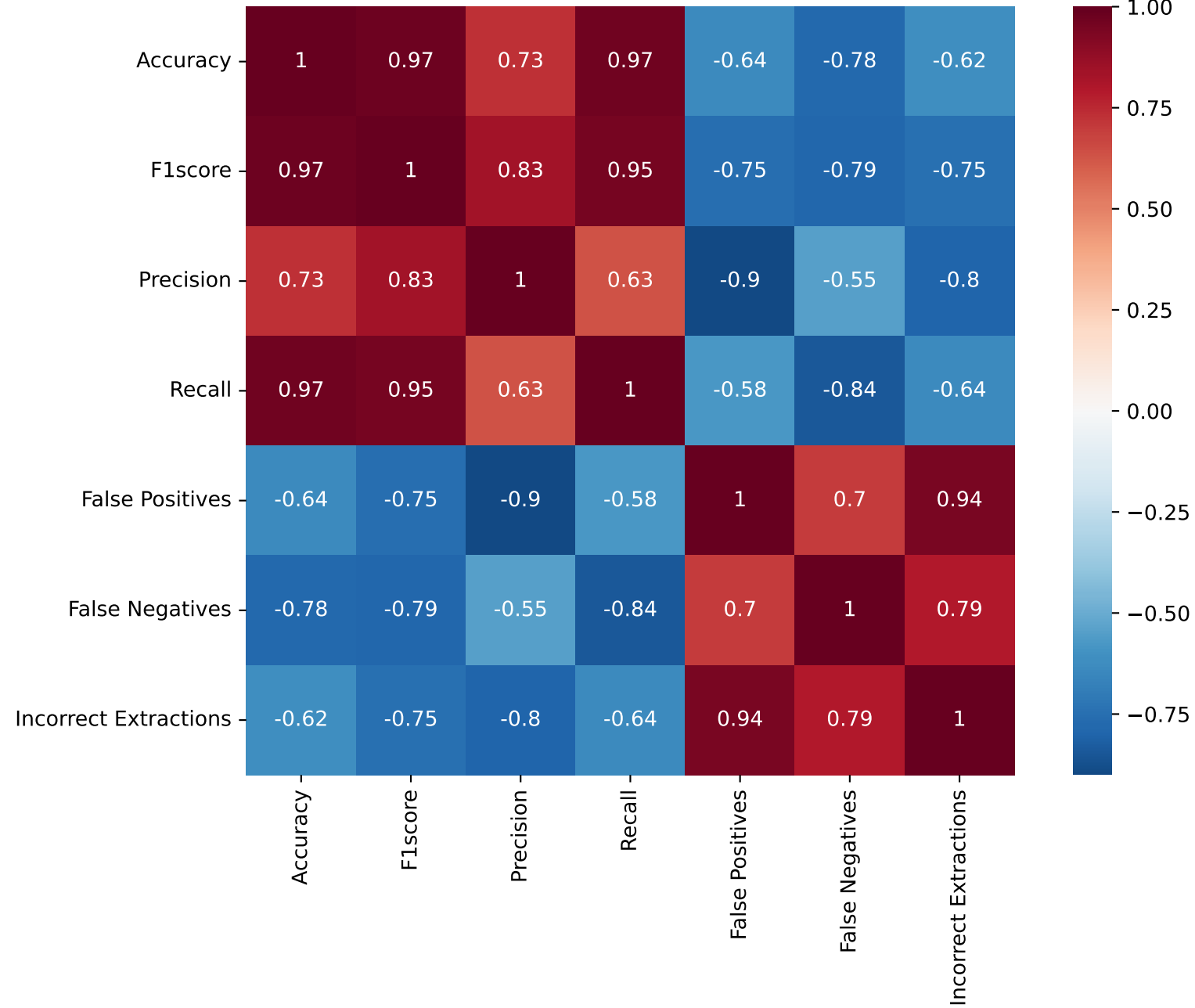
Error Distribution by Hospital



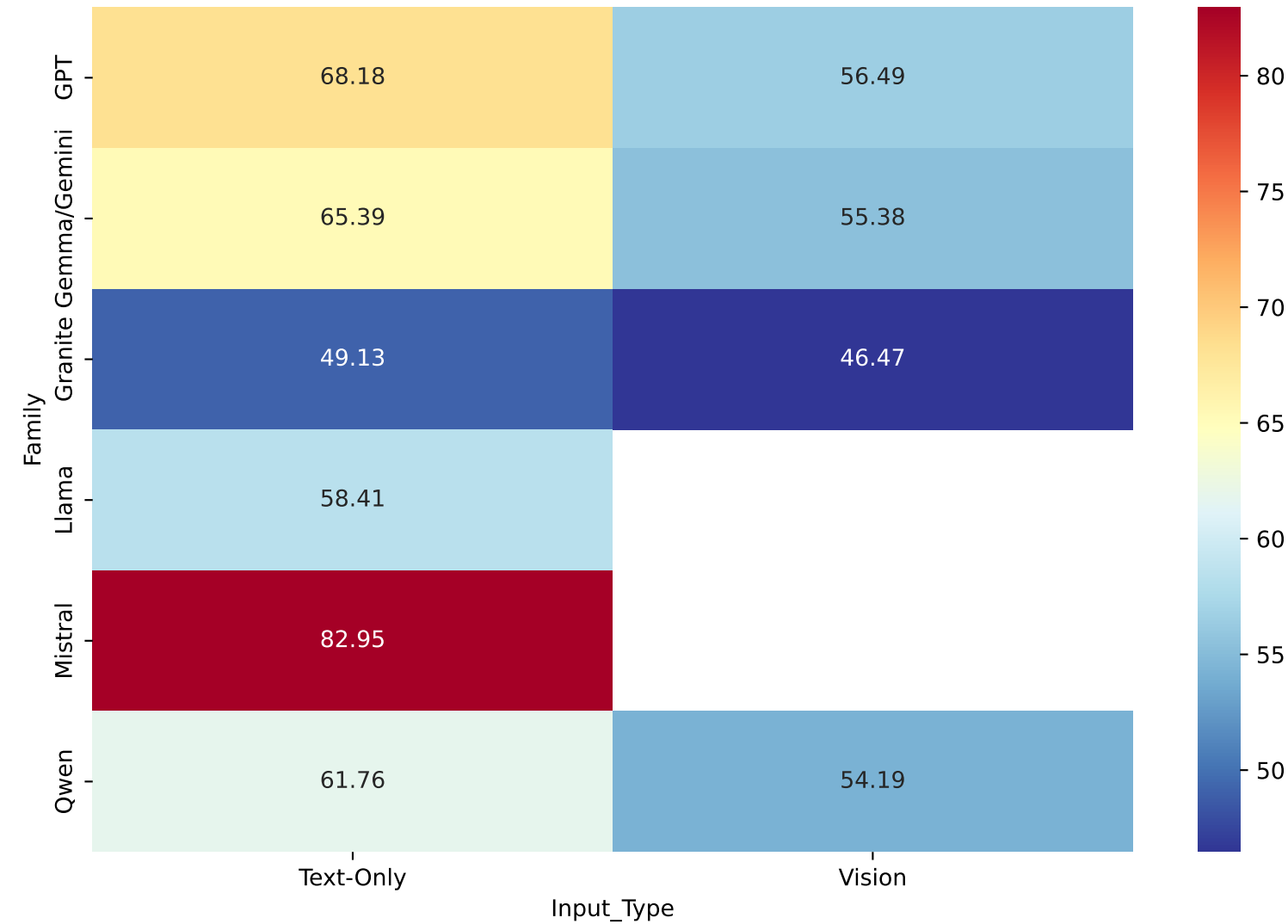
Error Distribution by Input Type



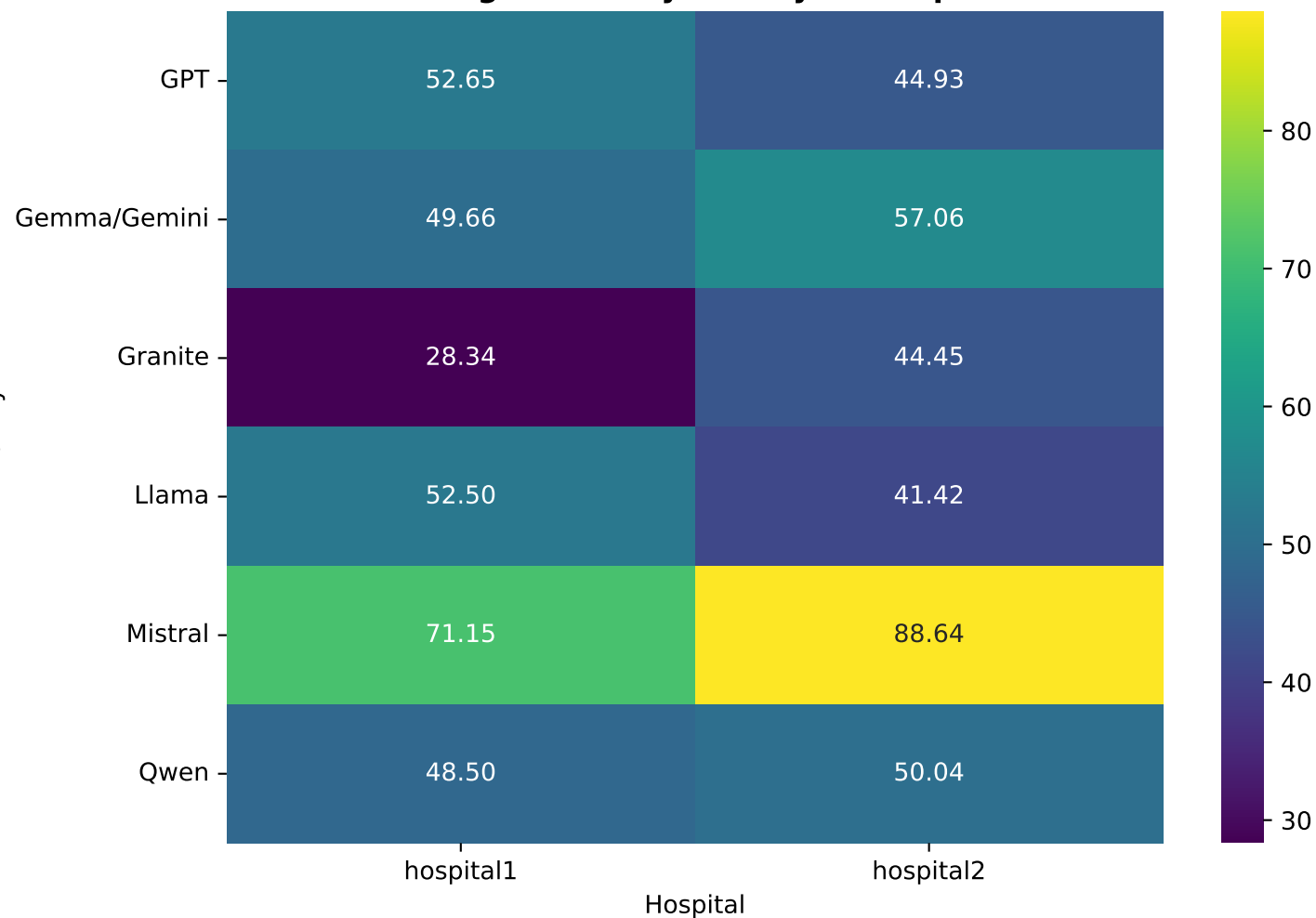
Performance Metrics Correlation Matrix



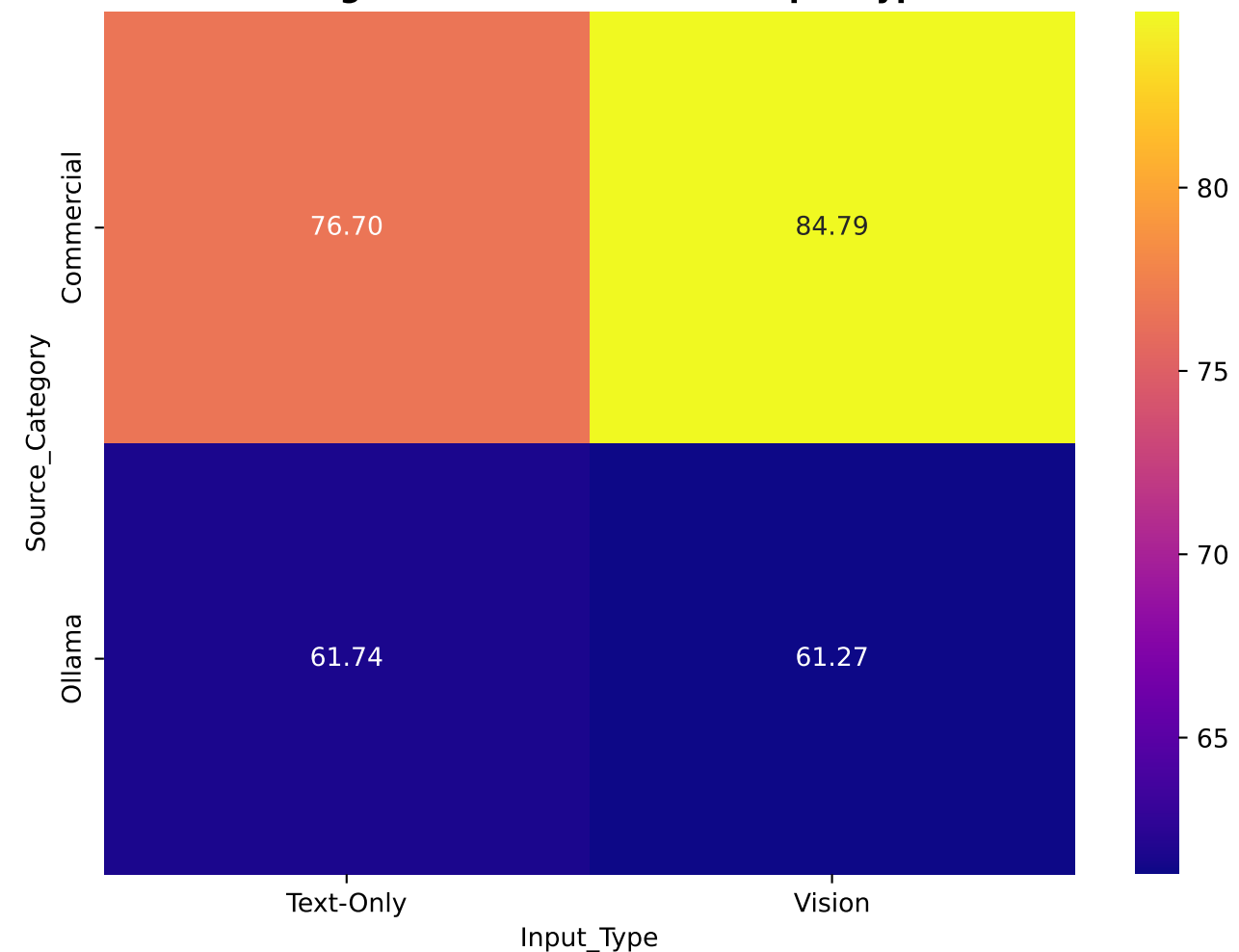
Average F1 Score: Family vs Input Type



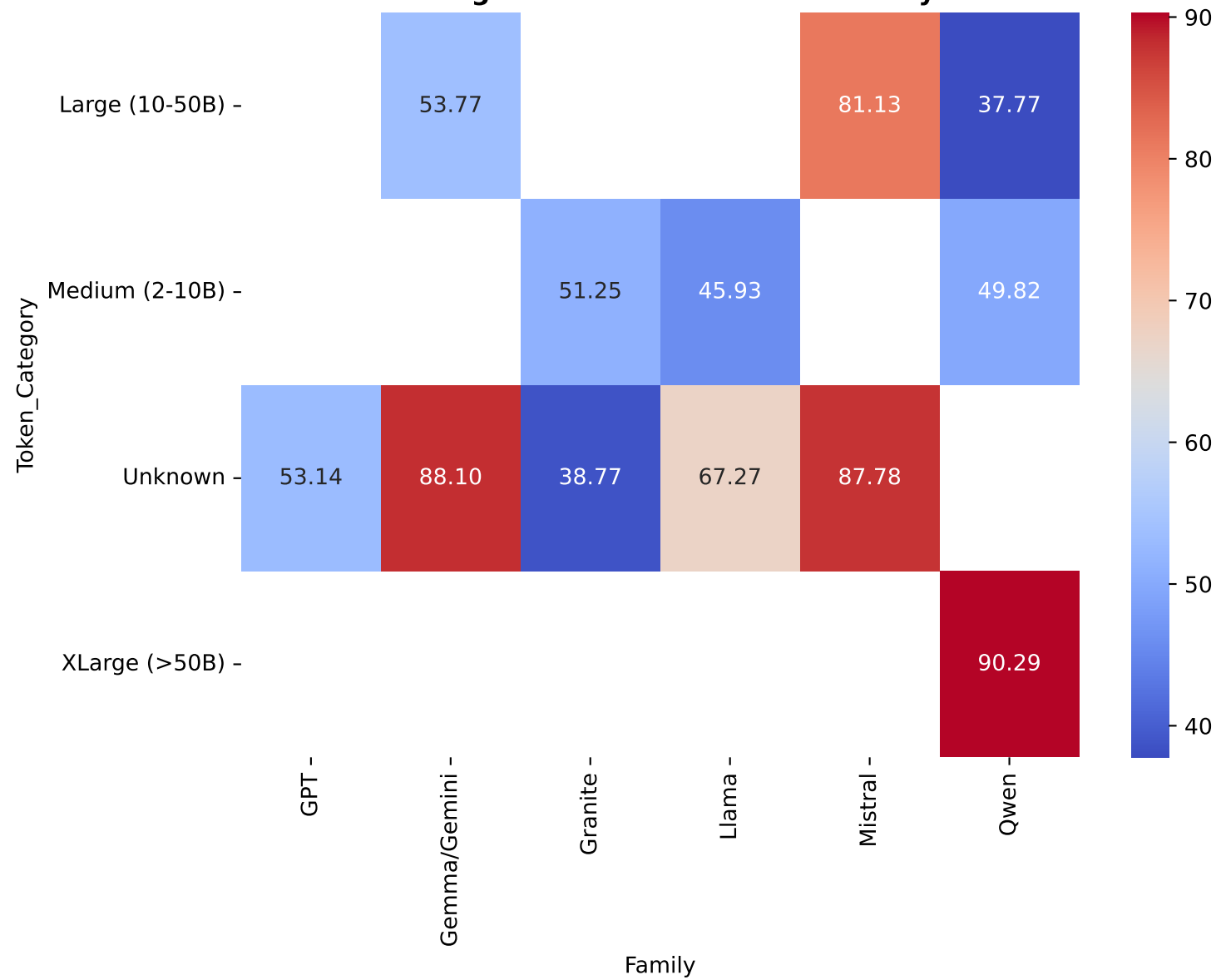
Average Accuracy: Family vs Hospital



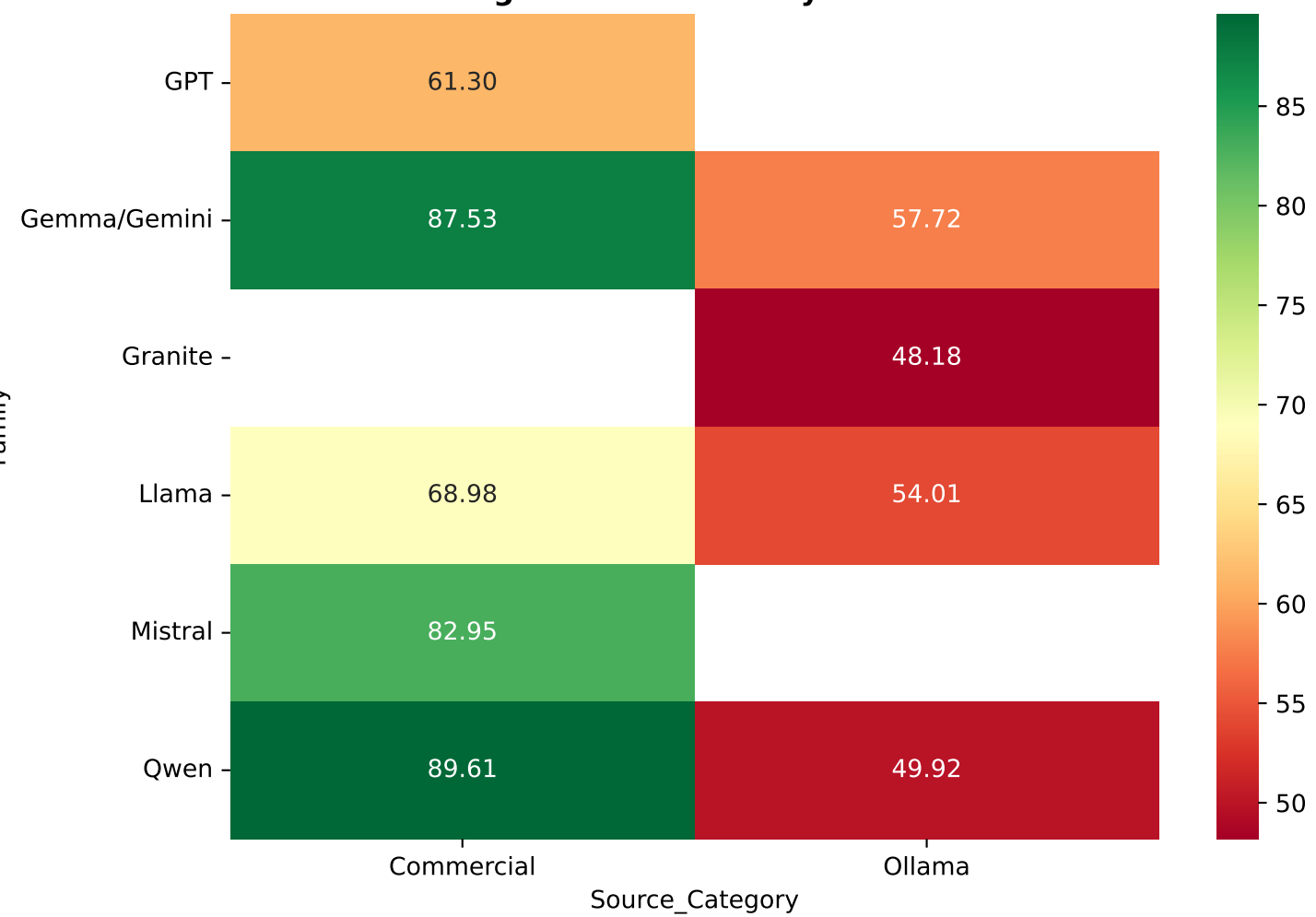
Average Precision: Source vs Input Type



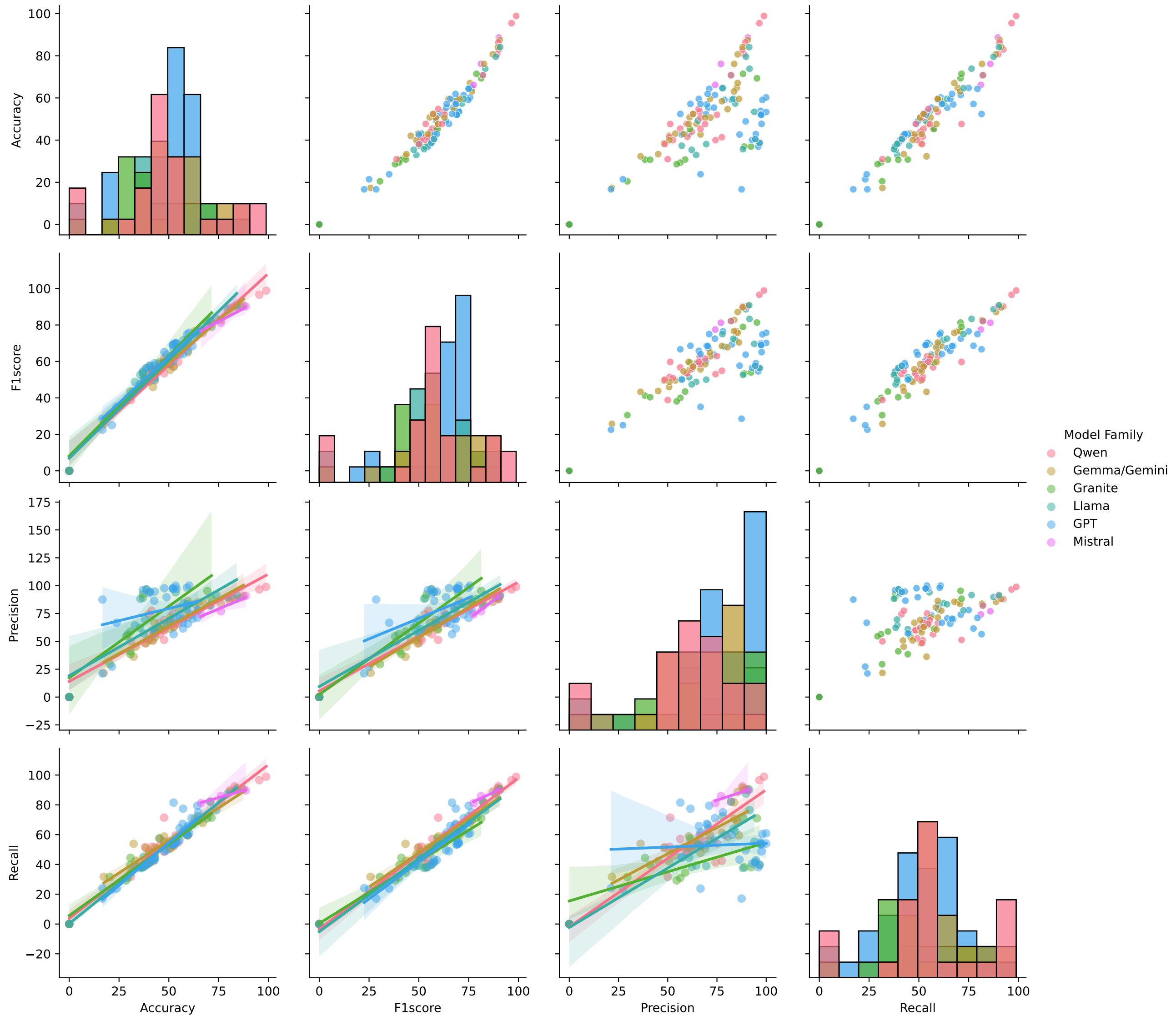
Average Recall: Token Size vs Family



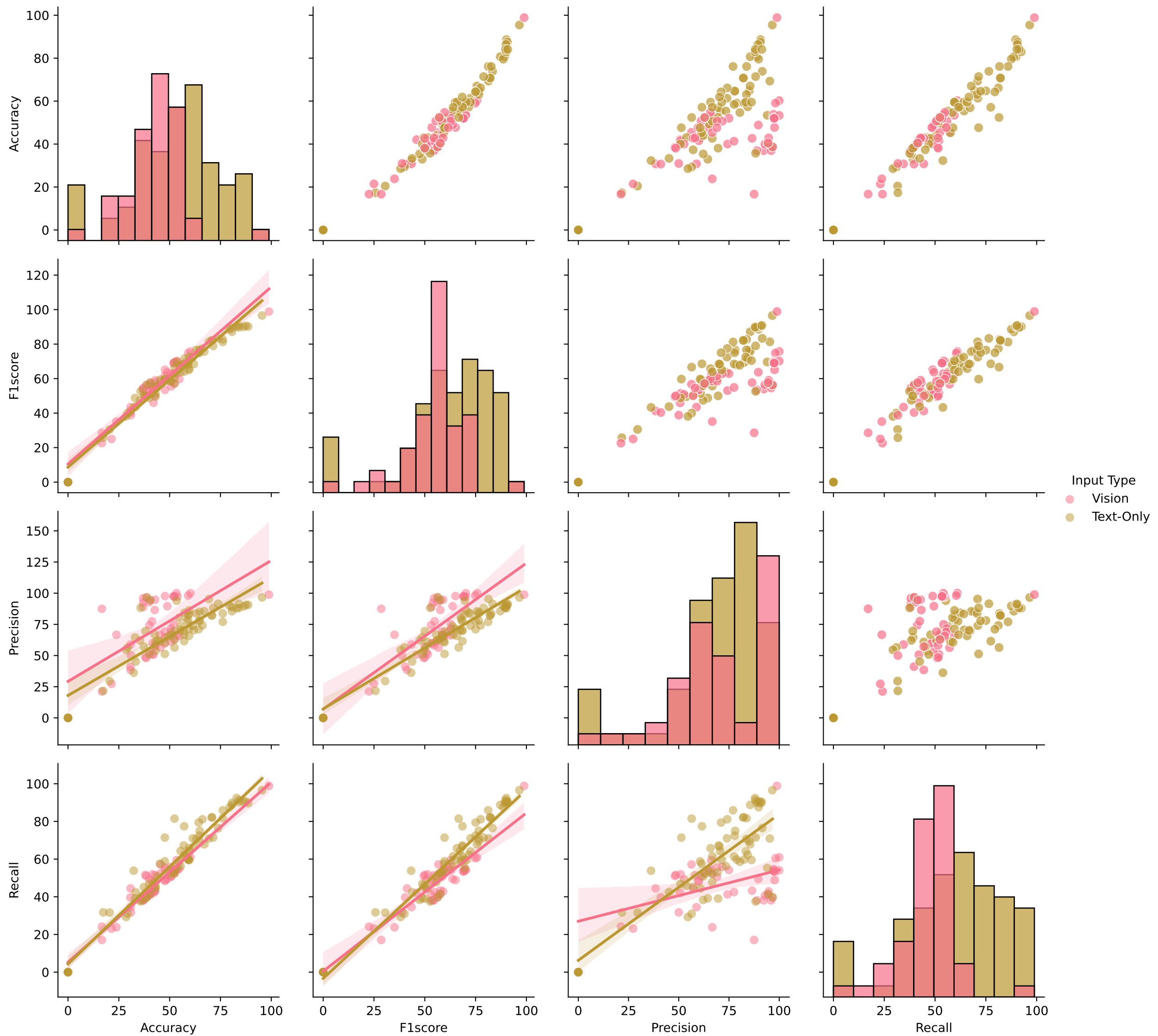
Average F1 Score: Family vs Source

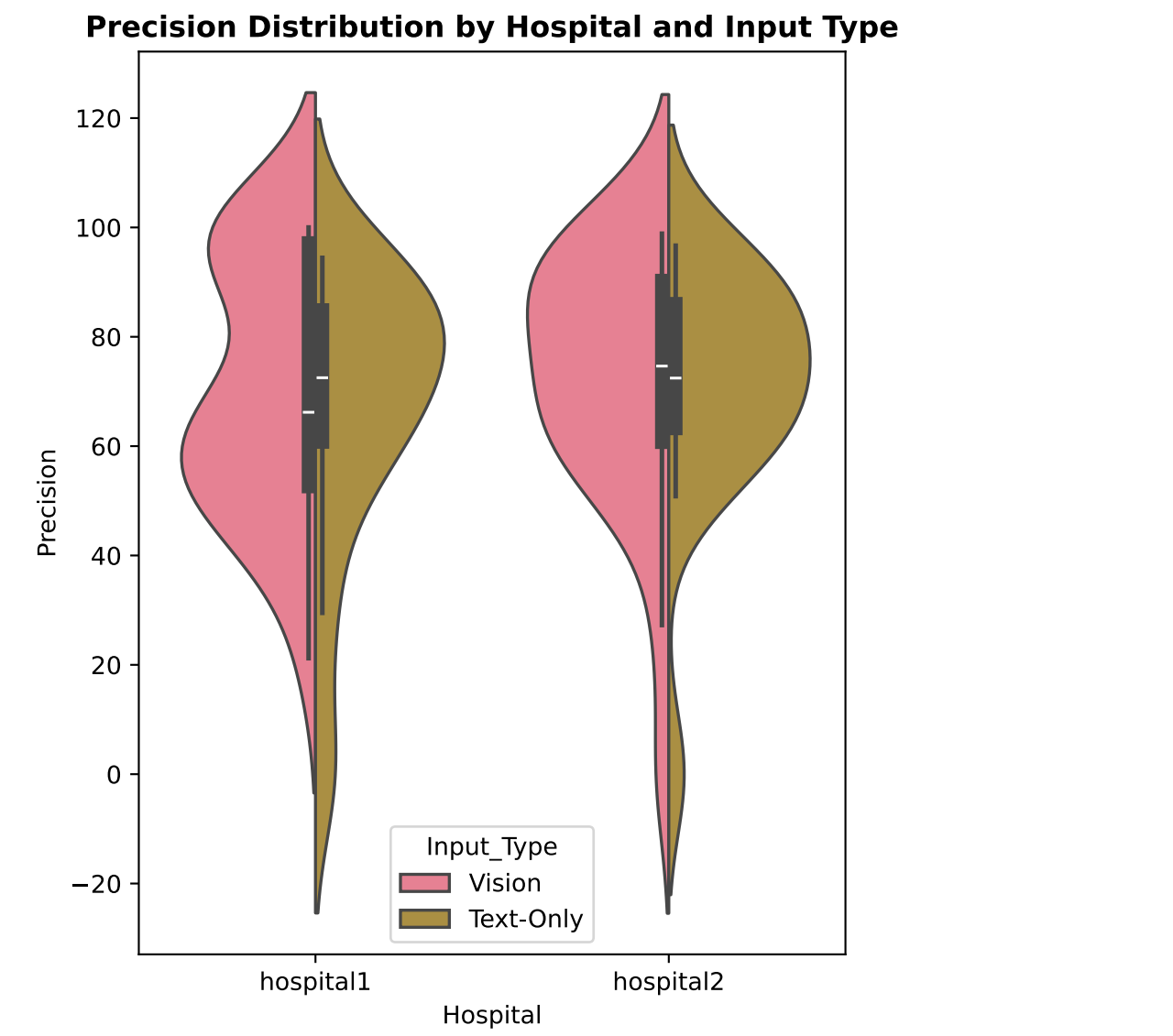
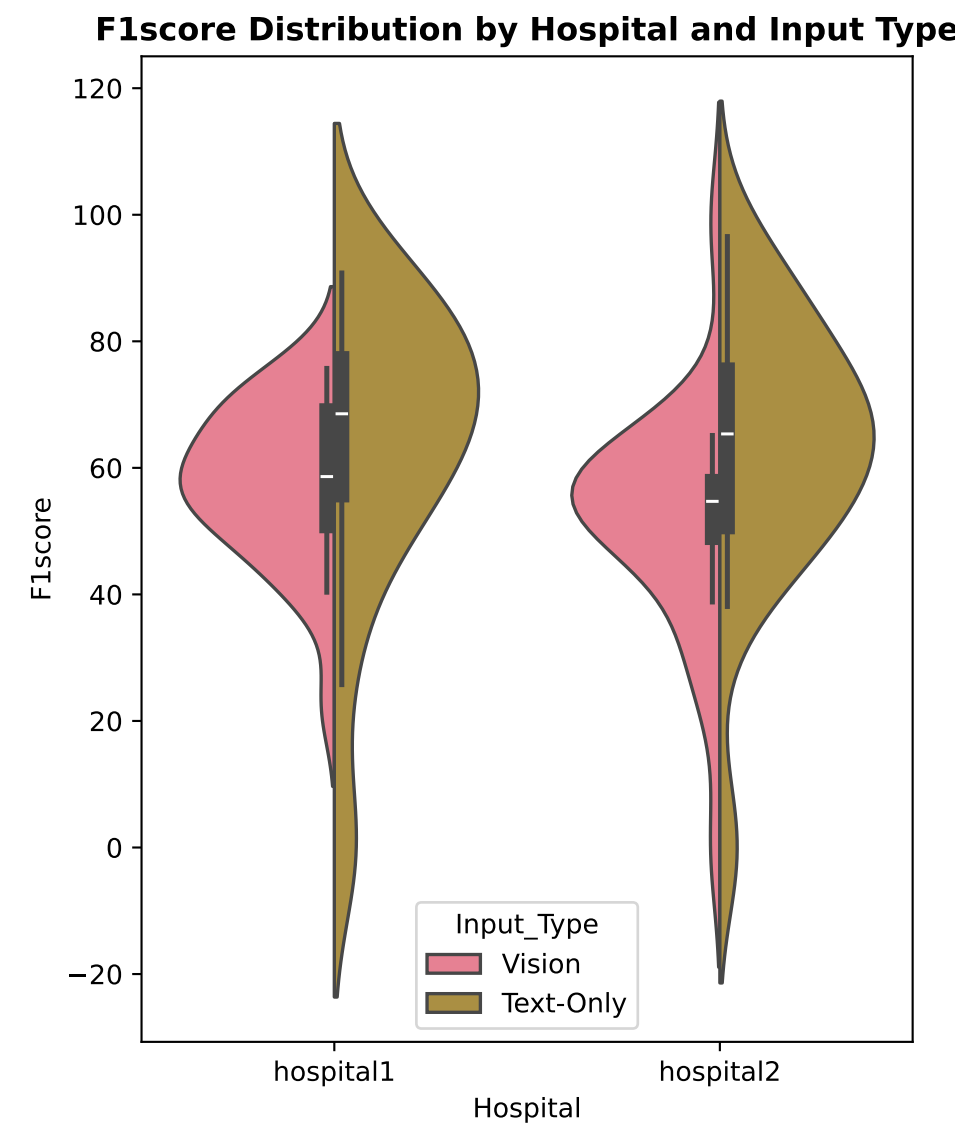
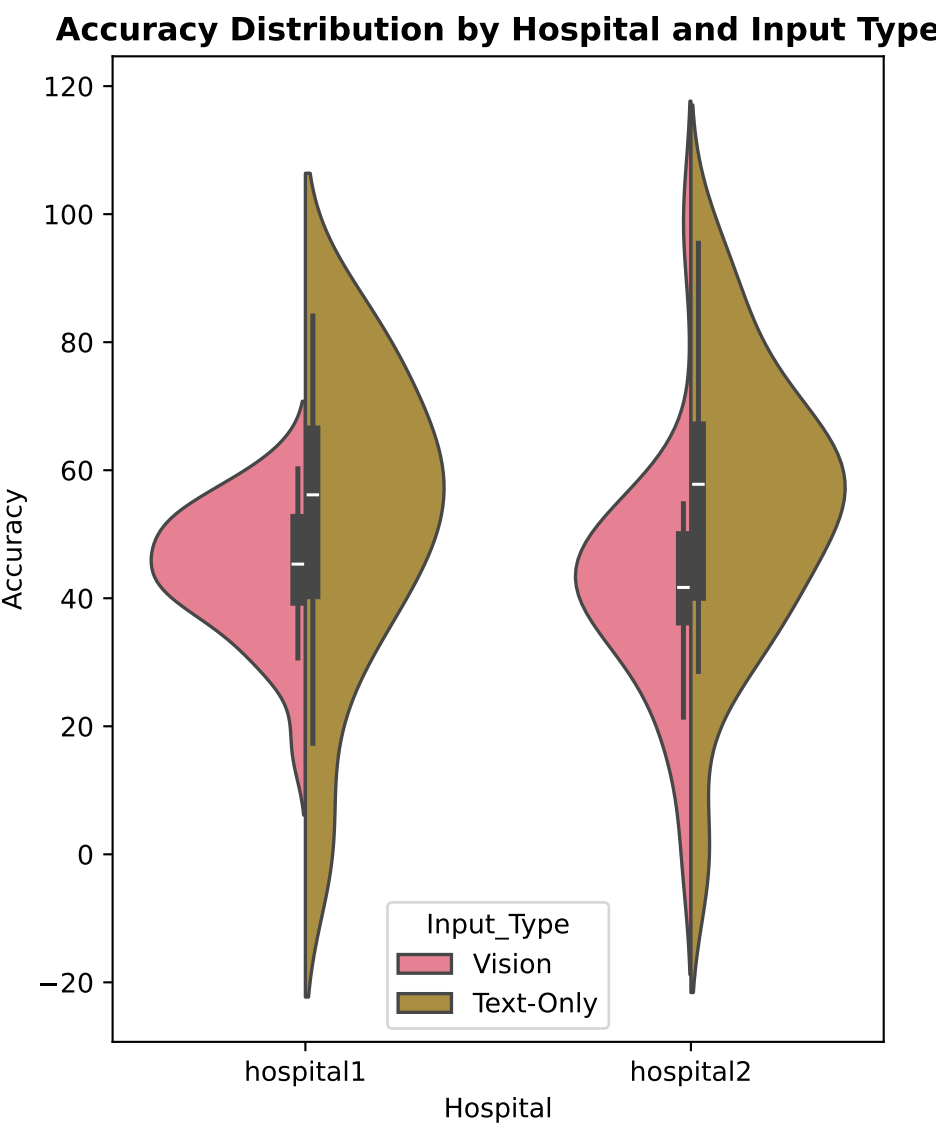
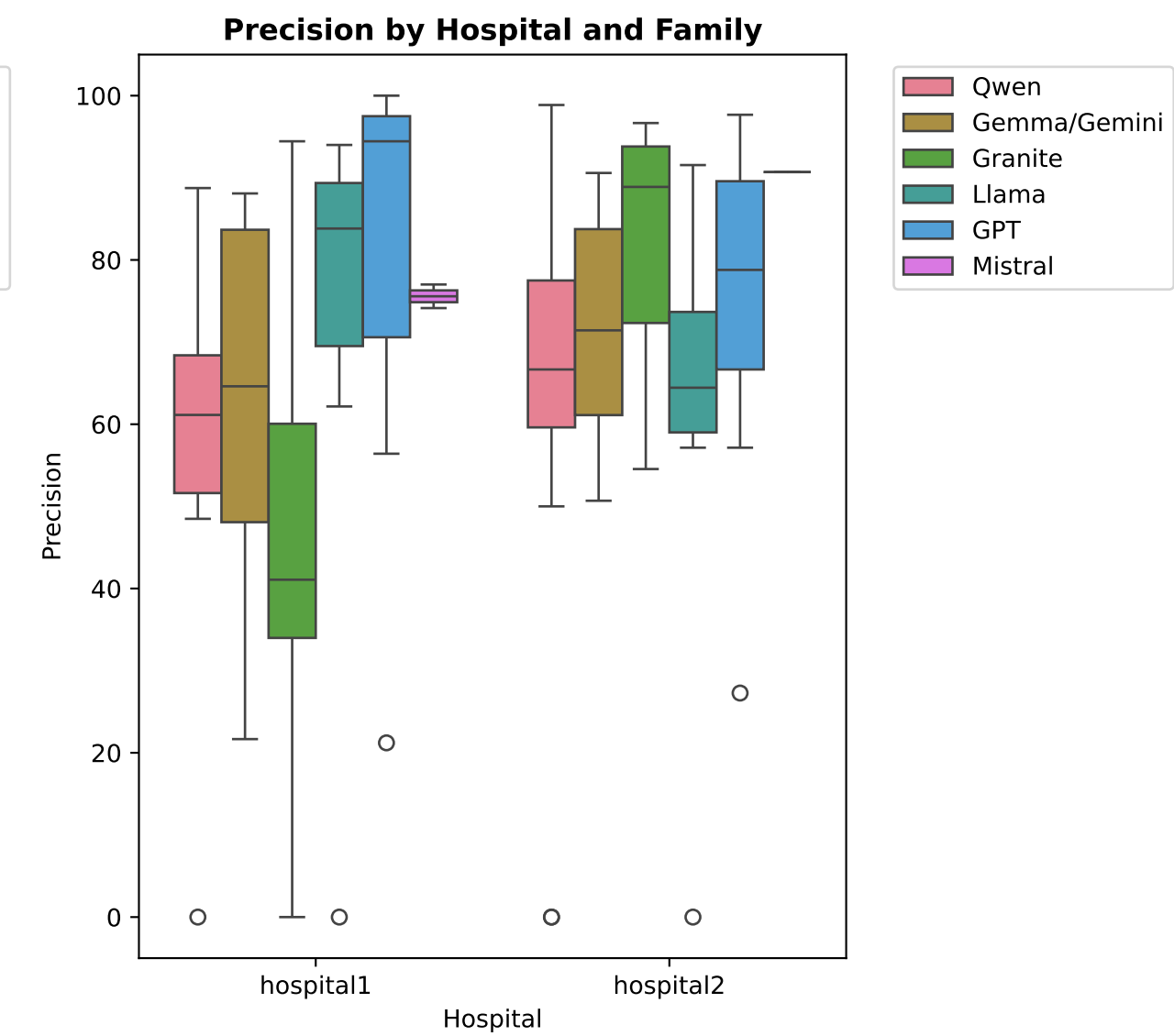
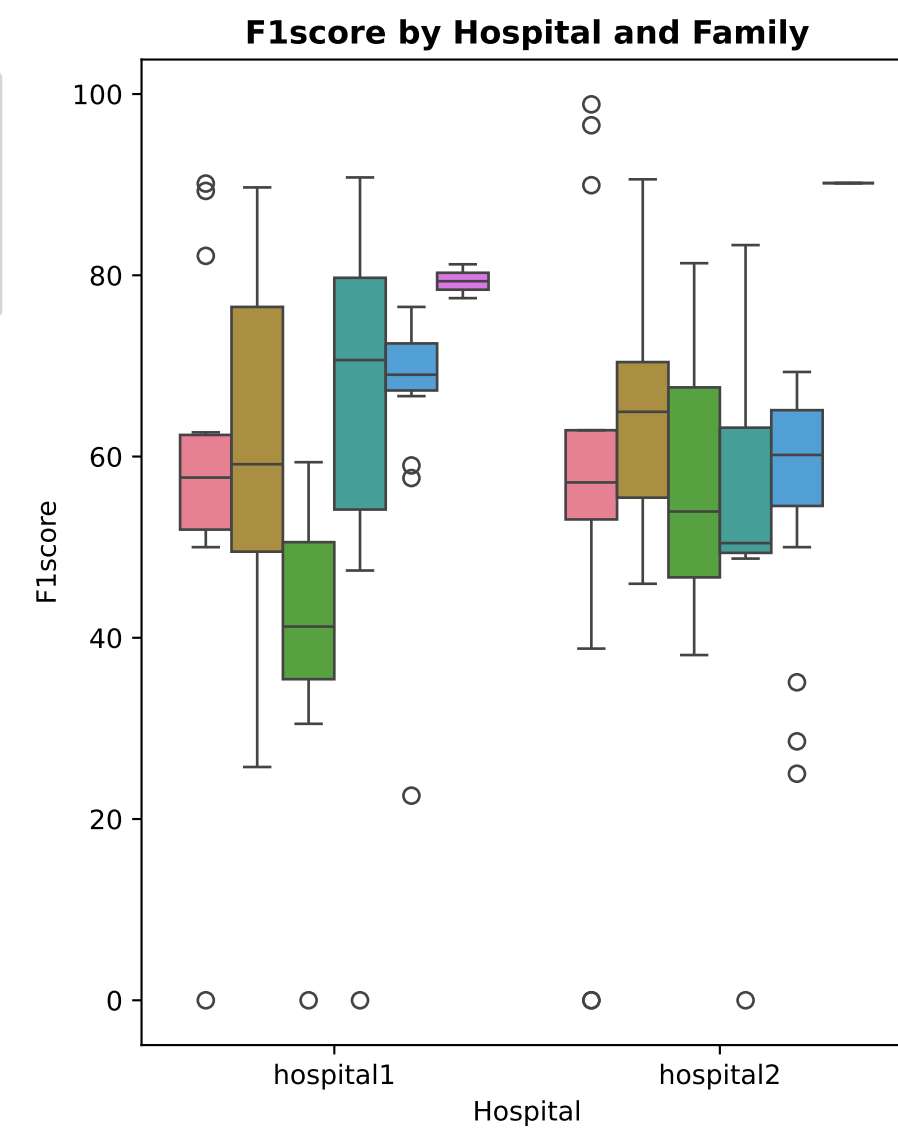
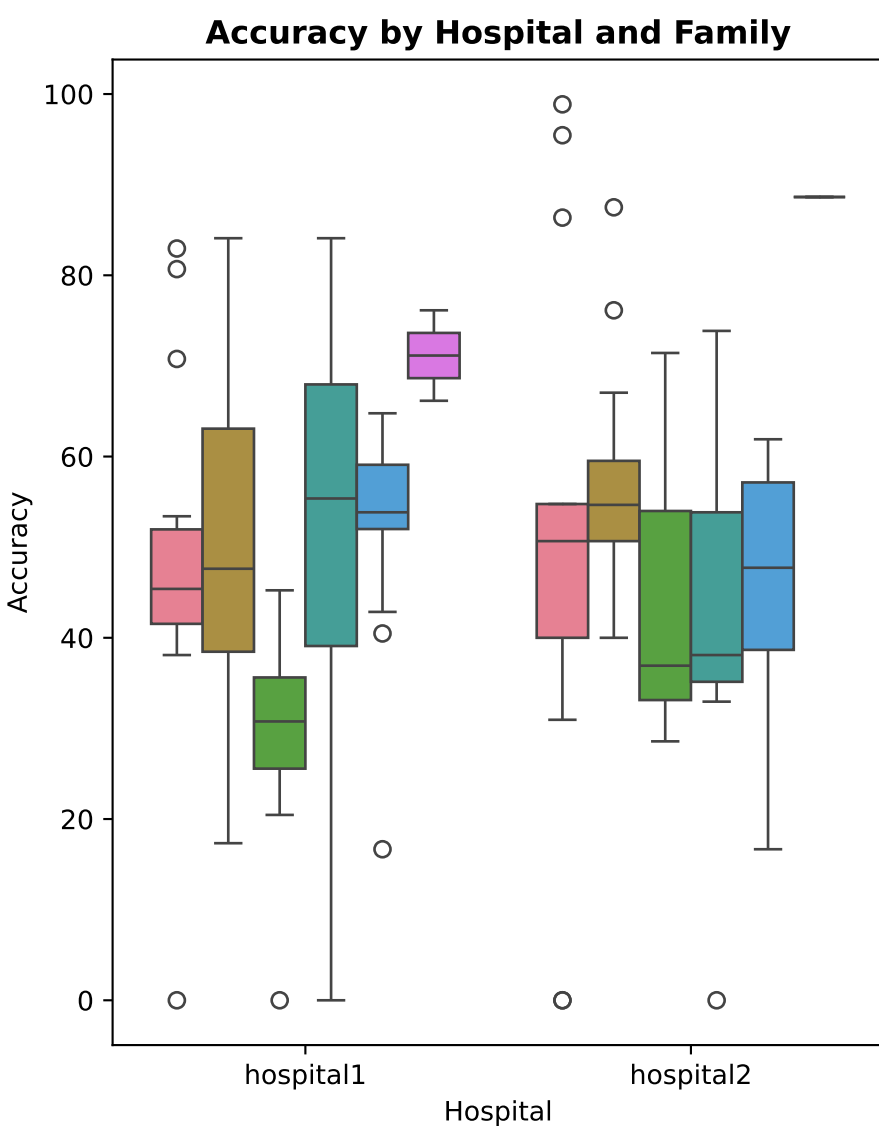


Performance Metrics Relationships by Model Family

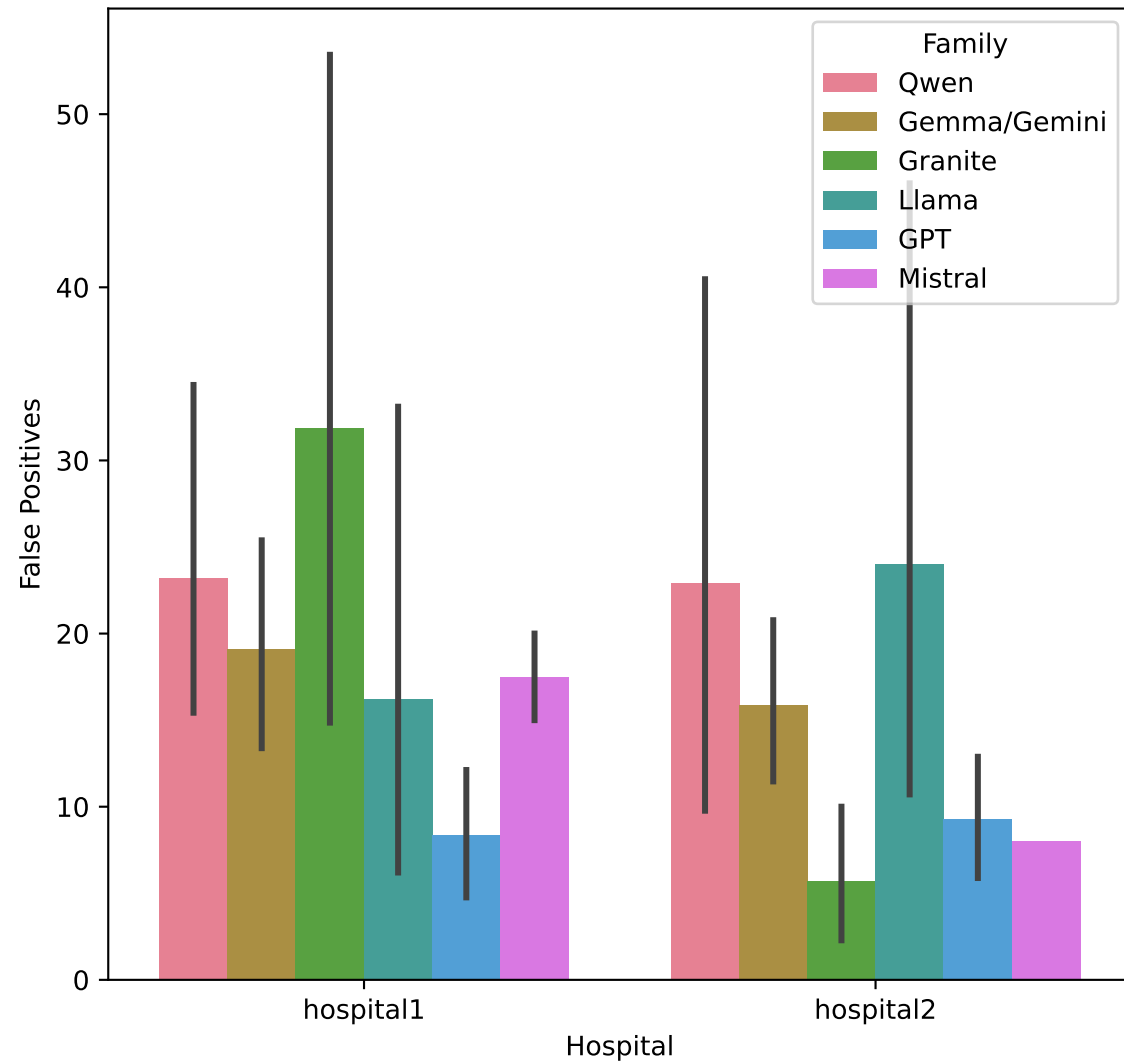


Performance Metrics Relationships by Input Type

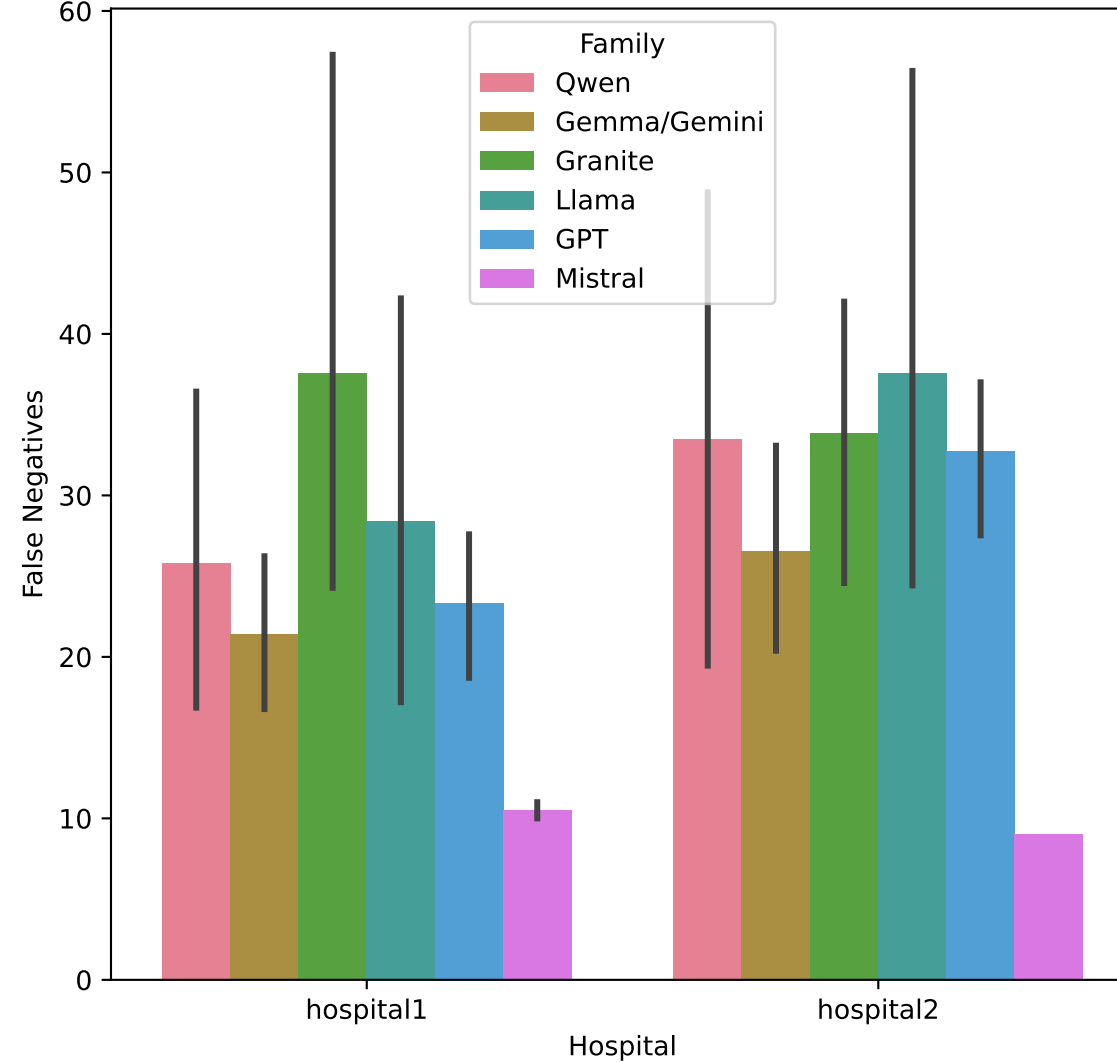




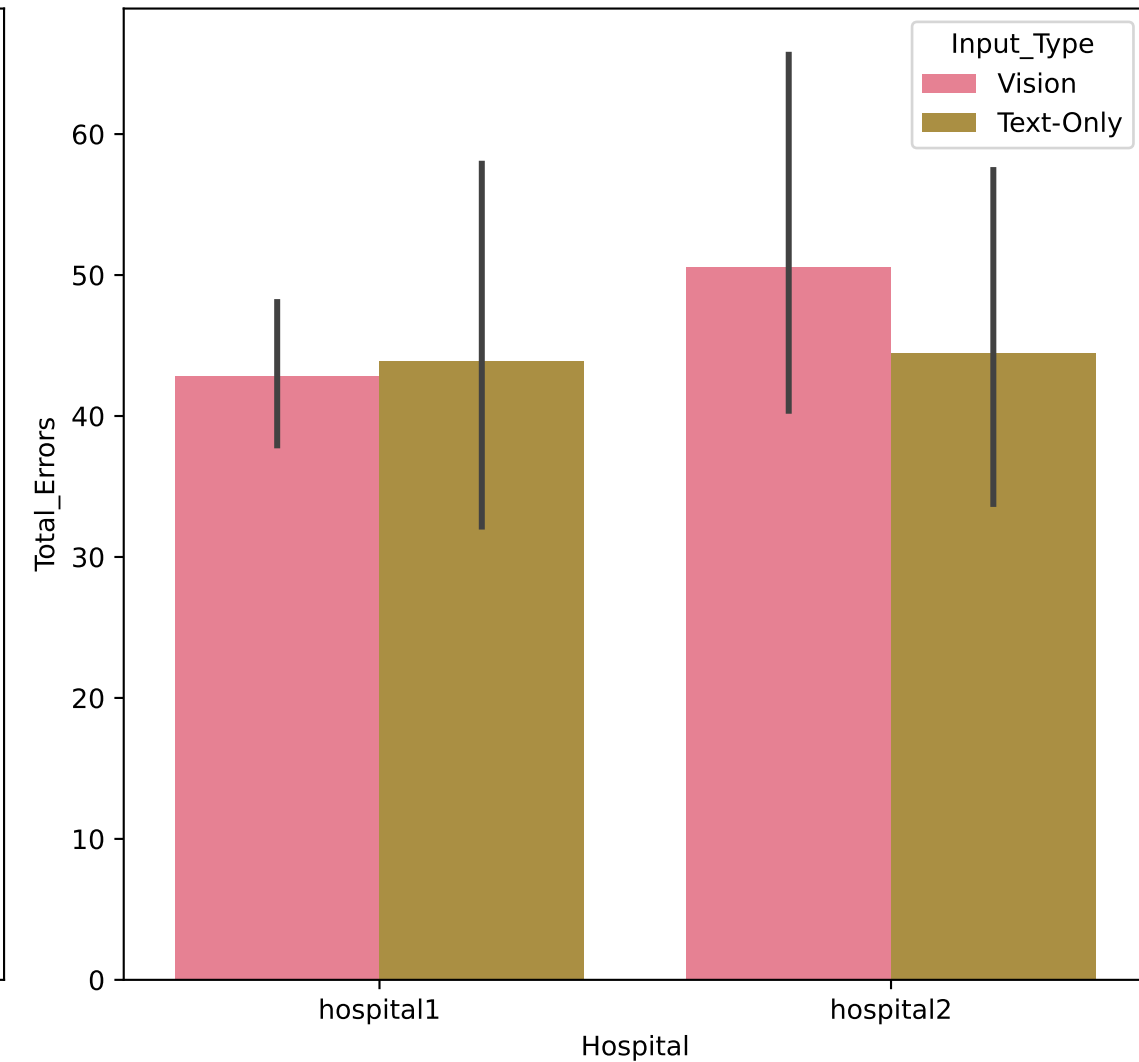
Average False Positives by Hospital and Family



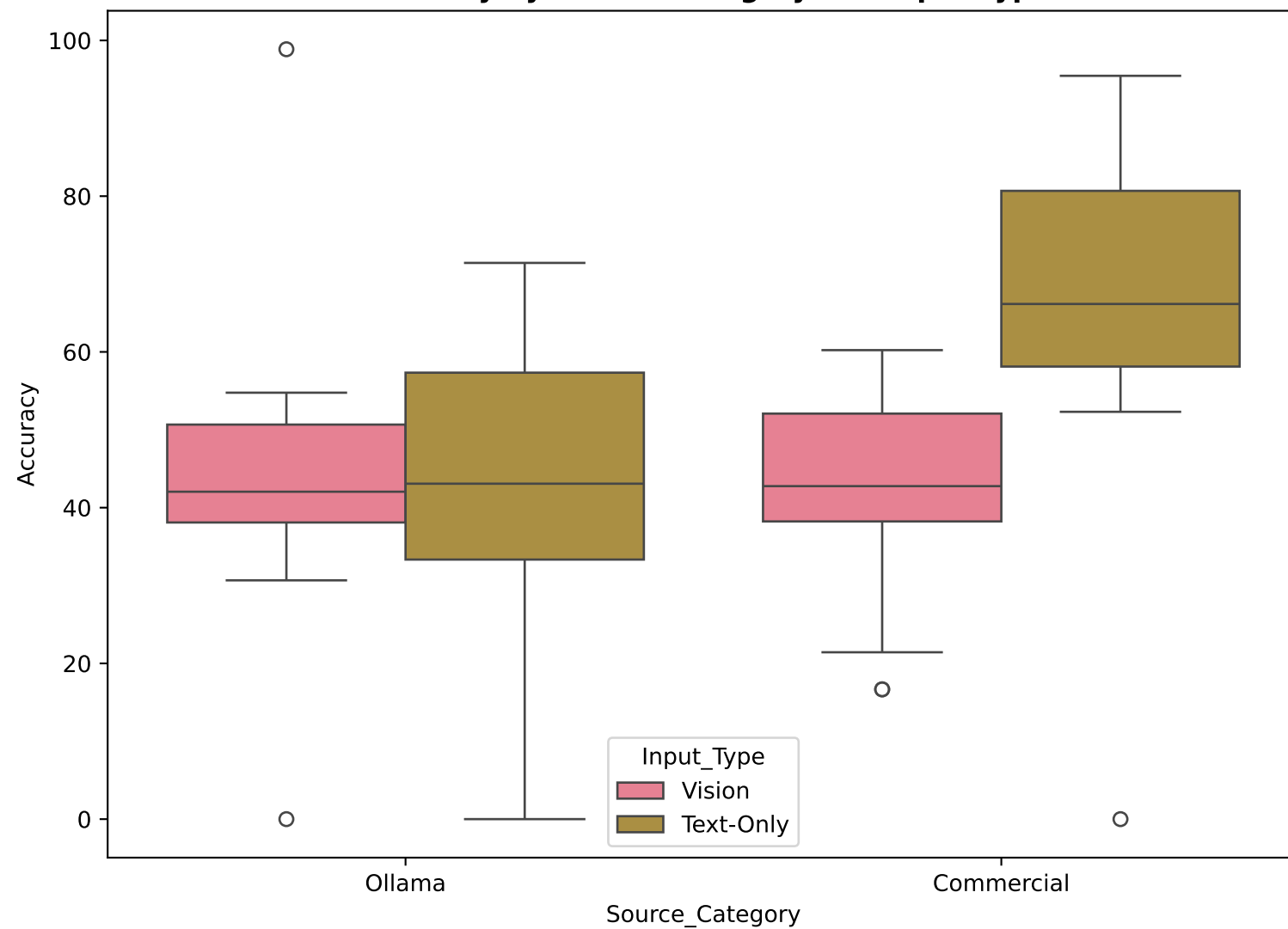
Average False Negatives by Hospital and Family



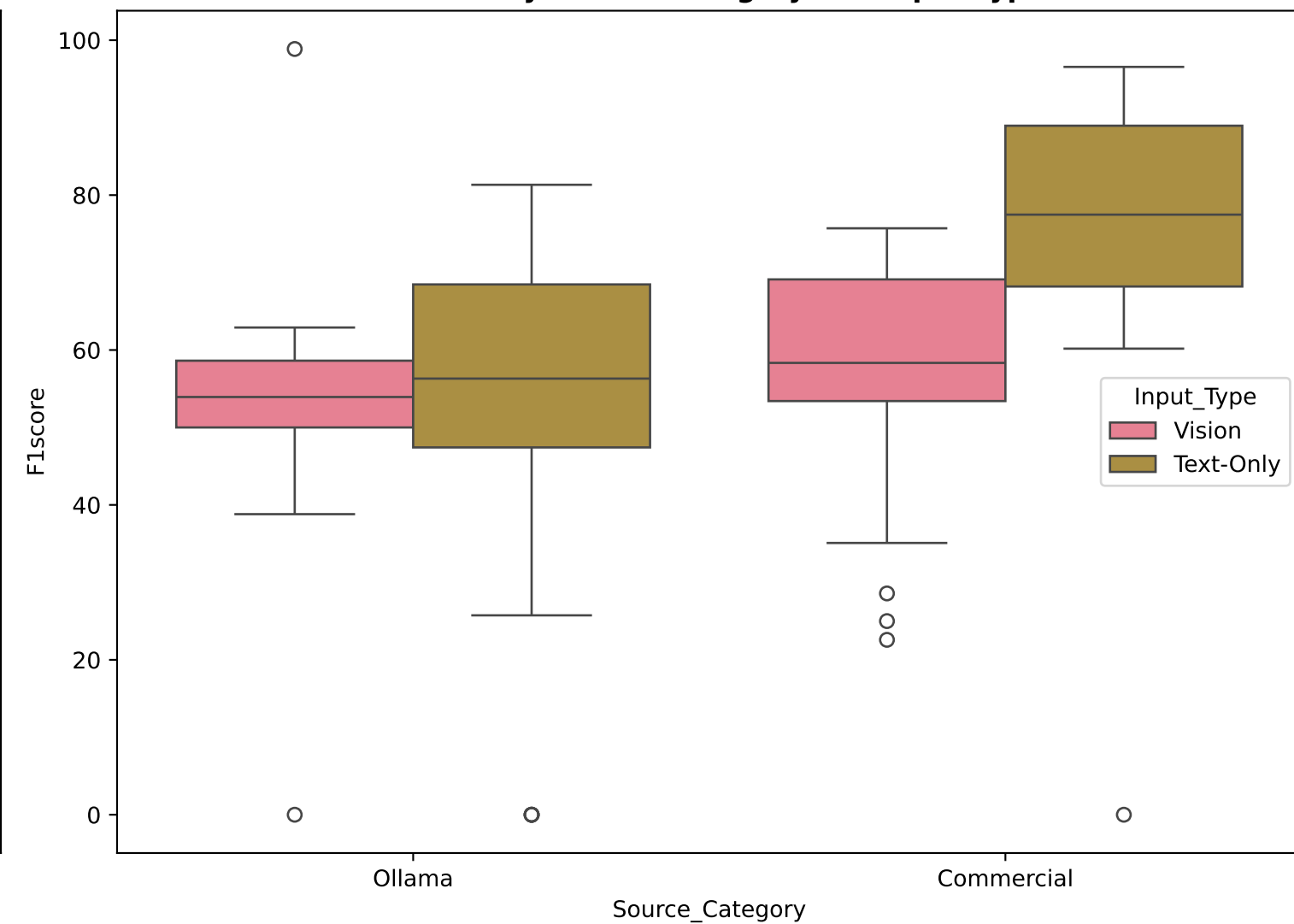
Total Errors by Hospital and Input Type



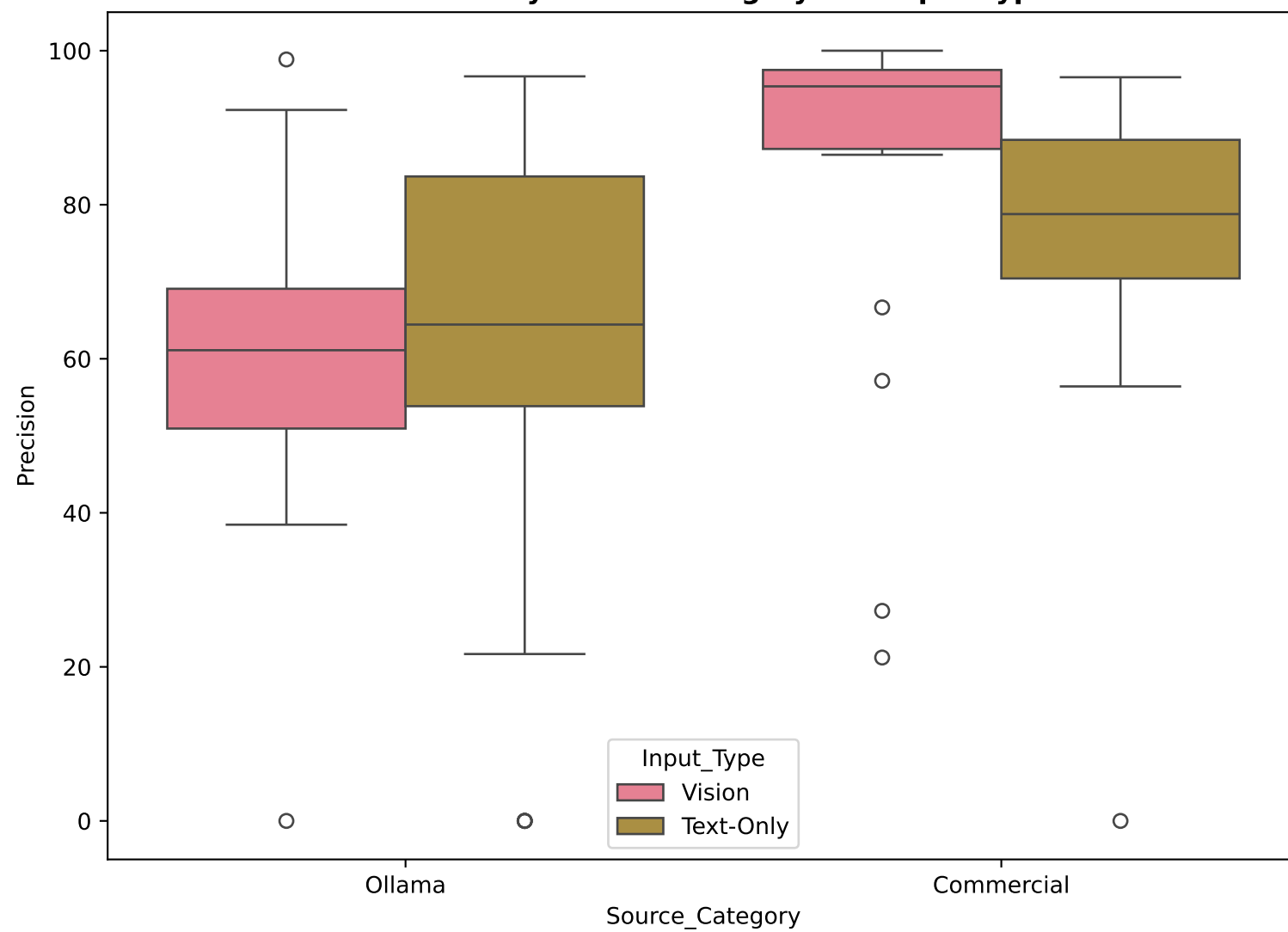
Accuracy by Source Category and Input Type



F1score by Source Category and Input Type



Precision by Source Category and Input Type



F1 Score vs Parameter Size by Source

