COMPREHENSIVE LLM BENCHMARKING ANALYSIS - SUMMARY REPORT
==

DATASET OVERVIEW:
- Total Records: 145
- Model Families: Qwen, Gemma/Gemini, Granite, Llama, GPT, NuExtract
- Hospitals: hospital1, hospital2
- Vision-enabled Models: 63
- Text-only Models: 82
- Ollama Models: 91
- Commercial Models: 54

TOP PERFORMERS:
- qwen2.5vl3b*ImageInput* (Qwen, 3.0B, Vision)
  F1: 98.86, Accuracy: 98.86
- NuExtract:0.5B (NuExtract, 0.5B, Text-only)
  F1: 97.62, Accuracy: 97.62
- NuExtract:2B (NuExtract, 2.0B, Text-only)
  F1: 97.62, Accuracy: 97.62
- NuExtract:2B (NuExtract, 2.0B, Text-only)
  F1: 97.62, Accuracy: 97.62
- NuExtract:0.5B (NuExtract, 0.5B, Text-only)
  F1: 97.62, Accuracy: 97.62

FAMILY PERFORMANCE RANKING (by F1 Score):
- NuExtract: 97.62 (n=8.0)
- GPT: 59.17 (n=46.0)
- Gemma/Gemini: 58.13 (n=29.0)
- Llama: 54.88 (n=15.0)
- Qwen: 50.90 (n=28.0)
- Granite: 47.82 (n=19.0)

KEY INSIGHTS:
- Vision Models Avg F1: 53.33
- Text-only Models Avg F1: 60.80
- Hospital 1 Avg F1: 57.70
- Hospital 2 Avg F1: 57.40
- Ollama Models Avg F1: 53.22
- Commercial Models Avg F1: 64.87

Grouped Model F1 Score Statistics:
- Unique Base Models: 13
- Total Test Instances: 145
- Best Performing Model: NuExtract:2B (F1: 97.62)
- Worst Performing Model: granite3.32b (F1: 37.28)
- Overall Average F1: 60.04
- Models with Vision: 6

Top 5 Performers:
- gpt-4.1-nano (GPT, with Vision): F1 = 62.09 ± 11.01
- llama3.23b (Llama, Text-only): F1 = 62.60 ± 11.41
- gemma31b (Gemma/Gemini, Text-only): F1 = 64.94 ± 11.08
- NuExtract:0.5B (NuExtract, Text-only): F1 = 97.62 ± 0.00
- NuExtract:2B (NuExtract, 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

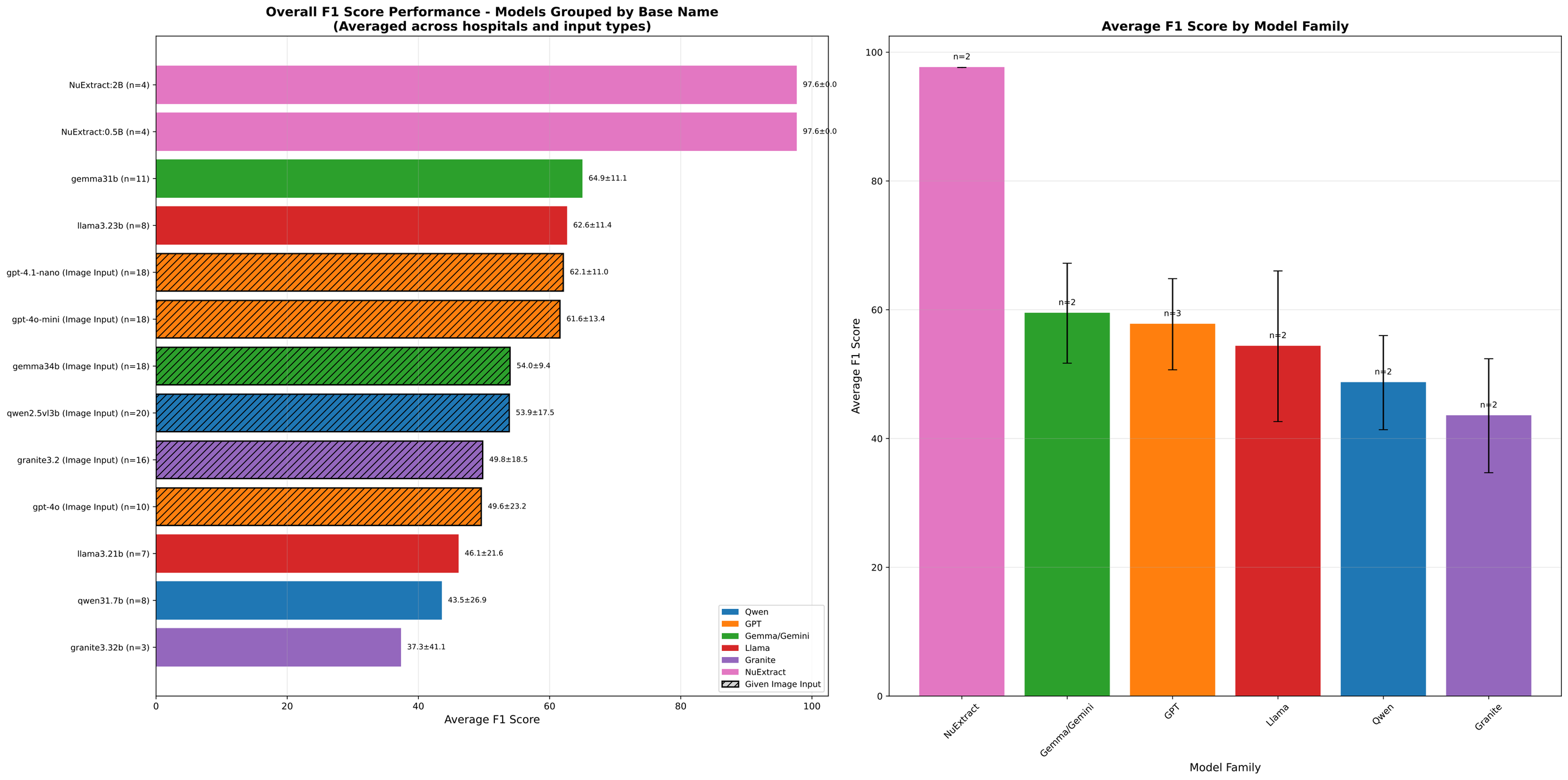
Error Analysis Summary:
- Average False Positives: 14.80
- Average False Negatives: 26.49
- Models with more FP than FN: 20
- Models with more FN than FP: 109

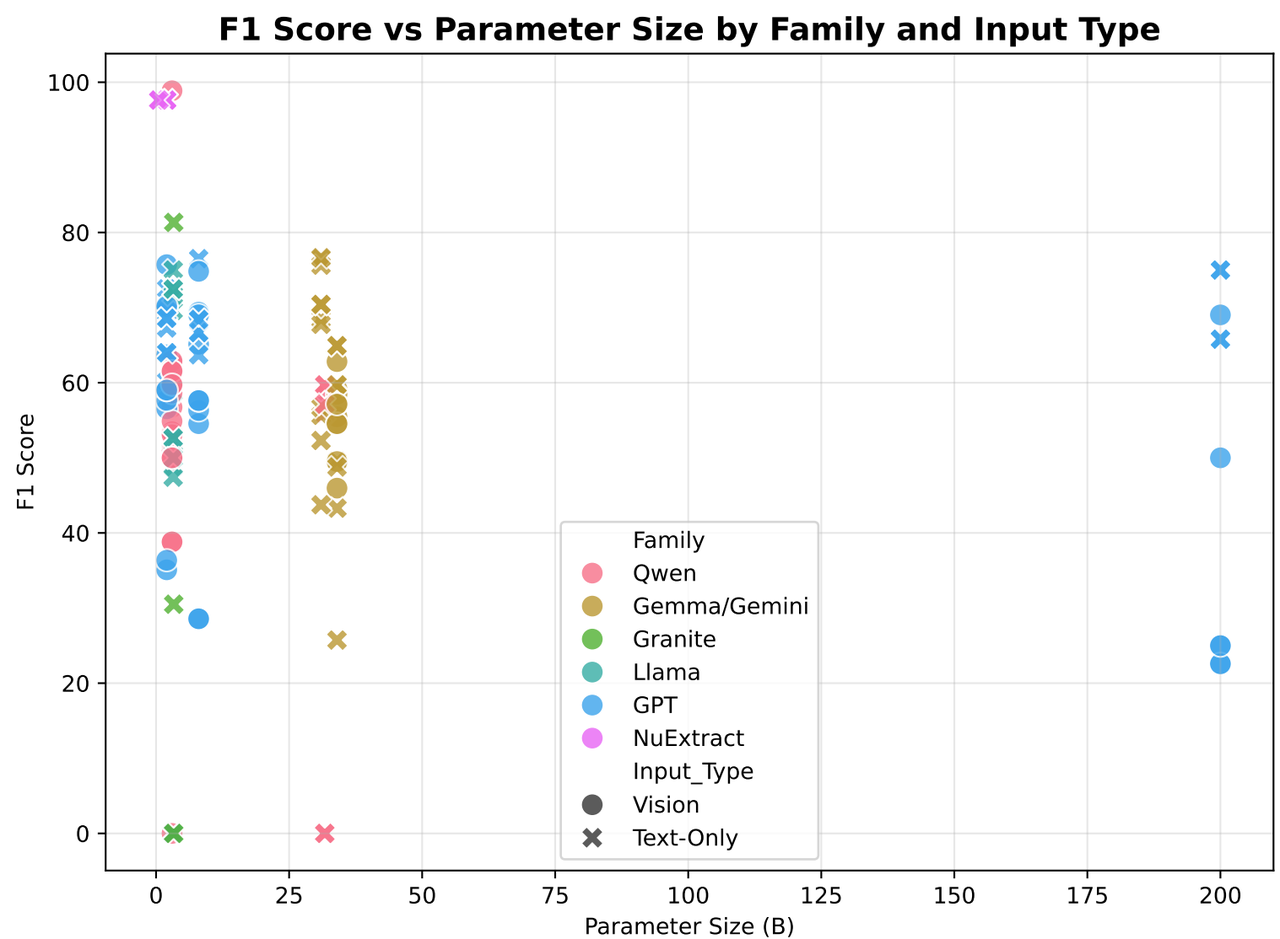
SOURCE CATEGORY DETAILED STATISTICS:

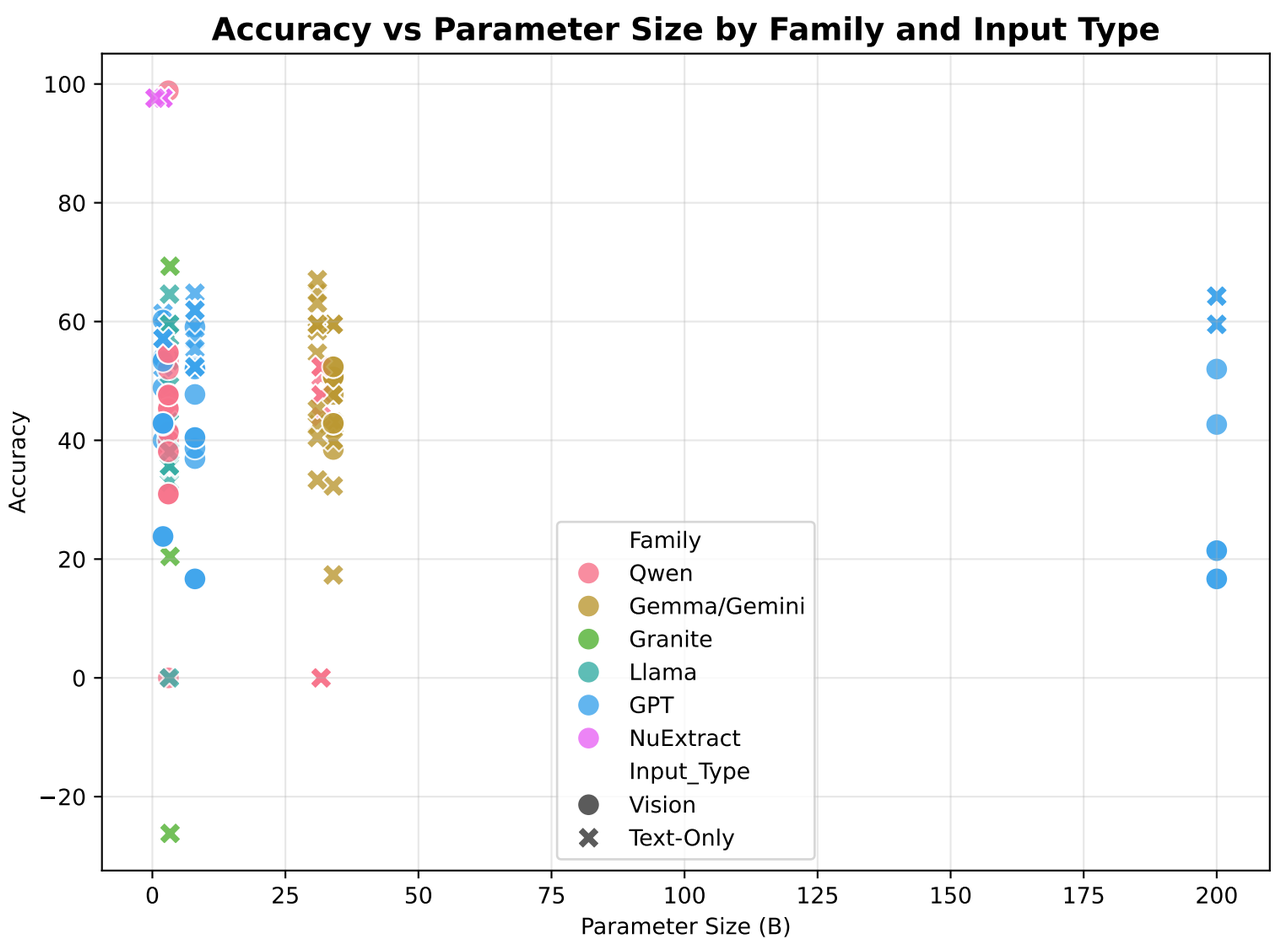
COMMERCIAL MODELS:
  GPT:
    Accuracy: 47.2±14.8 (n=46)
    F1 Score: 59.2±15.7
    Precision: 75.6±21.1
    False Positives: 9.5
    False Negatives: 25.6
  NuExtract:
    Accuracy: 97.6±0.0 (n=8)
    F1 Score: 97.6±0.0
    Precision: 97.6±0.0
    False Positives: 1.0
    False Negatives: 1.0

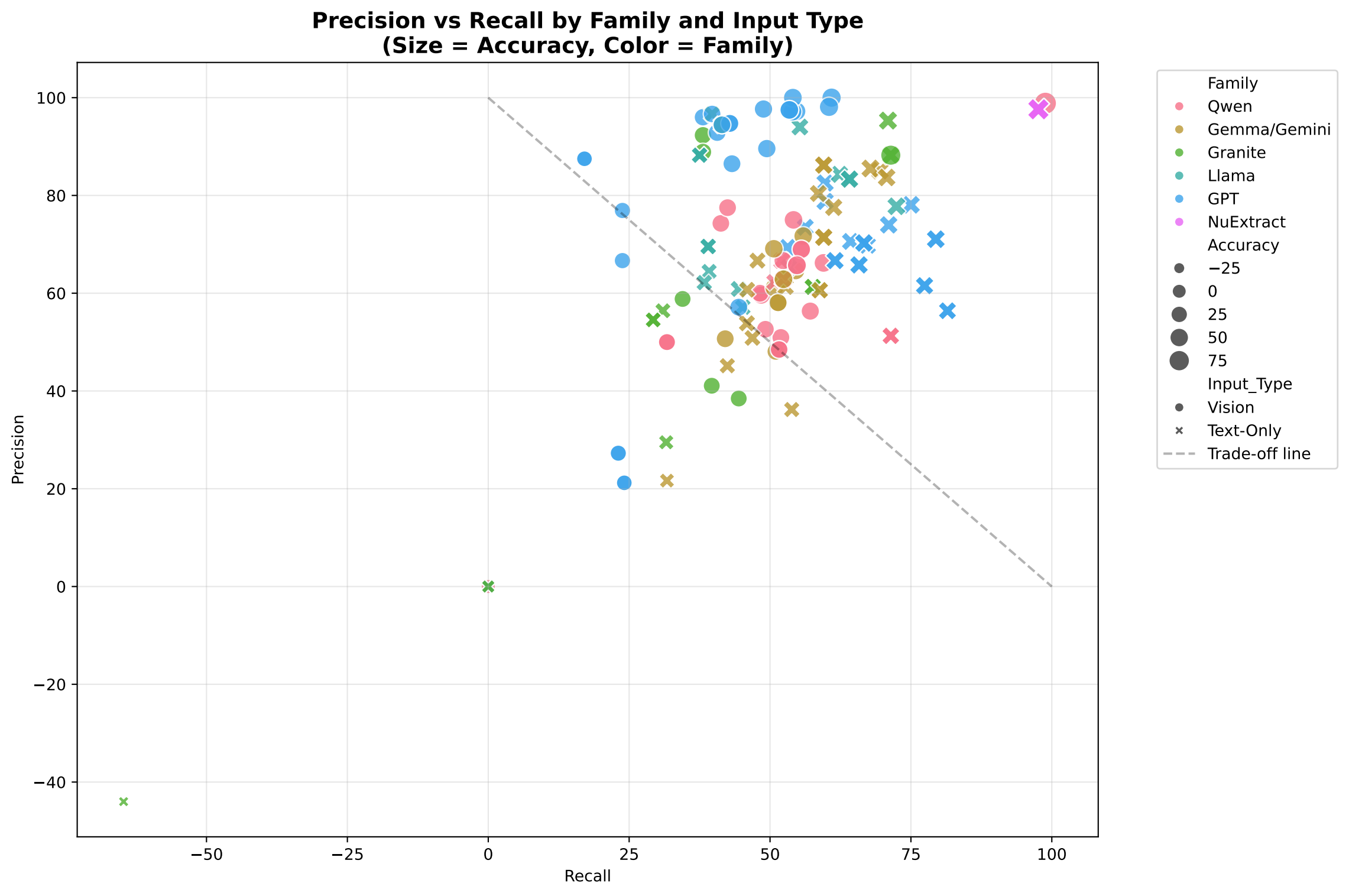
OLLAMA MODELS:
  Gemma/Gemini:
    Accuracy: 48.5±11.0 (n=29)
    F1 Score: 58.1±11.3
    Precision: 63.9±15.4
    False Positives: 16.8
    False Negatives: 24.3
  Granite:
    Accuracy: 35.2±22.9 (n=19)
    F1 Score: 47.8±22.2
    Precision: 60.6±37.0
    False Positives: 17.2
    False Negatives: 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

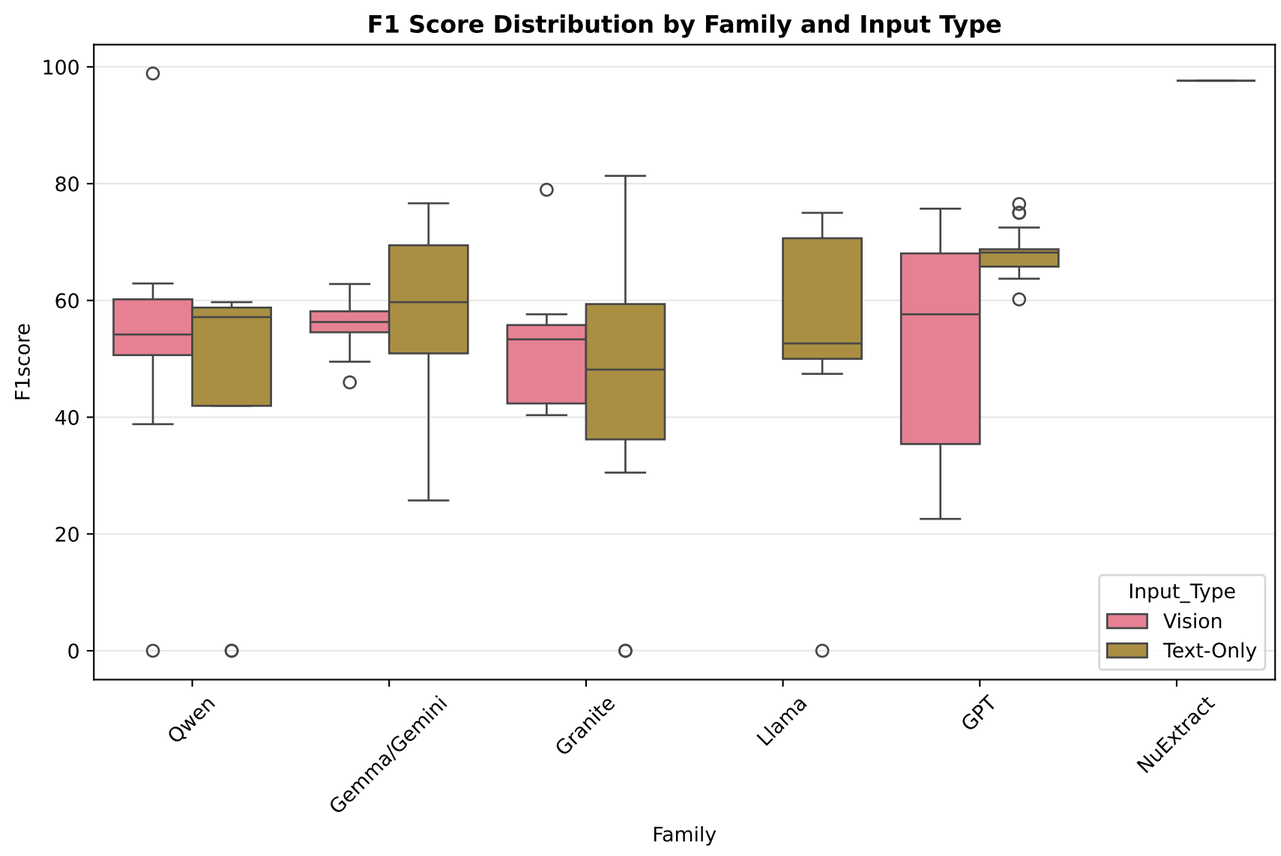
==

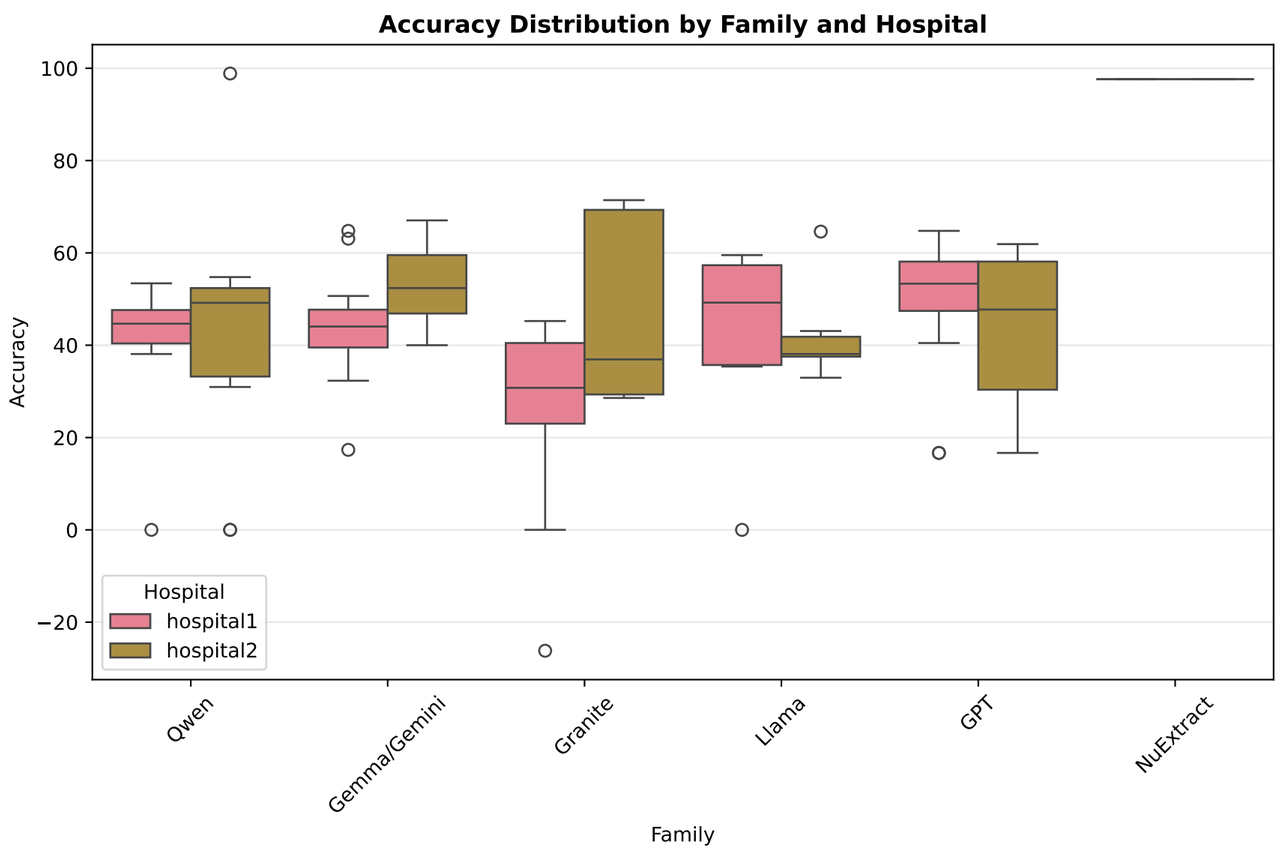

Overall F1 Score Performance - Models Grouped by Base Name
(Averaged across hospitals and input types)
NuExtract:2B (n=4)
NuExtract:0.5B (n=4)
gemma31b (n=11)
llama3.23b (n=8)
gpt-4.1-nano (Image Input) (n=18)
gpt-4o-mini (Image Input) (n=18)
gemma34b (Image Input) (n=18)
qwen2.5vl3b (Image Input) (n=20)
granite3.2 (Image Input) (n=16)
gpt-4o (Image Input) (n=10)
llama3.21b (n=7)
qwen31.7b (n=8)
granite3.32b (n=3)
97.6±0.0
97.6±0.0
64.9±11.1
62.6±11.4
62.1±11.0
61.6±13.4
54.0±9.4
53.9±17.5
49.8±18.5
49.6±23.2
46.1±21.6
43.5±26.9
37.3±41.1
Qwen
GPT
Gemma/Gemini
Llama
Granite
NuExtract
Given Image Input
0
20
40
60
80
100
Average F1 Score
Average F1 Score by Model Family
n=2
n=2
n=3
n=2
n=2
n=2
Average F1 Score
NuExtract
Gemma/Gemini
GPT
Llama
Qwen
Granite
Model Family

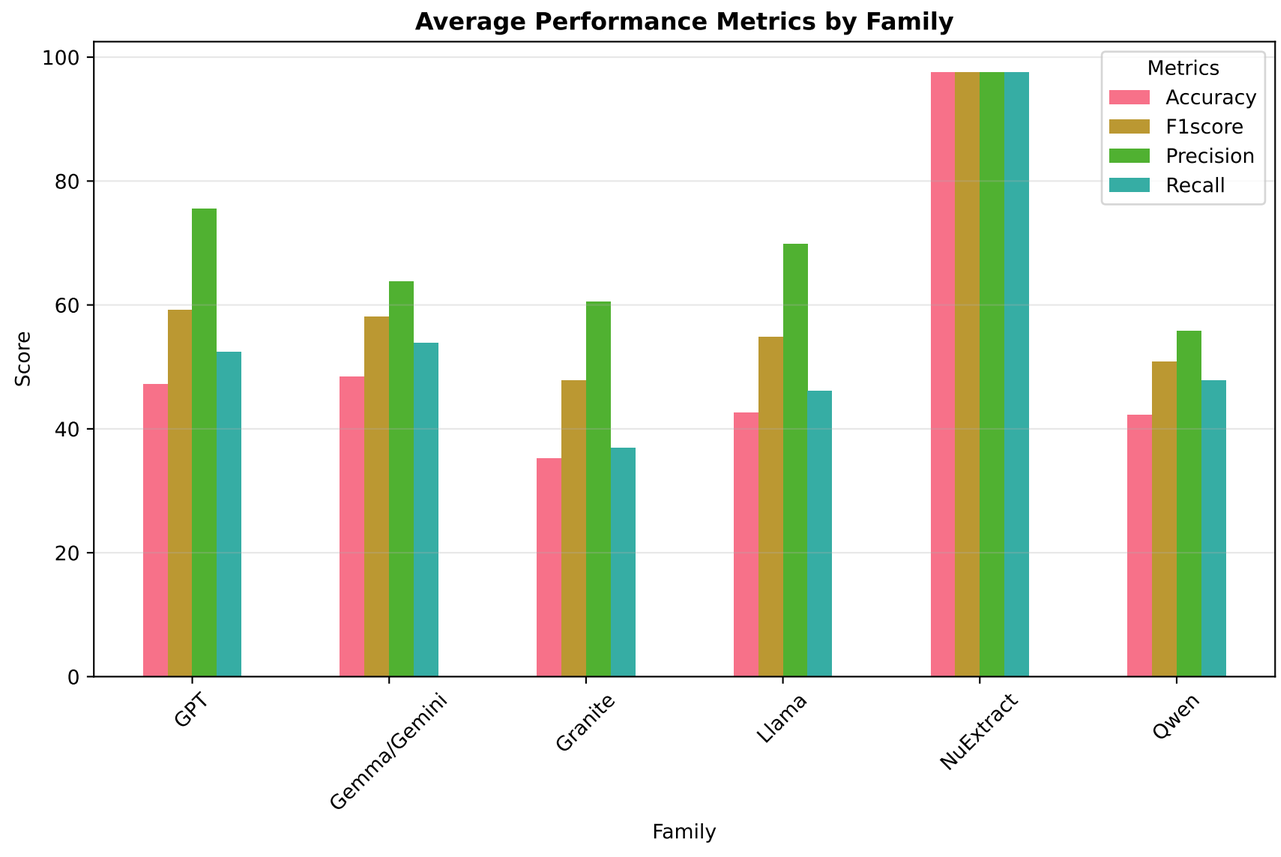

F1 Score vs Parameter Size by Family and Input Type
F1 Score
Parameter Size (B)
Family
Qwen
Gemma/Gemini
Granite
Llama
GPT
NuExtract
Input_Type
Vision
Text-Only

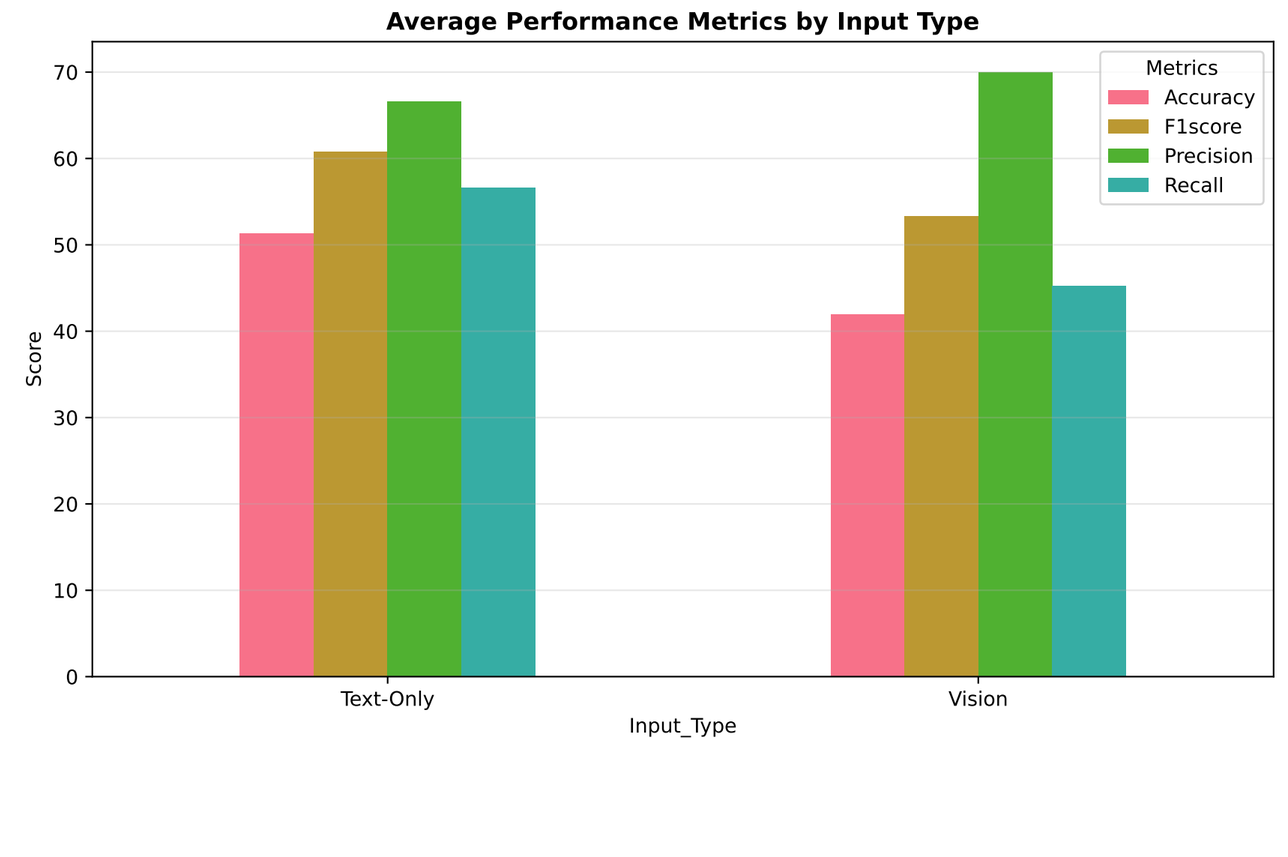

Accuracy vs Parameter Size by Family and Input Type
Accuracy
Parameter Size (B)
Family
Qwen
Gemma/Gemini
Granite
Llama
GPT
NuExtract
Input_Type
Vision
Text-Only

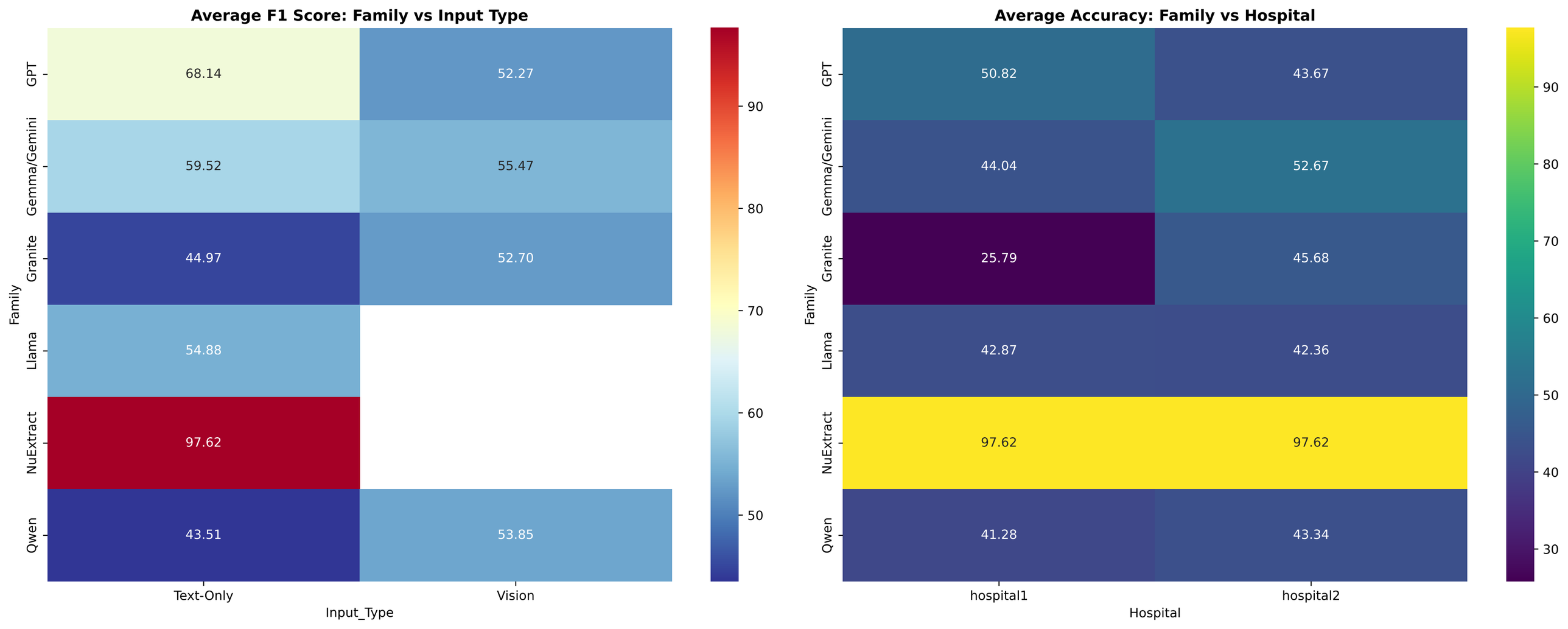

Precision vs Recall by Family and Input Type
(Size = Accuracy, Color = Family)
Precision
Recall
100
80
60
40
20
0
−20
−40
−50
−25
0
25
50
75
100
Family
Qwen
Gemma/Gemini
Granite
Llama
GPT
NuExtract
Accuracy
−25
0
25
50
75
Input_Type
Vision
Text-Only
Trade-off line

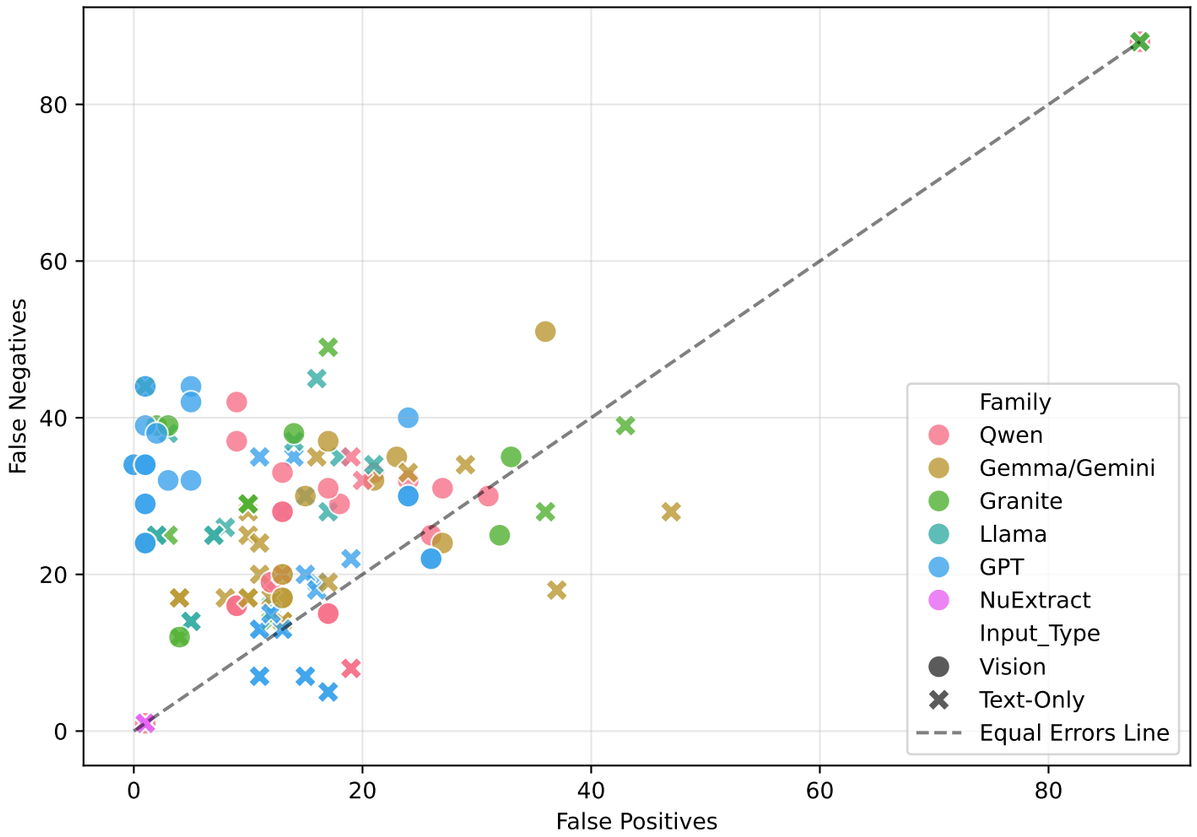

F1 Score Distribution by Family and Input Type
F1score
Family
Input_Type
Vision
Text-Only
Qwen
Gemma/Gemini
Granite
Llama
GPT
NuExtract

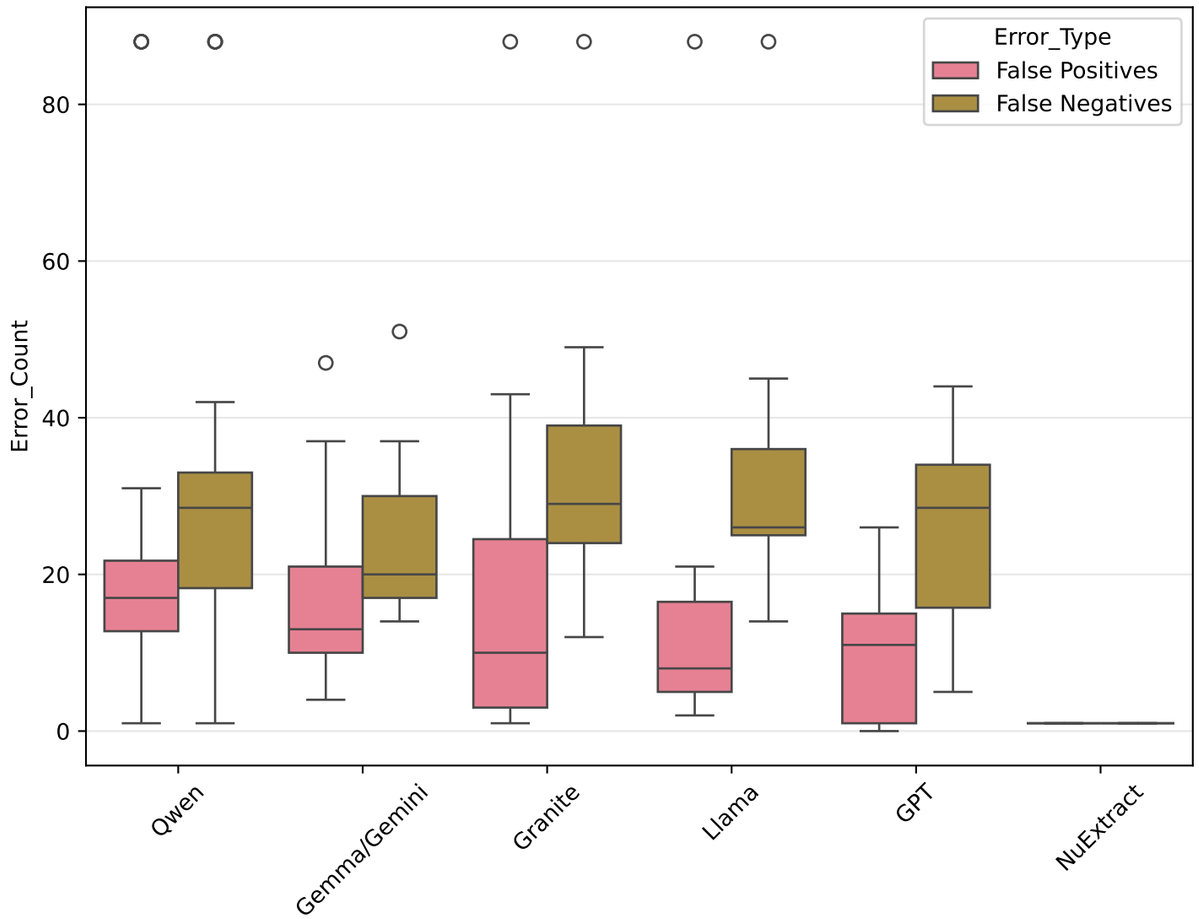

Accuracy Distribution by Family and Hospital
Accuracy
Family
Hospital
hospital1
hospital2
Qwen
Gemma/Gemini
Granite
Llama
GPT
NuExtract

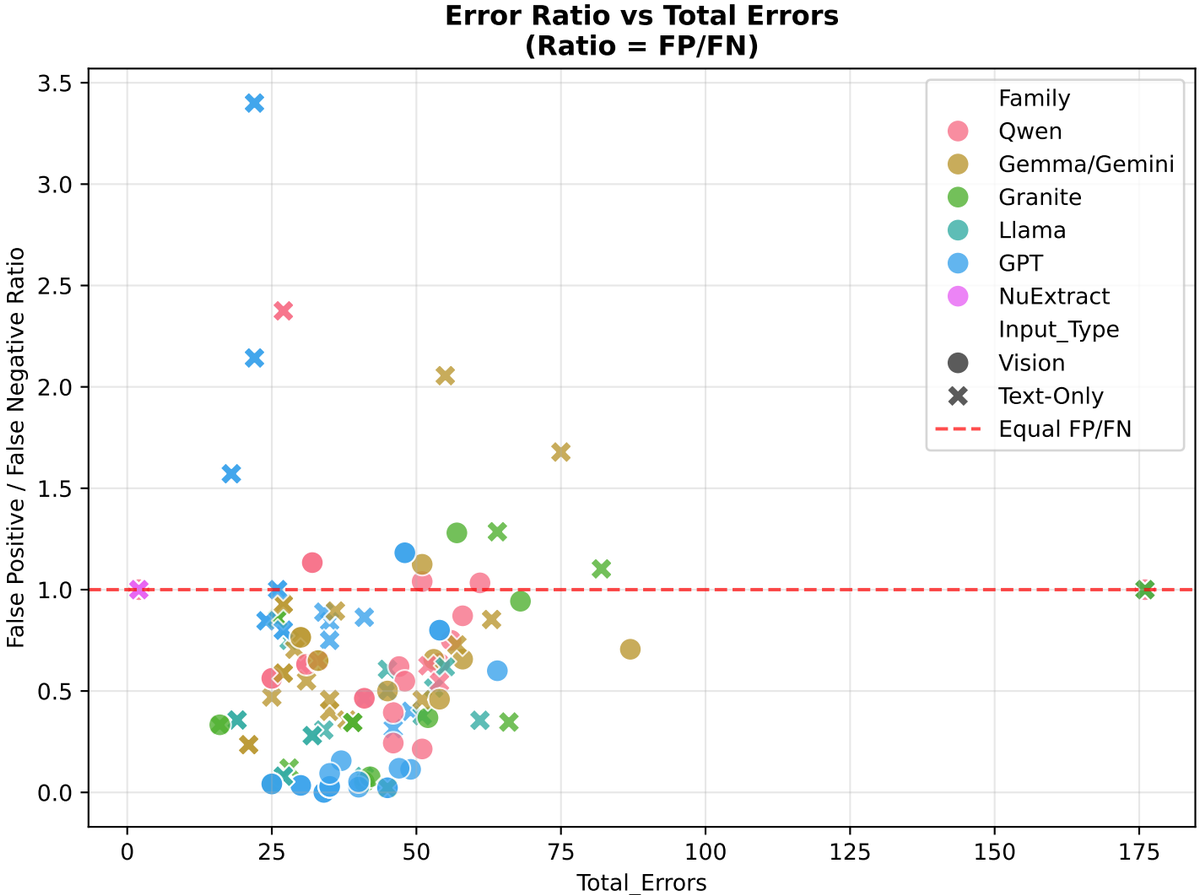

Average Performance Metrics by Family
Score
Family
Metrics
Accuracy
F1score
Precision
Recall
GPT
Gemma/Gemini
Granite
Llama
NuExtract
Qwen

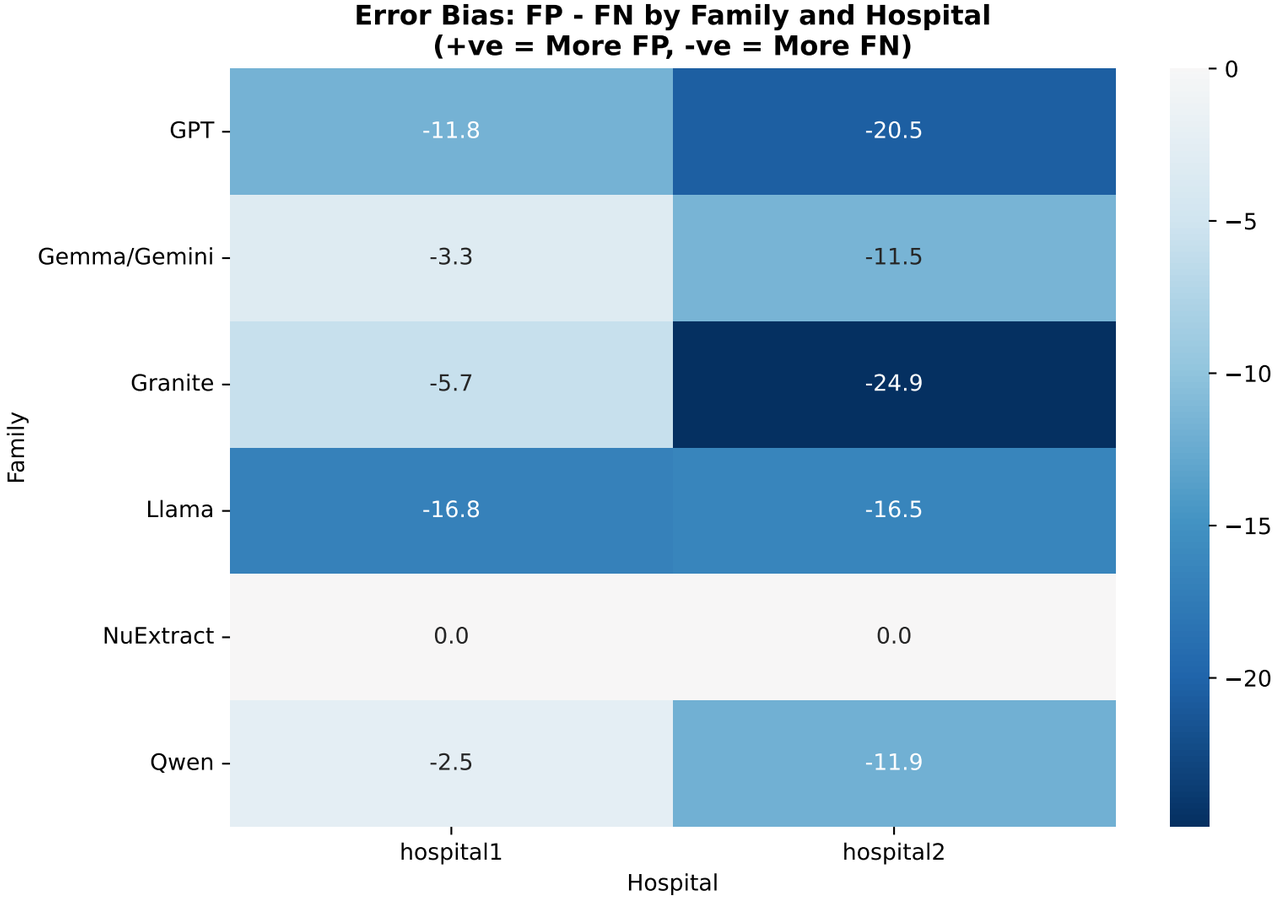

Average Performance Metrics by Input Type
Score
Input_Type
Metrics
Accuracy
F1score
Precision
Recall
Text-Only
Vision

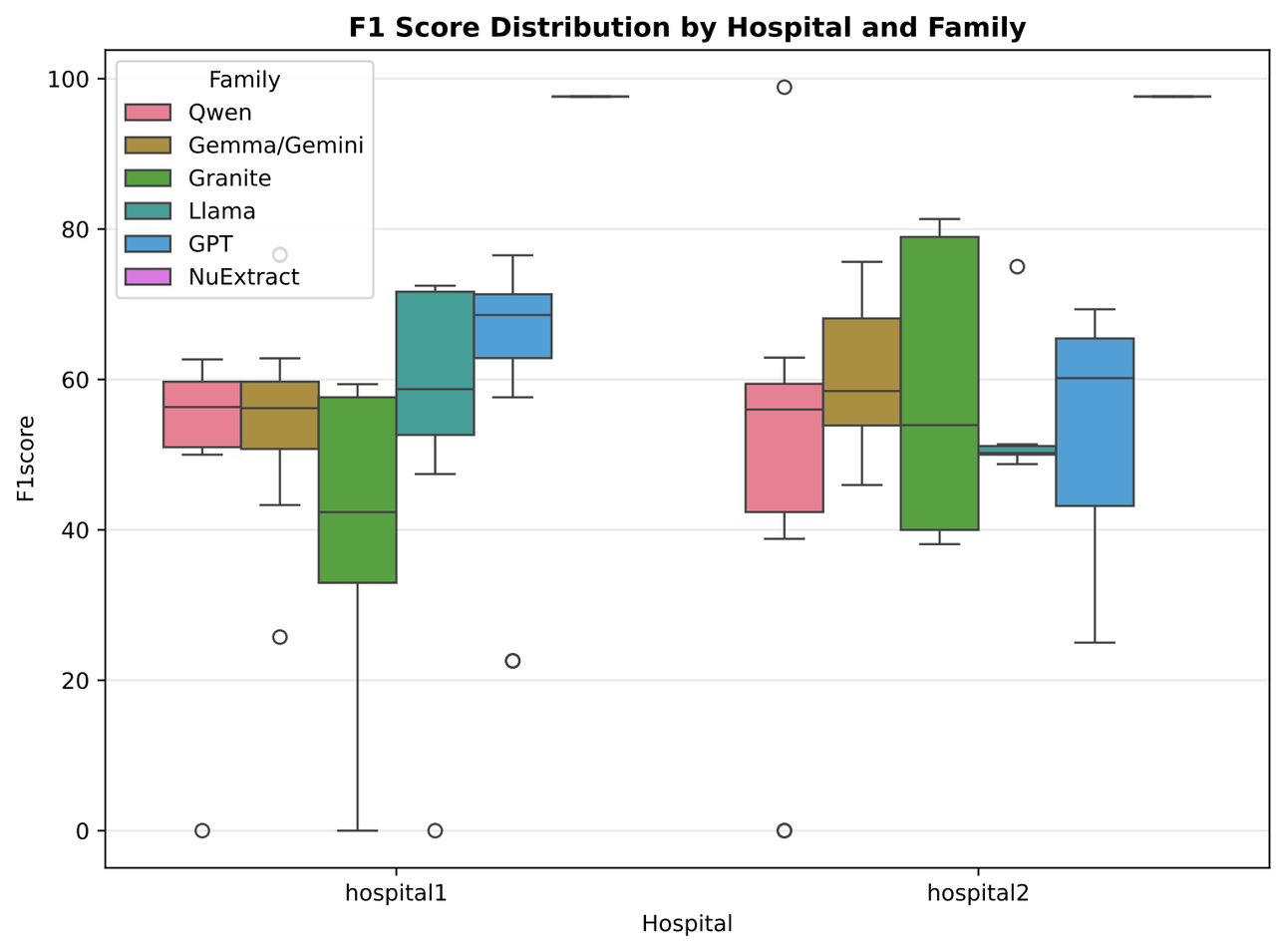

Average F1 Score: Family vs Input Type
Family
GPT
Gemma/Gemini
Granite
Llama
NuExtract
Qwen
68.14
52.27
59.52
55.47
44.97
52.70
54.88
97.62
43.51
53.85
Text-Only
Vision
Input_Type
90
80
70
60
50
Average Accuracy: Family vs Hospital
Family
GPT
Gemma/Gemini
Granite
Llama
NuExtract
Qwen
50.82
43.67
44.04
52.67
25.79
45.68
42.87
42.36
97.62
97.62
41.28
43.34
hospital1
hospital2
Hospital
90
80
70
60
50
40
30

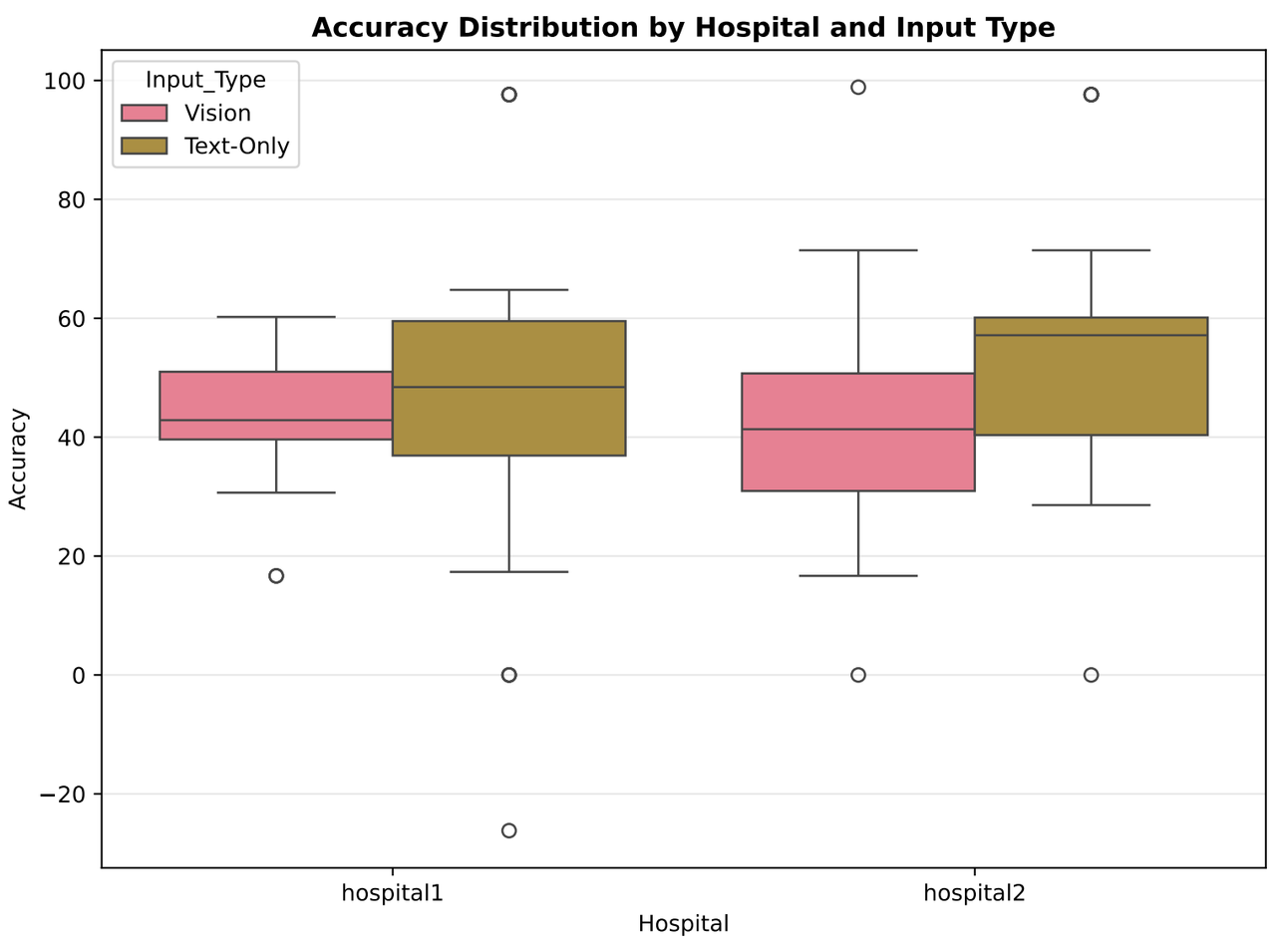

False Positives vs False Negatives by Family and Input Type
False Negatives
False Positives
0
20
40
60
80
Family
Qwen
Gemma/Gemini
Granite
Llama
GPT
NuExtract
Input_Type
Vision
Text-Only
Equal Errors Line

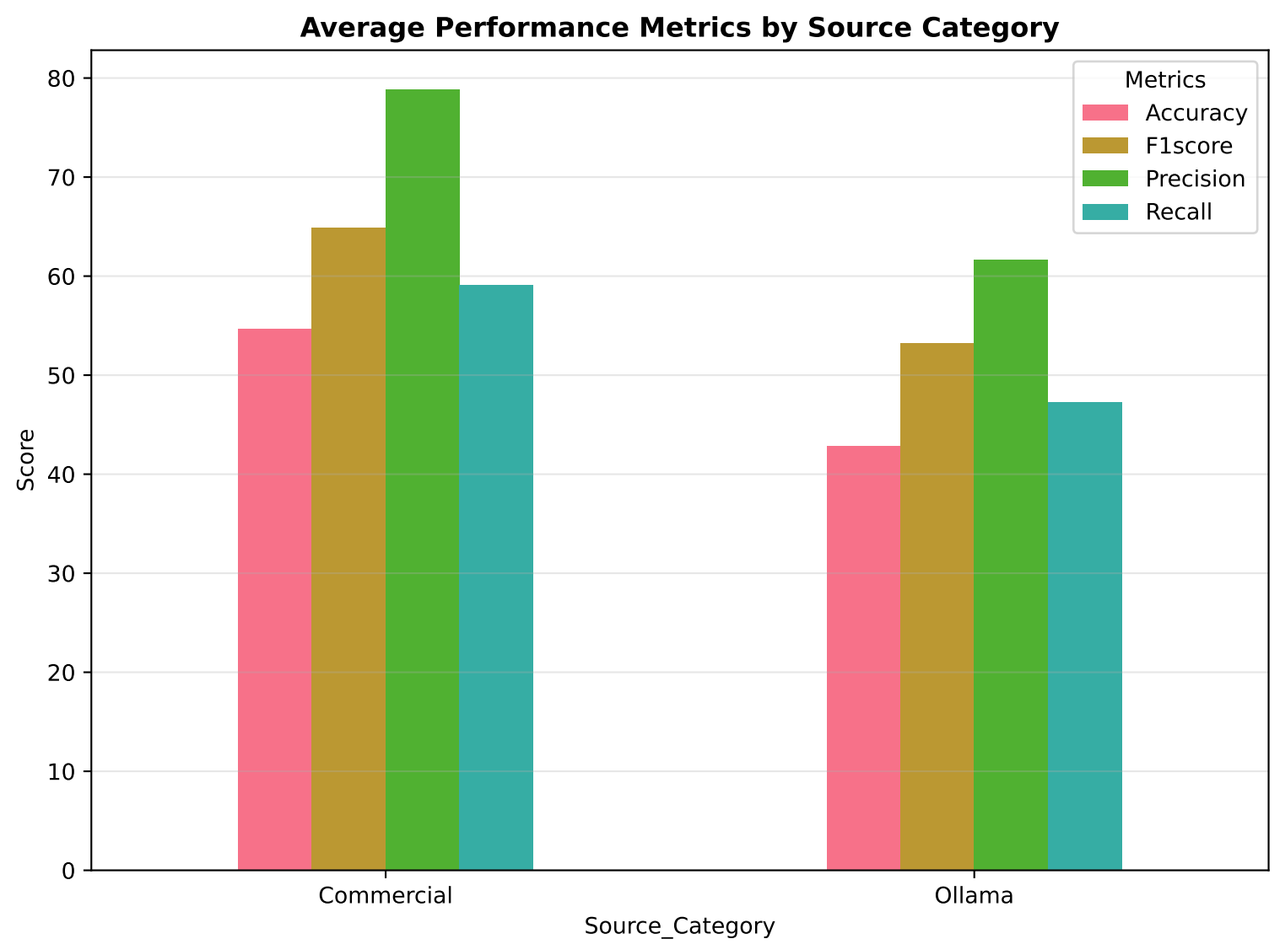

Error Distribution by Model Family
Error_Count
Error_Type
False Positives
False Negatives
Qwen
Gemma/Gemini
Granite
Llama
GPT
NuExtract
Family

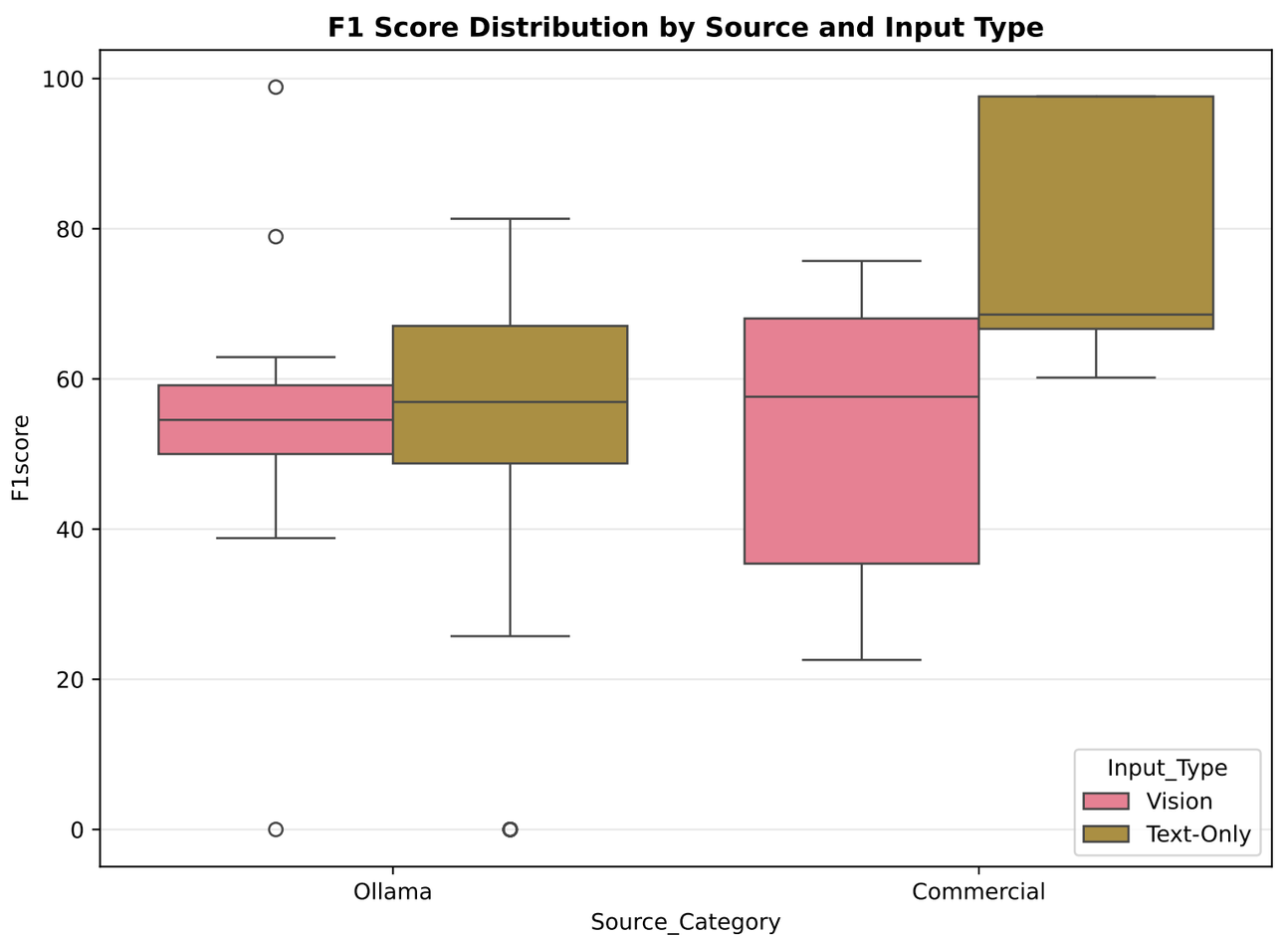

Error Ratio vs Total Errors
(Ratio = FP/FN)
False Positive / False Negative Ratio
Total_Errors
Family
Qwen
Gemma/Gemini
Granite
Llama
GPT
NuExtract
Input_Type
Vision
Text-Only
Equal FP/FN


Error Bias: FP - FN by Family and Hospital
(+ve = More FP, -ve = More FN)
Family
GPT
Gemma/Gemini
Granite
Llama
NuExtract
Qwen
-11.8
-20.5
-3.3
-11.5
-5.7
-24.9
-16.8
-16.5
0.0
0.0
-2.5
-11.9
hospital1
hospital2
Hospital


F1 Score Distribution by Hospital and Family
Family
Qwen
Gemma/Gemini
Granite
Llama
GPT
NuExtract
F1score
100
80
60
40
20
0
hospital1
hospital2
Hospital


Accuracy Distribution by Hospital and Input Type
Input_Type
Vision
Text-Only
Accuracy
100
80
60
40
20
0
−20
hospital1
hospital2
Hospital


Average Performance Metrics by Source Category
Metrics
Accuracy
F1score
Precision
Recall
Score
0
10
20
30
40
50
60
70
80
Commercial
Ollama
Source_Category


F1 Score Distribution by Source and Input Type
F1score
0
20
40
60
80
100
Ollama
Commercial
Source_Category
Input_Type
Vision
Text-Only