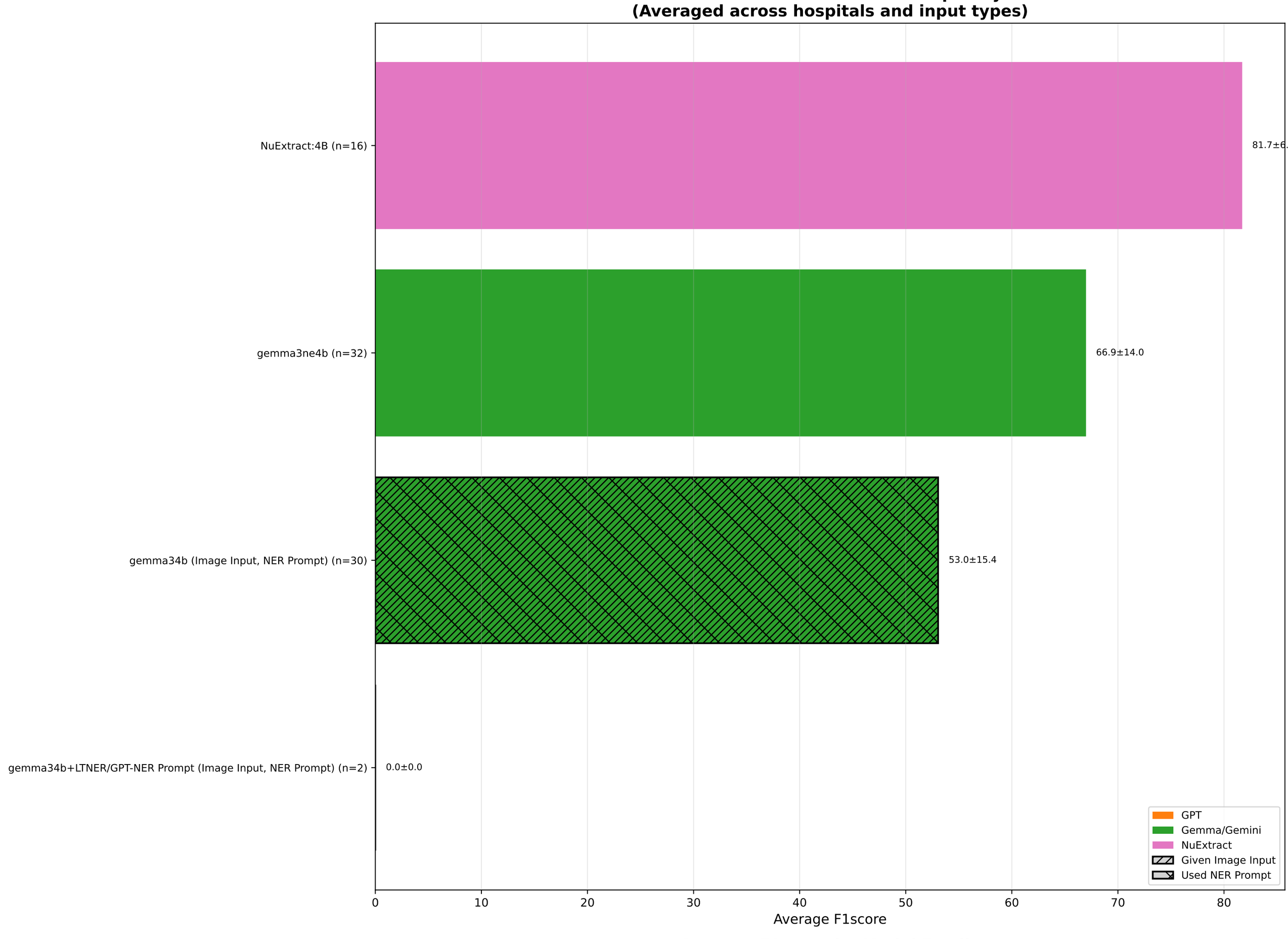
COMPREHENSIVE LLM PERFORMANCE ANALYSIS
==========
Report Generated: 2025-07-07 22:50:26

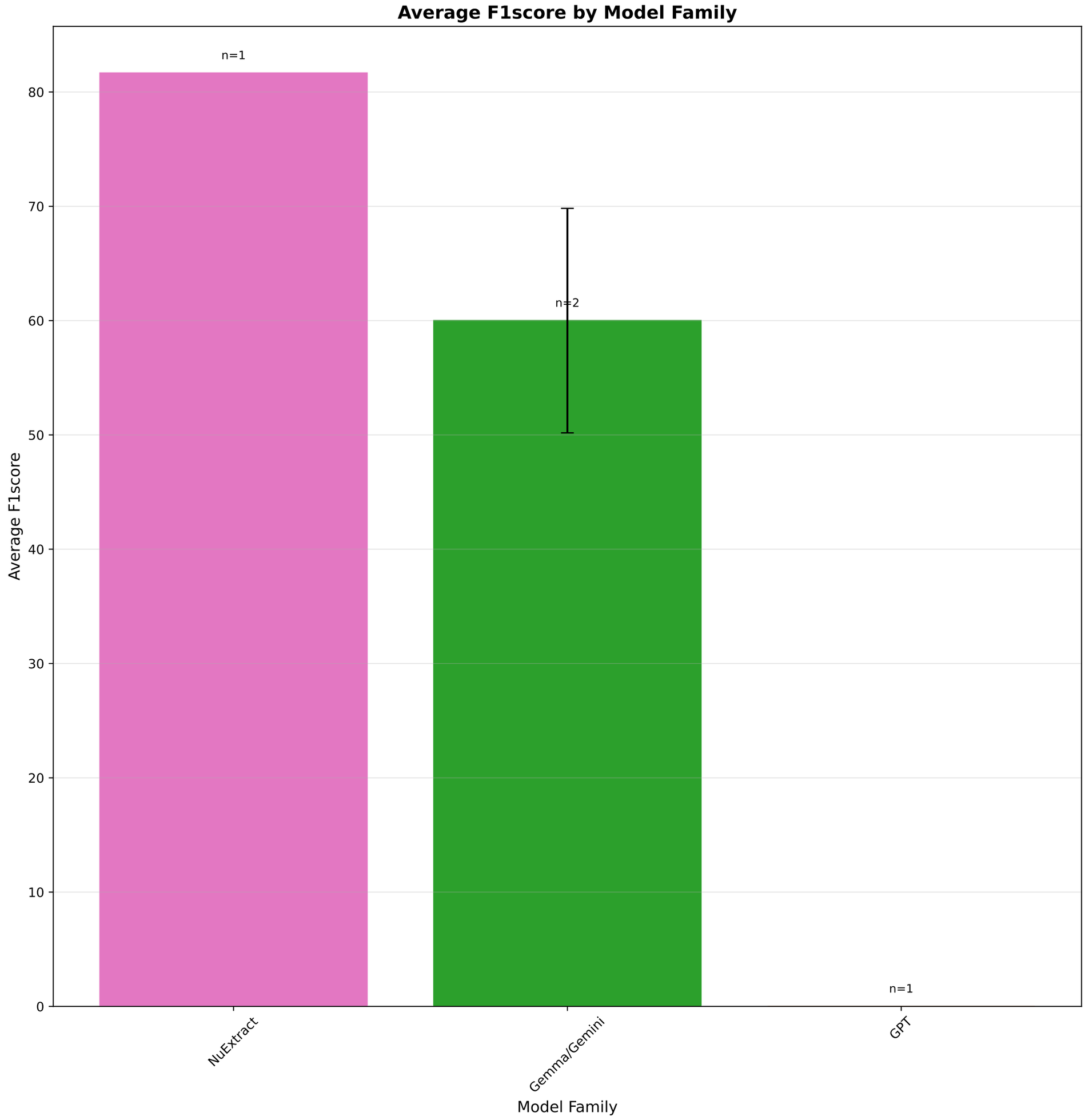
OVERALL PERFORMANCE METRICS:
- Average F1 Score: 62.999 (Std: 19.634)
- Average Accuracy: 48.572 (Std: 18.590)
- Average Precision: 70.781 (Std: 21.963)
- Average Recall: 59.142 (Std: 21.967)

GROUPED MODEL F1 SCORE STATISTICS:
- Unique Base Models: 4
- Total Test Instances: 80
- Best Performing Model: NuExtract:4B (F1: 81.66)
- Worst Performing Model: gemma34b+LTNER/GPT-NER Prompt (F1: 0.00)
- Overall Average F1: 50.41
- Models with Vision: 2

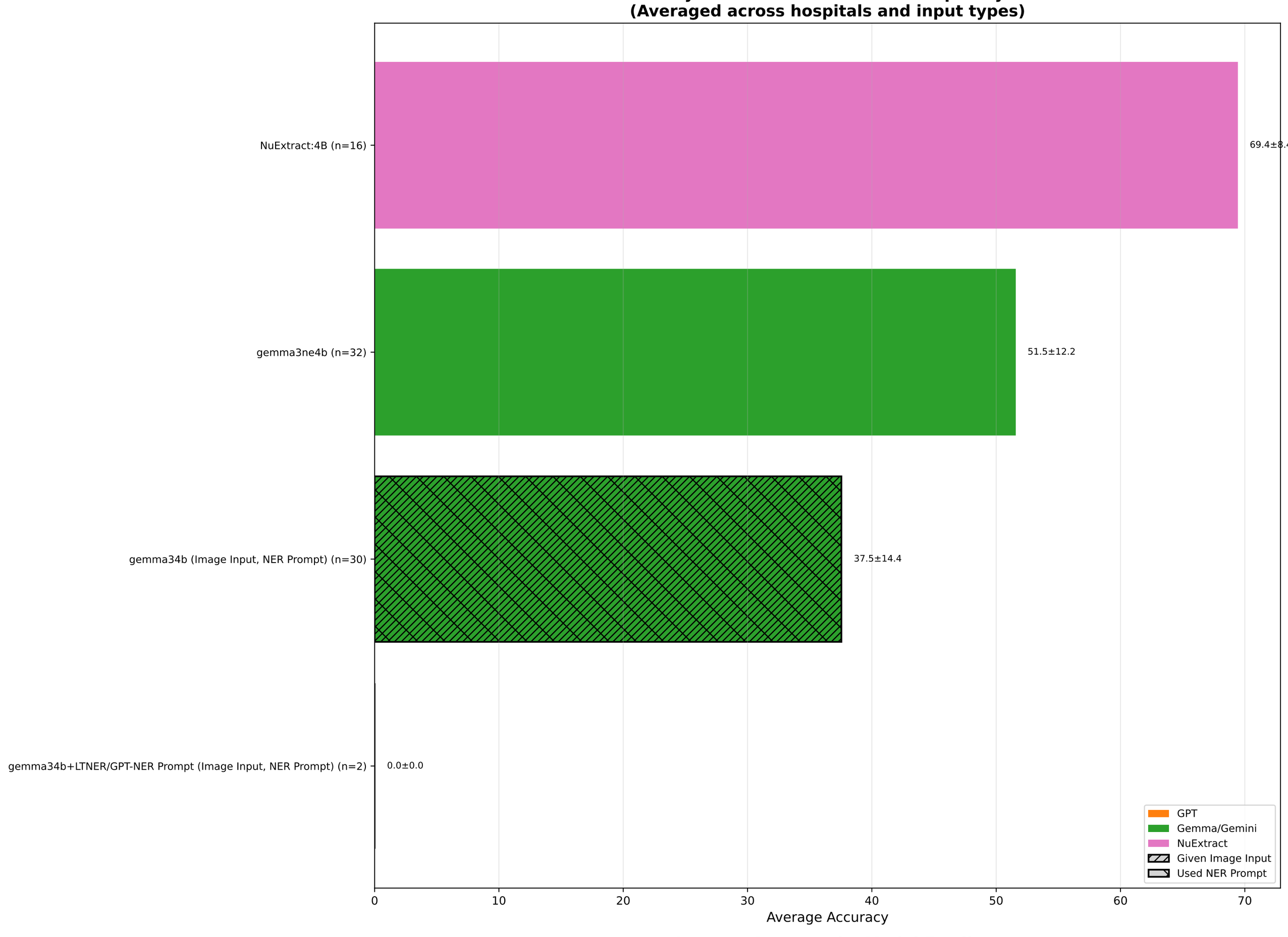
Overall F1score Performance - Models Grouped by Base Name (Averaged across hospitals and input types)



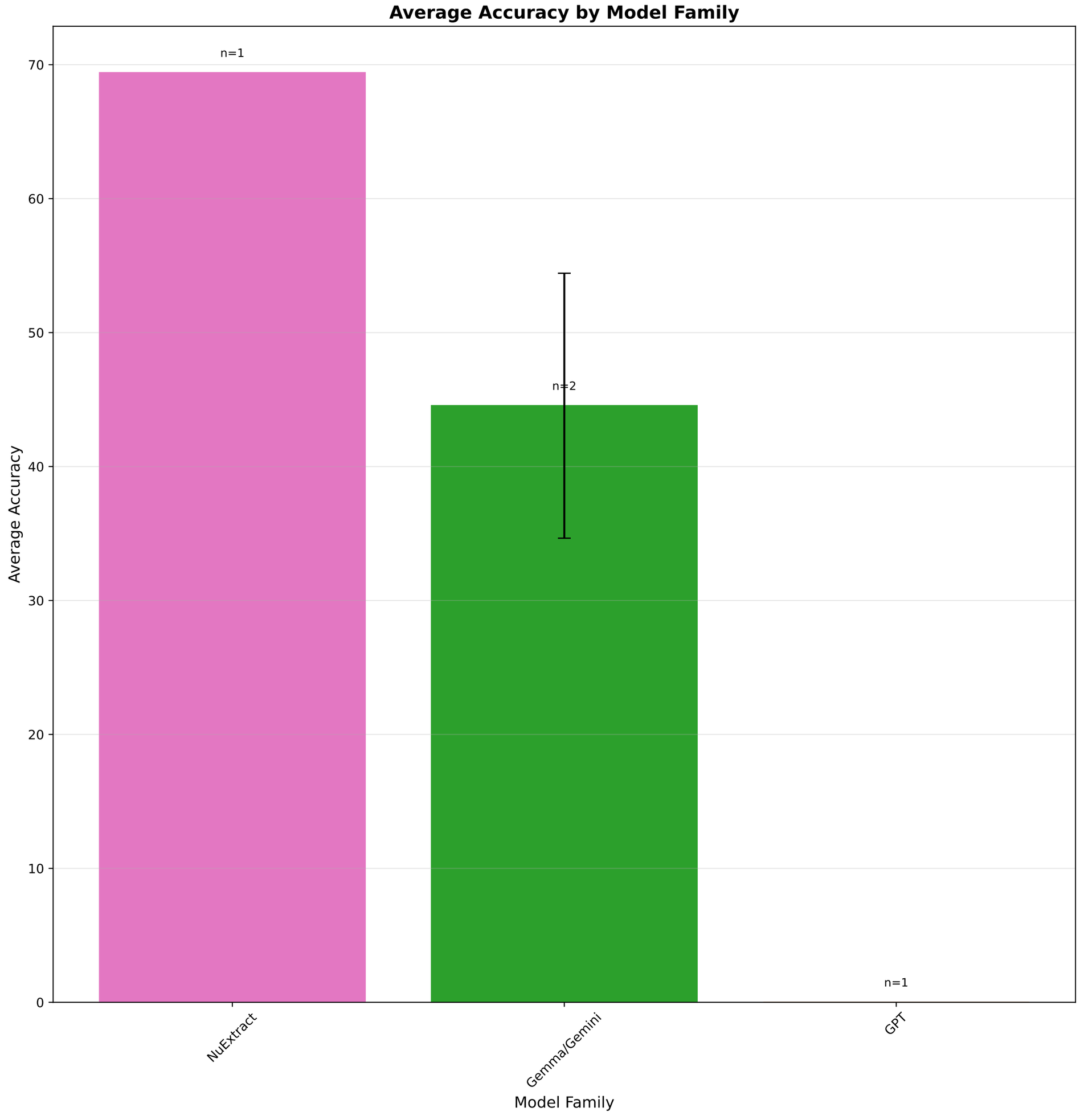
Average F1score by Model Family



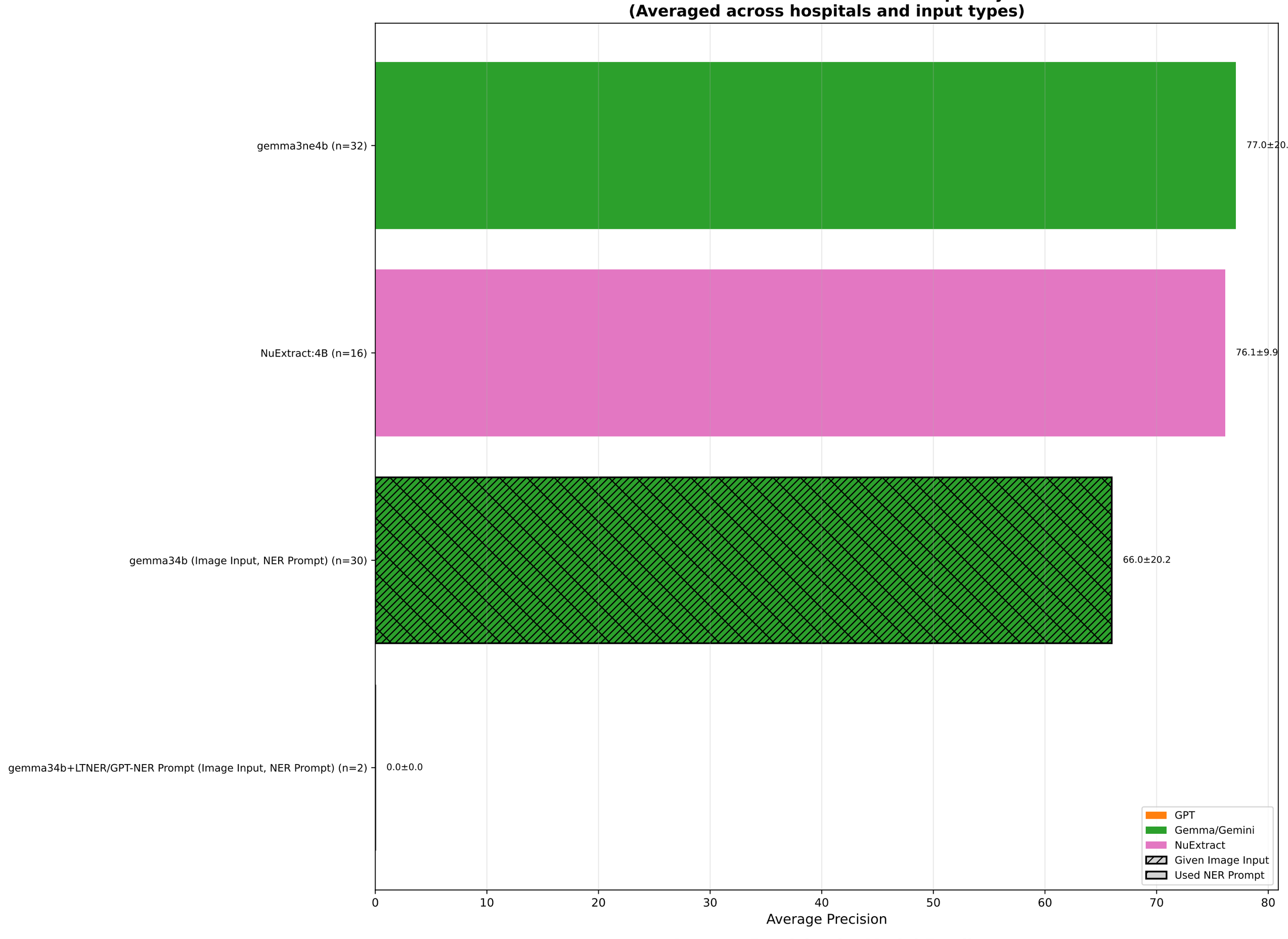
Overall Accuracy Performance - Models Grouped by Base Name
(Averaged across hospitals and input types)



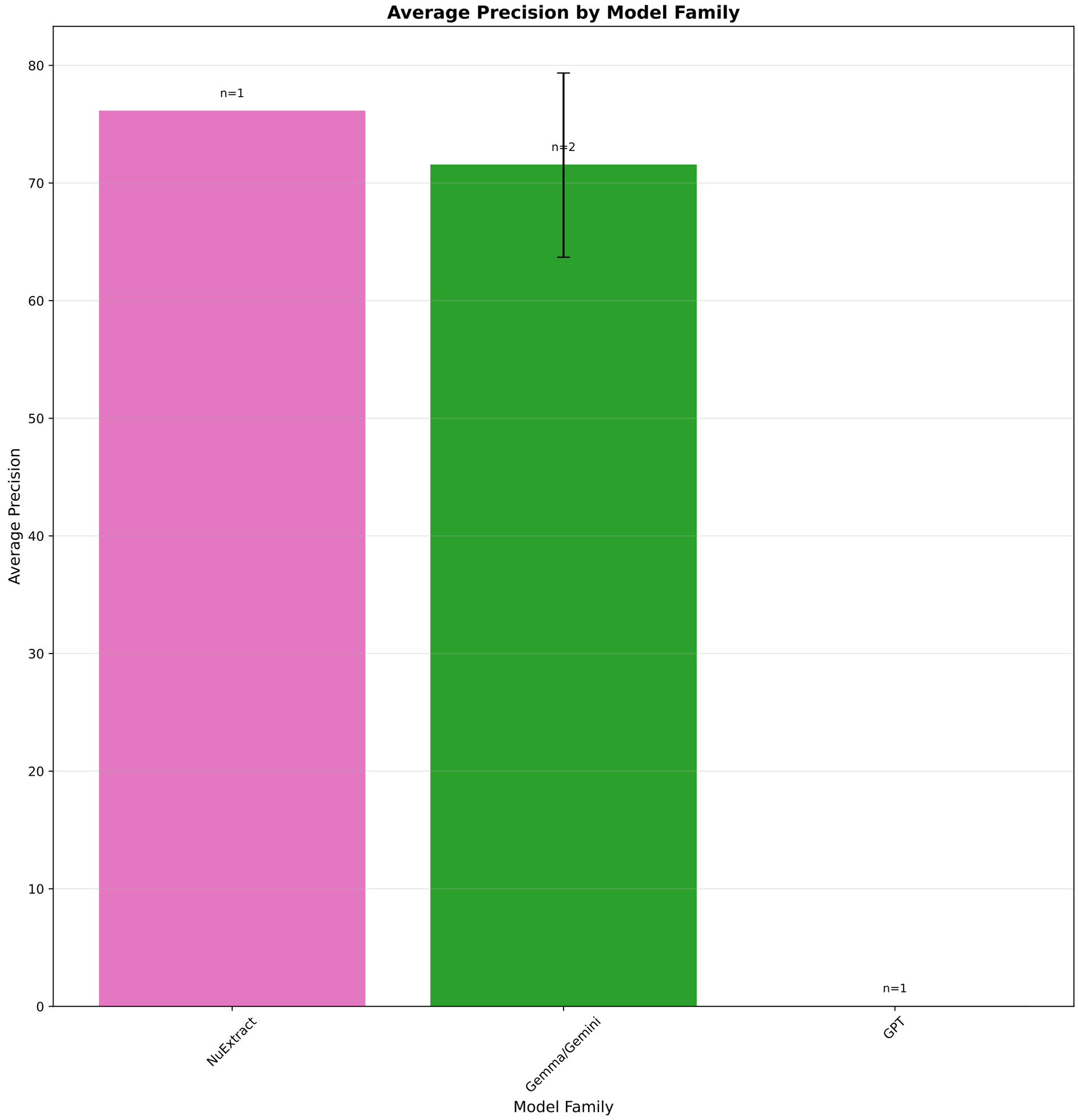
Average Accuracy by Model Family



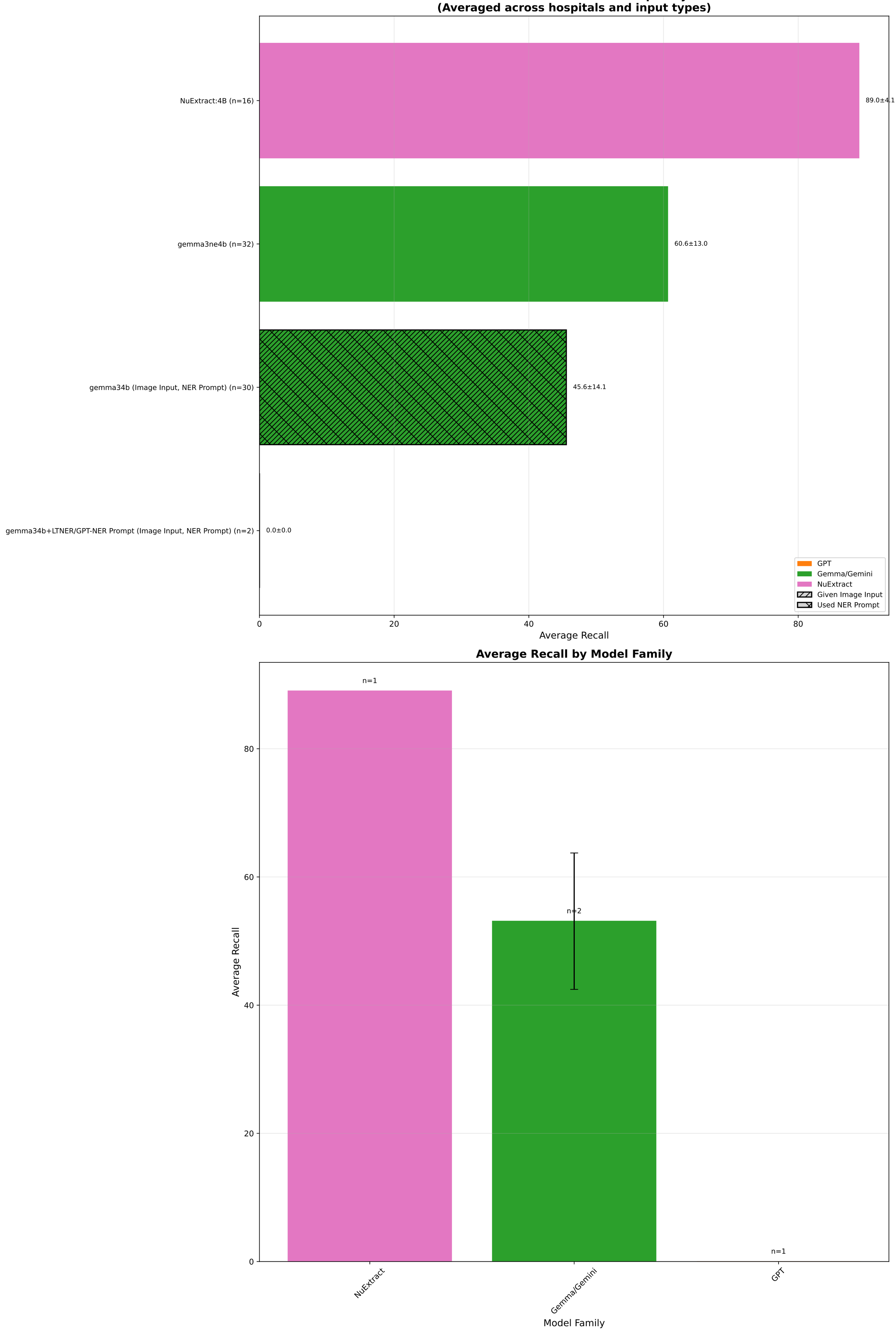
Overall Precision Performance - Models Grouped by Base Name
(Averaged across hospitals and input types)



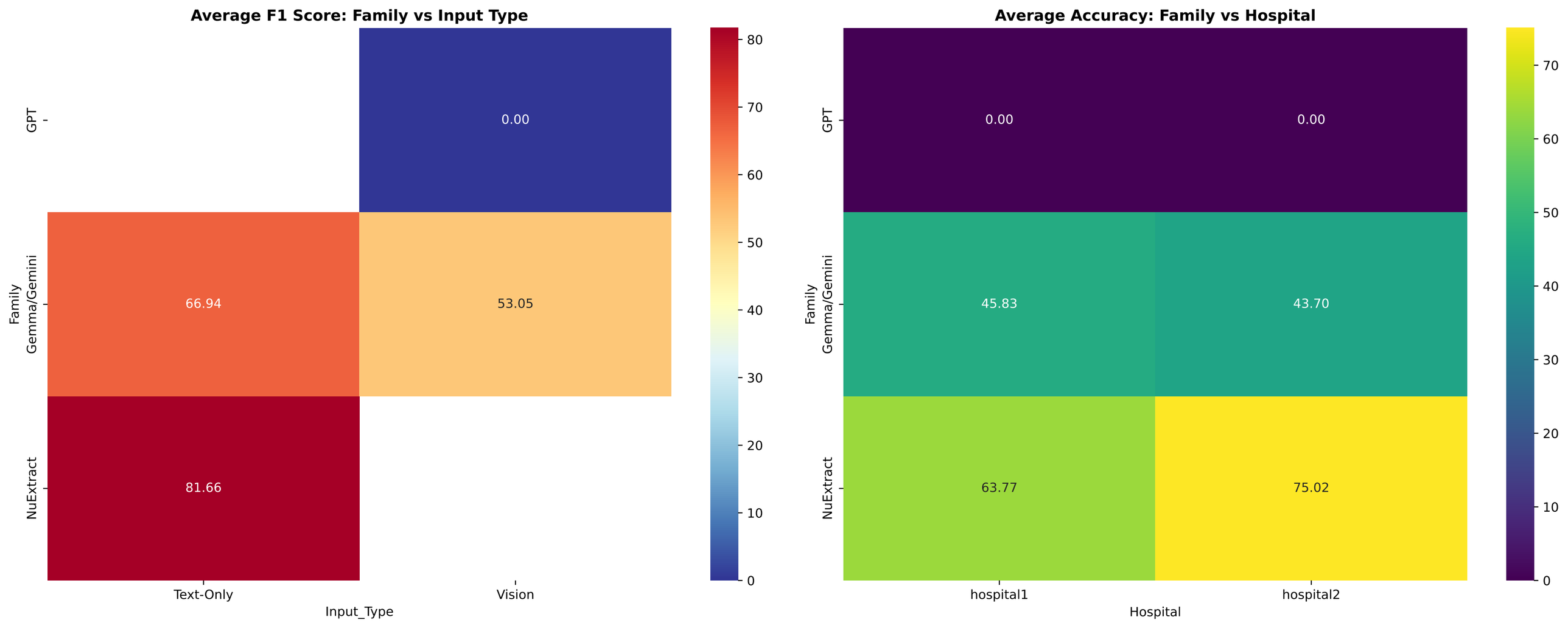
Average Precision by Model Family

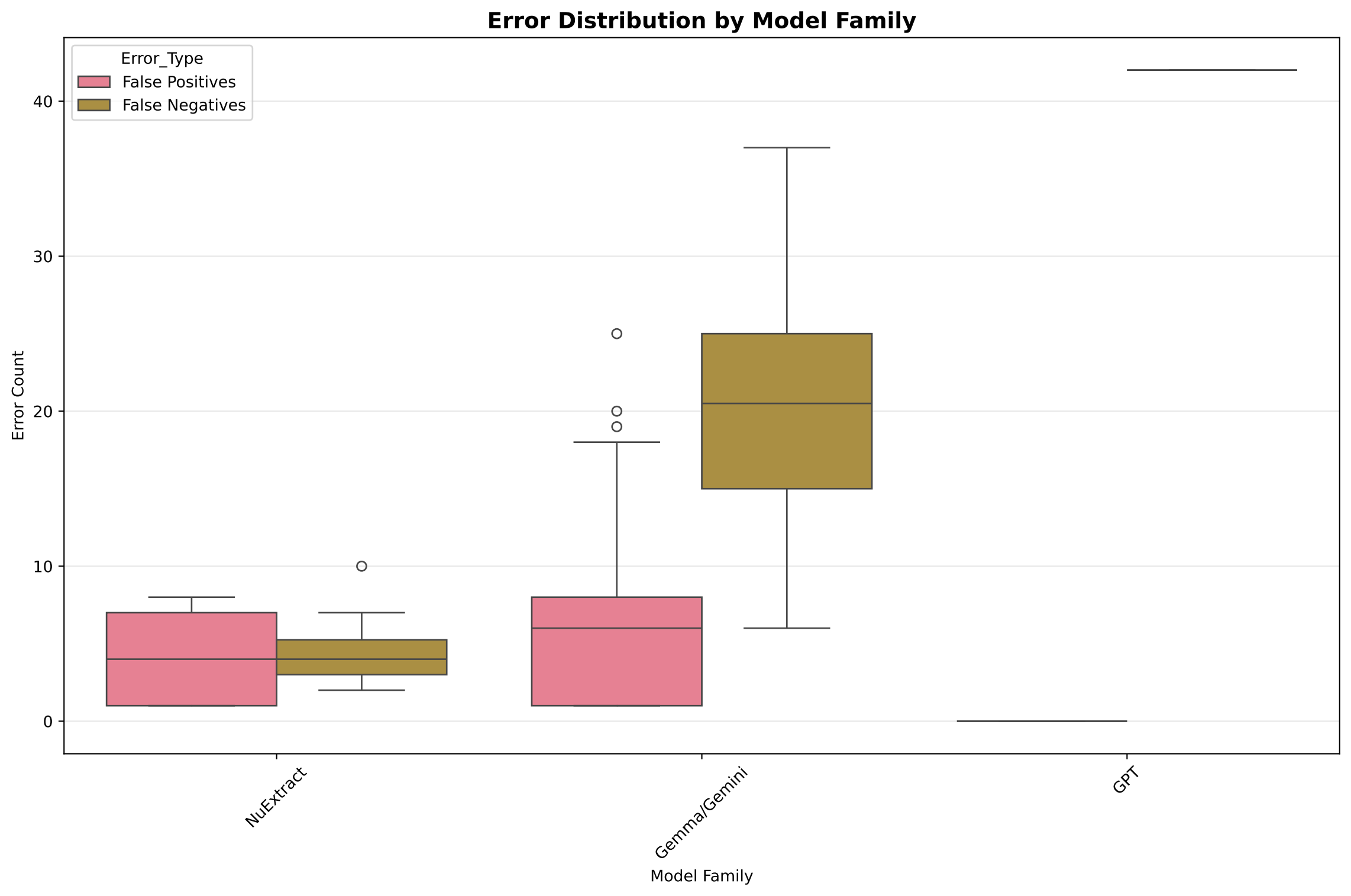


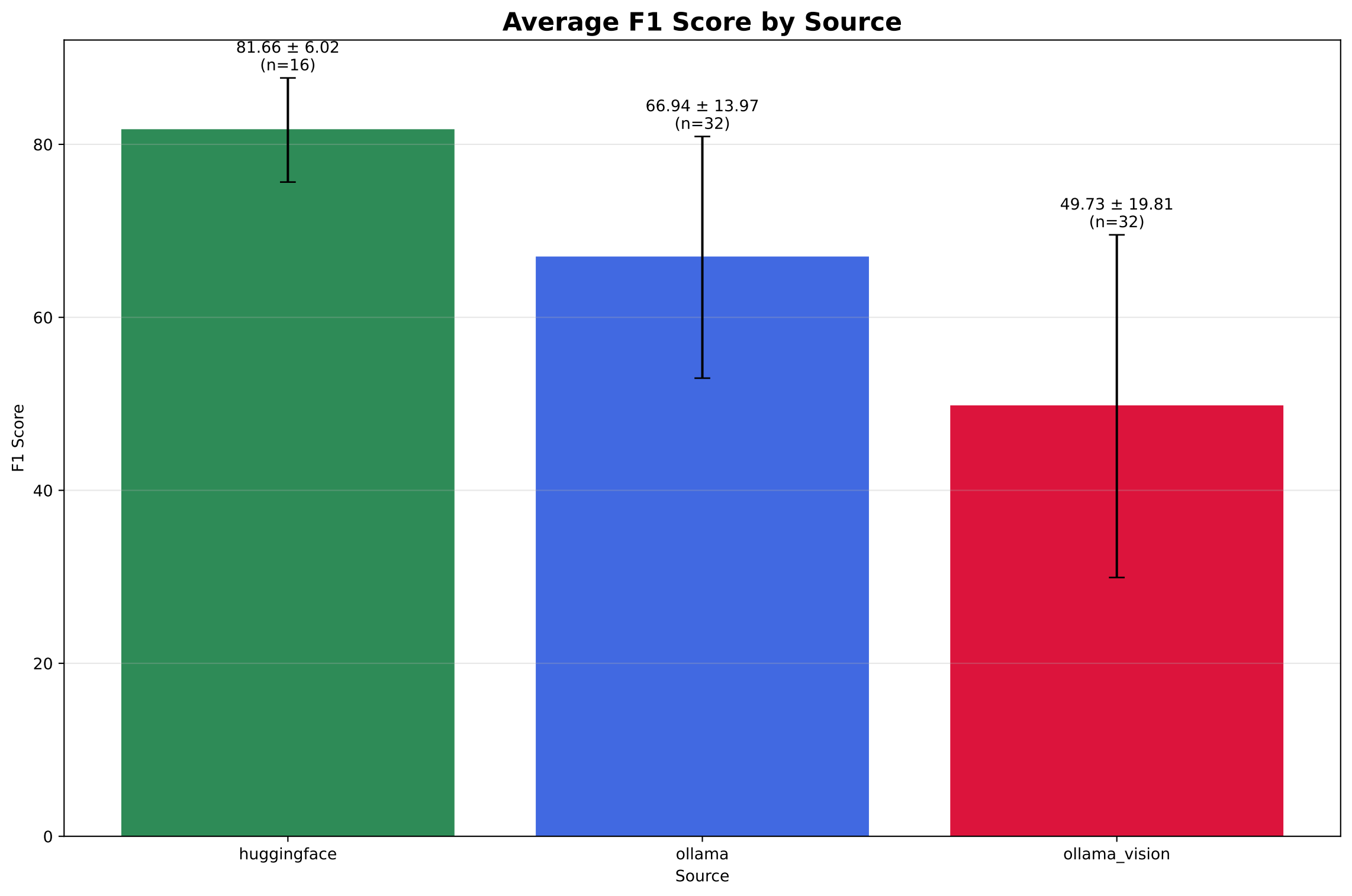
Overall Recall Performance - Models Grouped by Base Name
(Averaged across hospitals and input types)

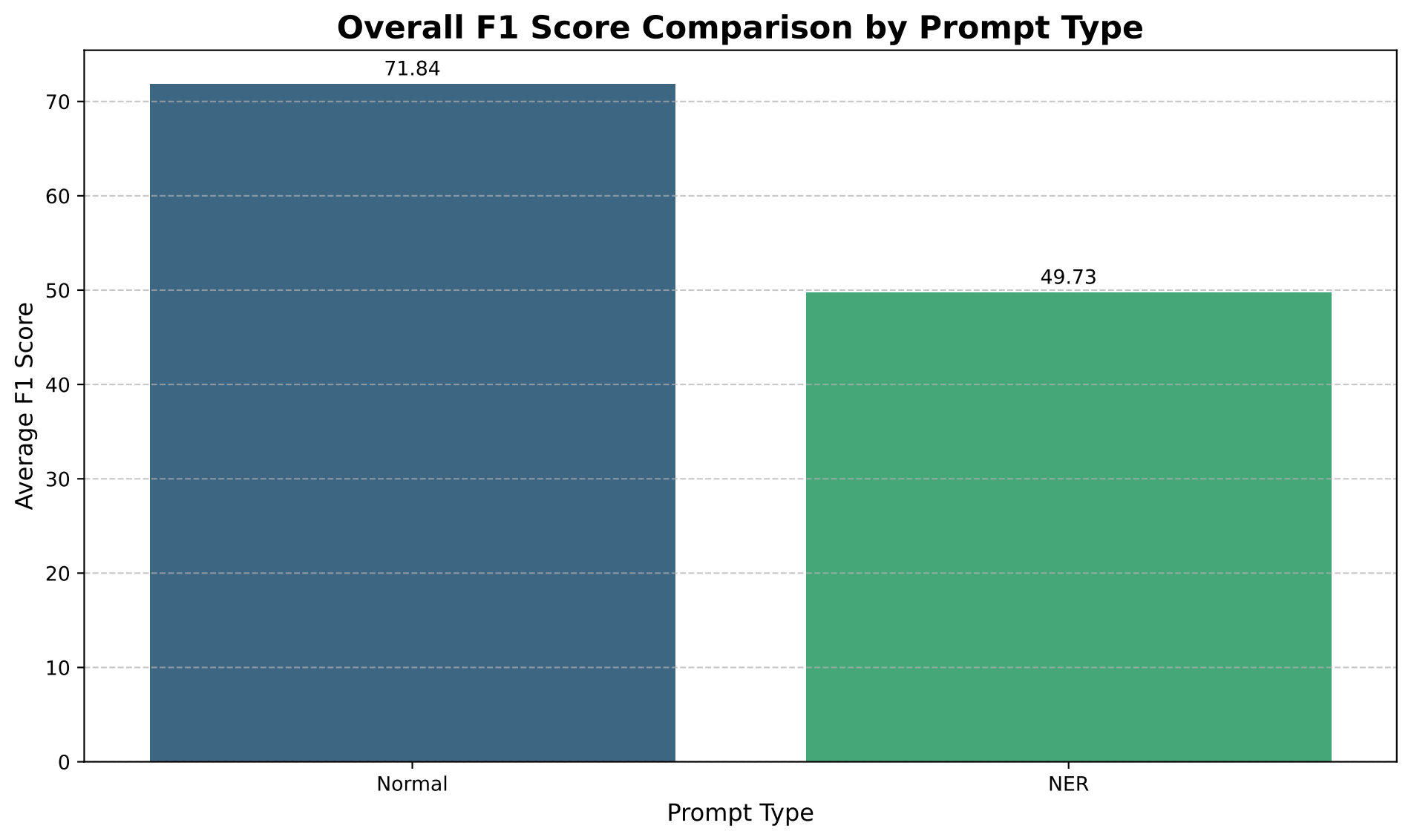


Average Recall by Model Family


Average F1 Score: Family vs Input Type
GPT
Gemma/Gemini
NuExtract
Family
0.00
66.94
53.05
81.66
Text-Only
Vision
Input_Type
80
70
60
50
40
30
20
10
0
Average Accuracy: Family vs Hospital
GPT
Gemma/Gemini
NuExtract
Family
0.00
0.00
45.83
43.70
63.77
75.02
hospital1
hospital2
Hospital
70
60
50
40
30
20
10
0


Error Distribution by Model Family
Error_Type
False Positives
False Negatives
Error Count
0
10
20
30
40
NuExtract
Gemma/Gemini
GPT
Model Family


Average F1 Score by Source
81.66 ± 6.02
(n=16)
66.94 ± 13.97
(n=32)
49.73 ± 19.81
(n=32)
F1 Score
0
20
40
60
80
huggingface
ollama
ollama_vision
Source


Overall F1 Score Comparison by Prompt Type
71.84
49.73
Average F1 Score
0
10
20
30
40
50
60
70
Normal
NER
Prompt Type