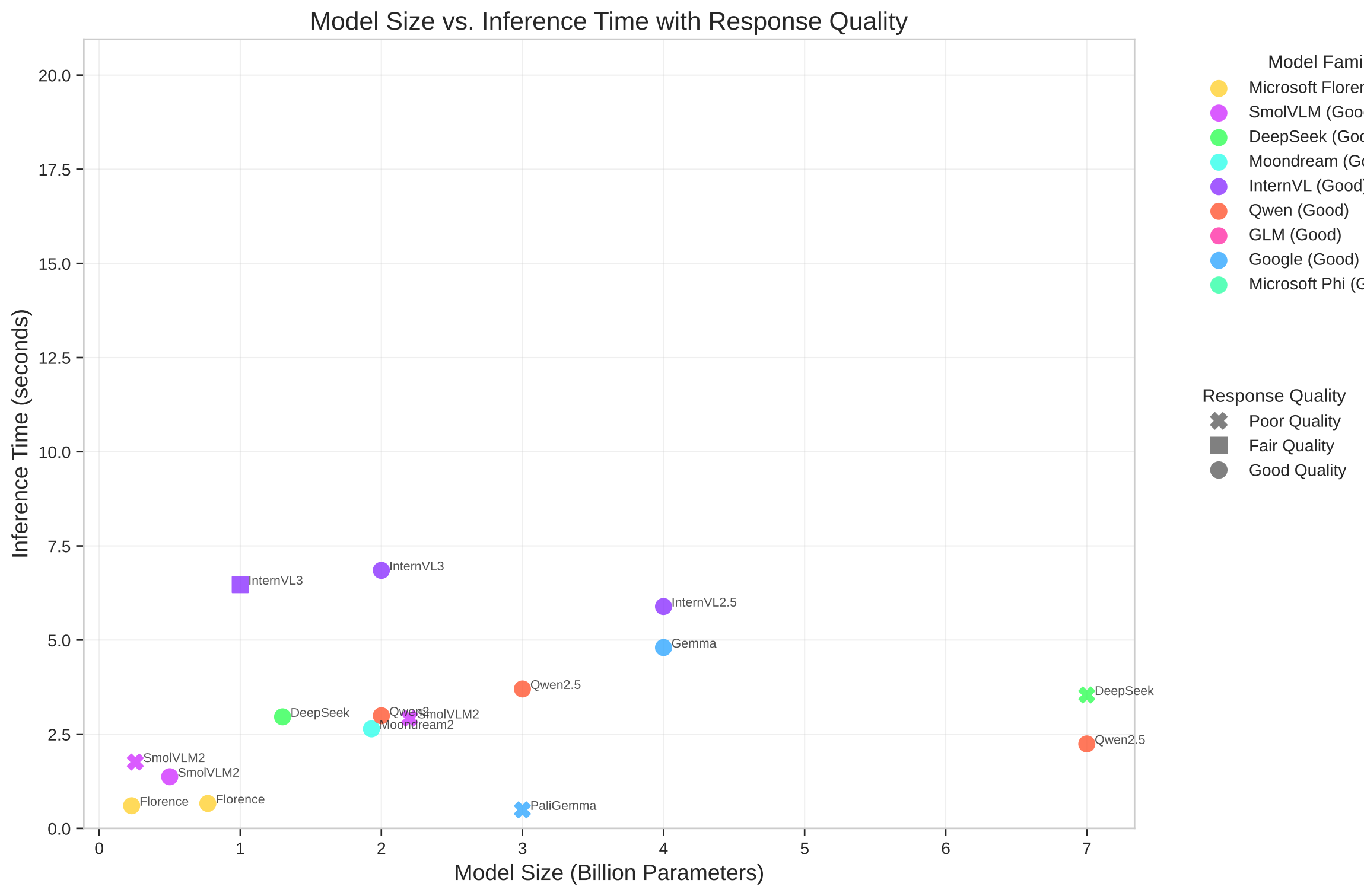
Model Size vs. Inference Time with Response Quality
Inference Time (seconds)
Model Size (Billion Parameters)
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
0
1
2
3
4
5
6
7
InternVL3
InternVL3
InternVL2.5
Gemma
Qwen2.5
DeepSeek
DeepSeek
Qwen2
SmolVLM2
Moondream2
Qwen2.5
SmolVLM2
SmolVLM2
Florence
Florence
PaliGemma
Model Fami
Microsoft Flore
SmolVLM (Goo
DeepSeek (Goo
Moondream (G
InternVL (Good)
Qwen (Good)
GLM (Good)
Google (Good)
Microsoft Phi (G
Response Quality
Poor Quality
Fair Quality
Good Quality