Figure 2: Temporal Prediction & Multi-Agent LLM Performance Validation
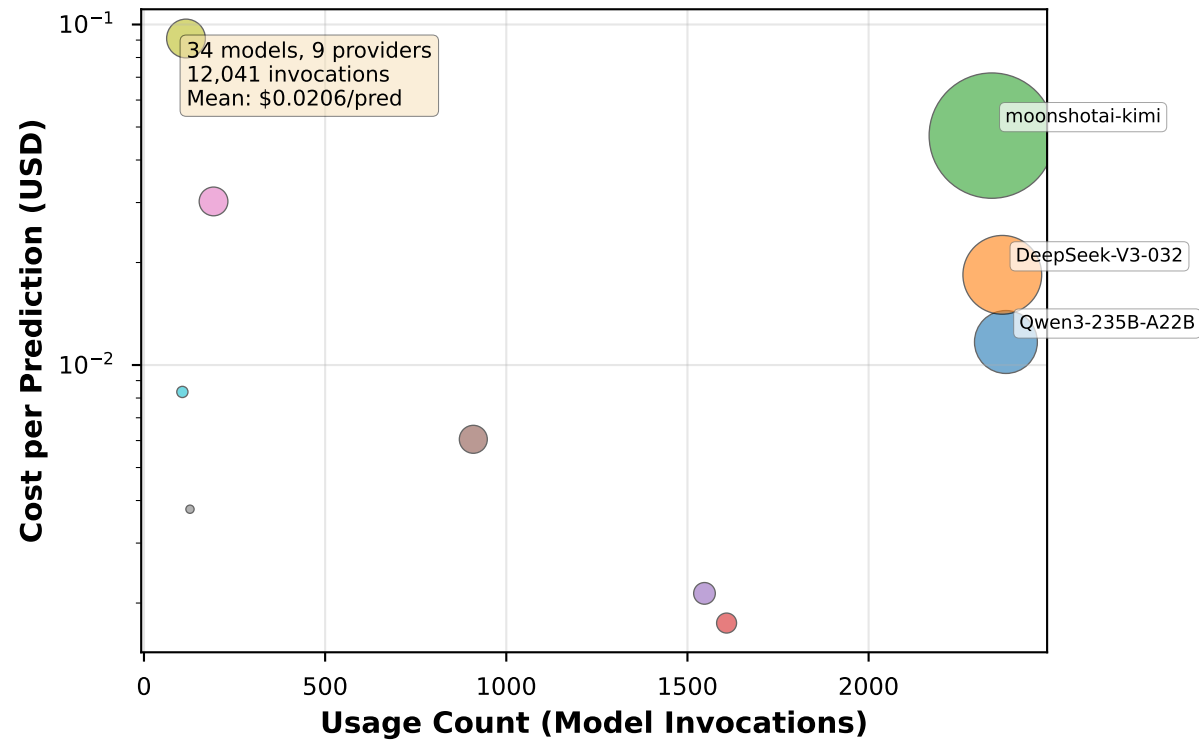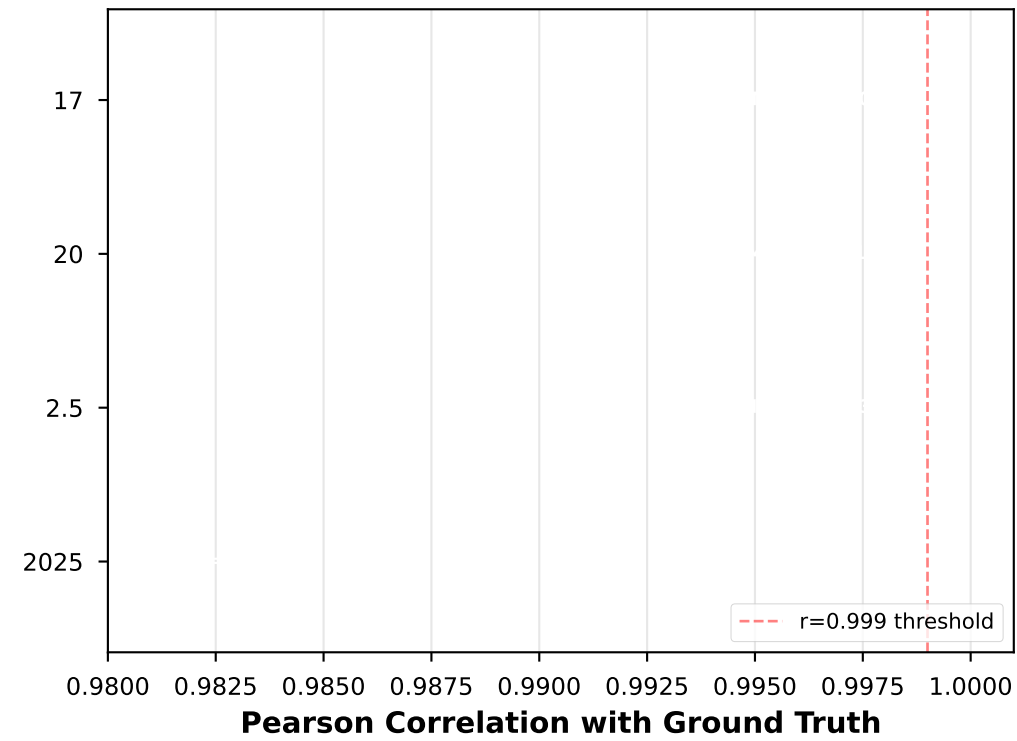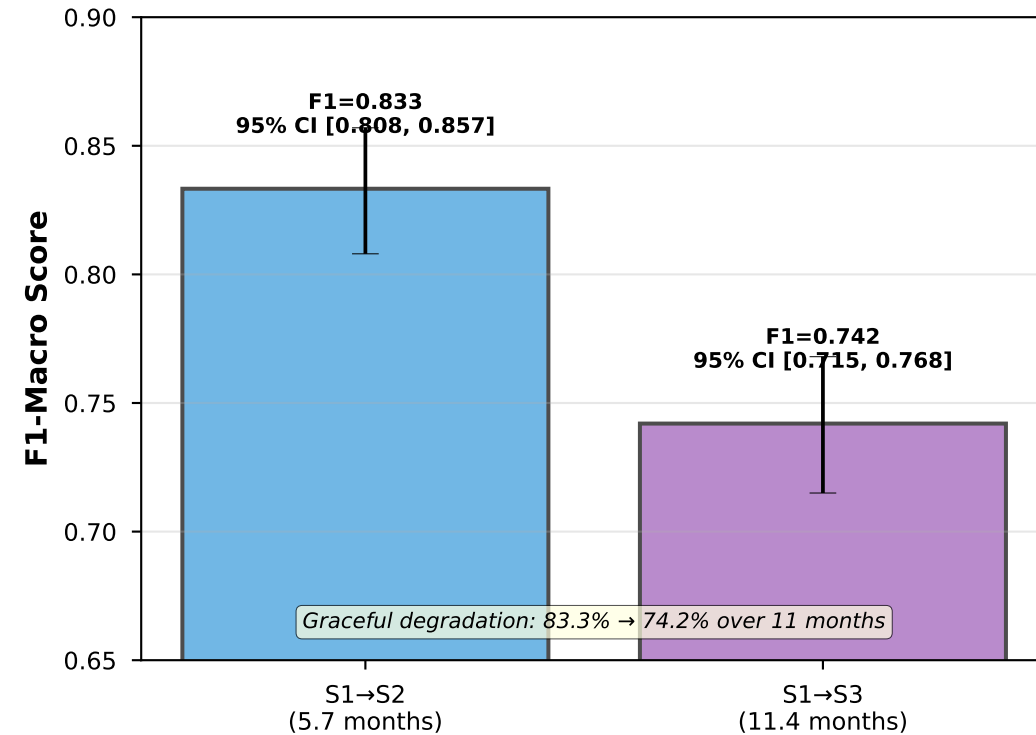
Panel A: Multi-Agent LLM System Cost-Efficiency

34 models, 9 providers
12,041 invocations
Mean: $0.0206/pred

moonshotai-kimi
DeepSeek-V3-032
Qwen3-235B-A22B

Cost per Prediction (USD)
Usage Count (Model Invocations)

Panel B: Individual Gemini Model Accuracy

r=0.999 threshold

Pearson Correlation with Ground Truth

Panel C: Temporal Prediction Accuracy Decay

F1=0.833
95% CI [0.808, 0.857]

F1=0.742
95% CI [0.715, 0.768]

Graceful degradation: 83.3% → 74.2% over 11 months

F1-Macro Score

S1→S2
(5.7 months)

S1→S3
(11.4 months)

Panel D: Per-Domain Temporal Stability

Others   0.993 (n=70)
Art   0.993 (n=69)
Academic   0.986 (n=68)
Service   0.970 (n=67)
Sport   0.903 (n=59)
Leadership   0.860 (n=52)
Technology   0.129 (n=19)

Excellent (F1>0.95)
Good (0.85<F1<0.95)
Poor (F1<0.85)

S1→S2 F1 Score (n=70 test children)