DrugDevBench Pipeline Test Report

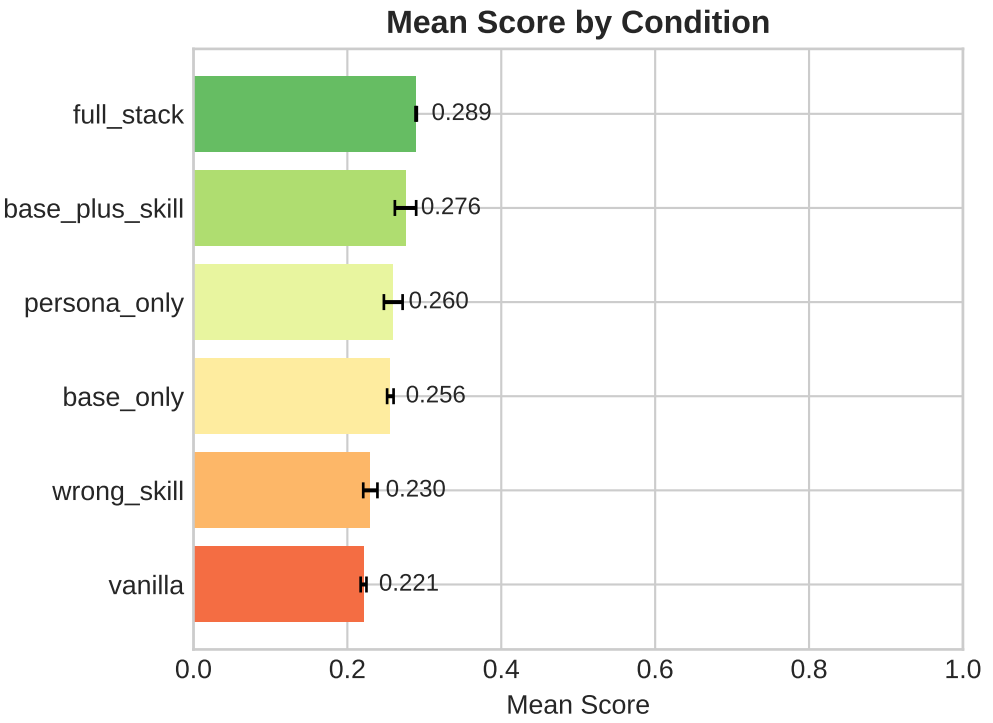
Evaluating LLM Interpretation of Drug Development Figures

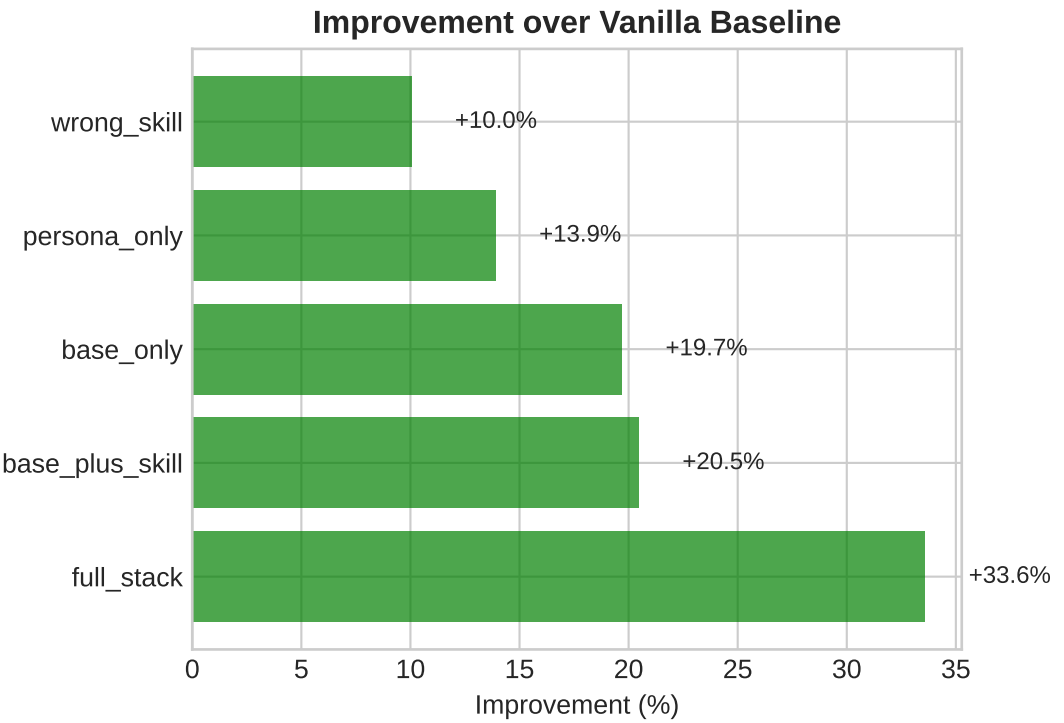
Generated: 2026-01-07 08:30

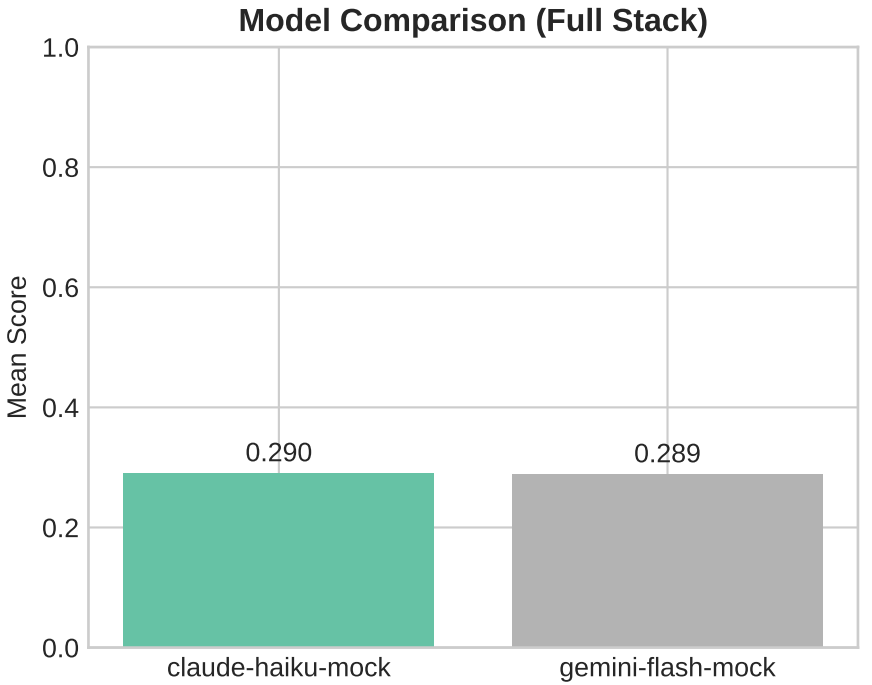
Models: claude-haiku-mock, gemini-flash-mock

Total Evaluations: 3,576

Executive Summary



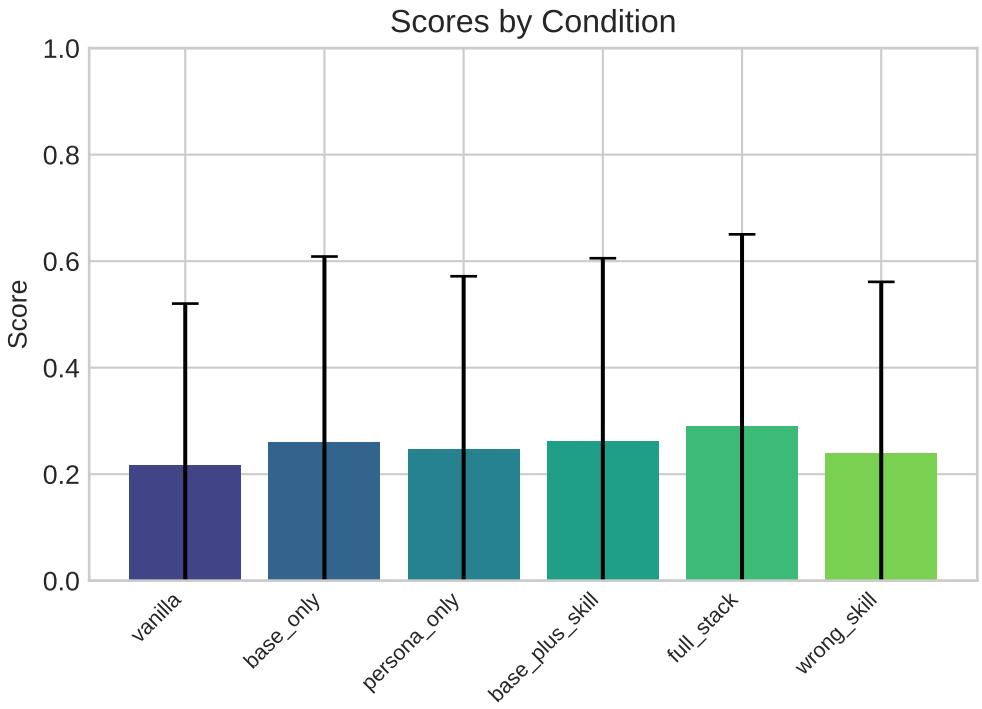


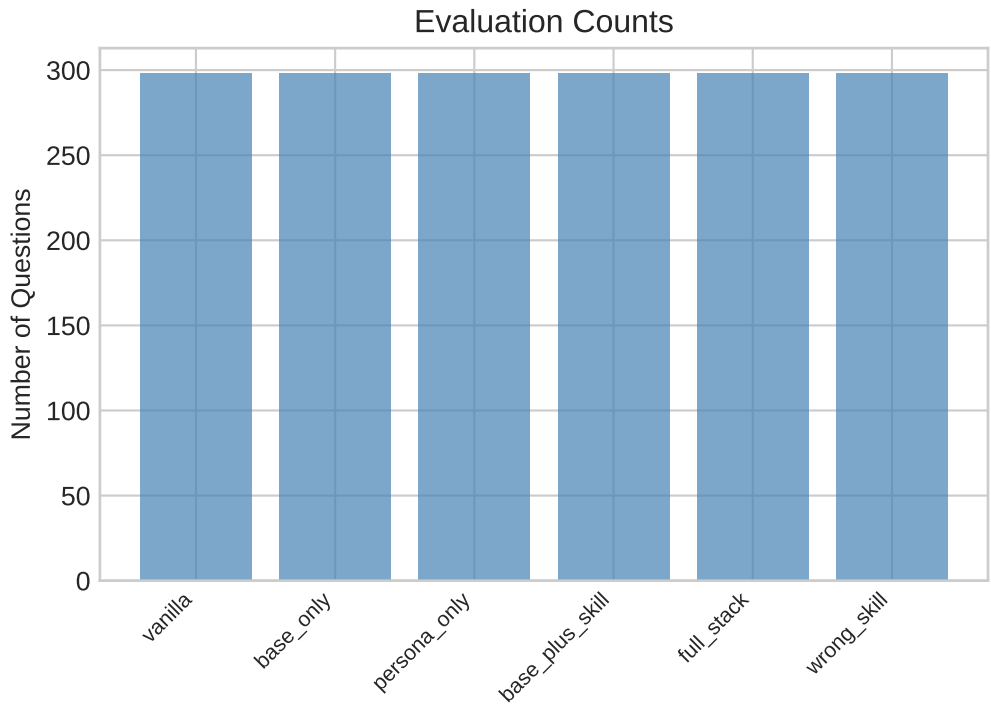


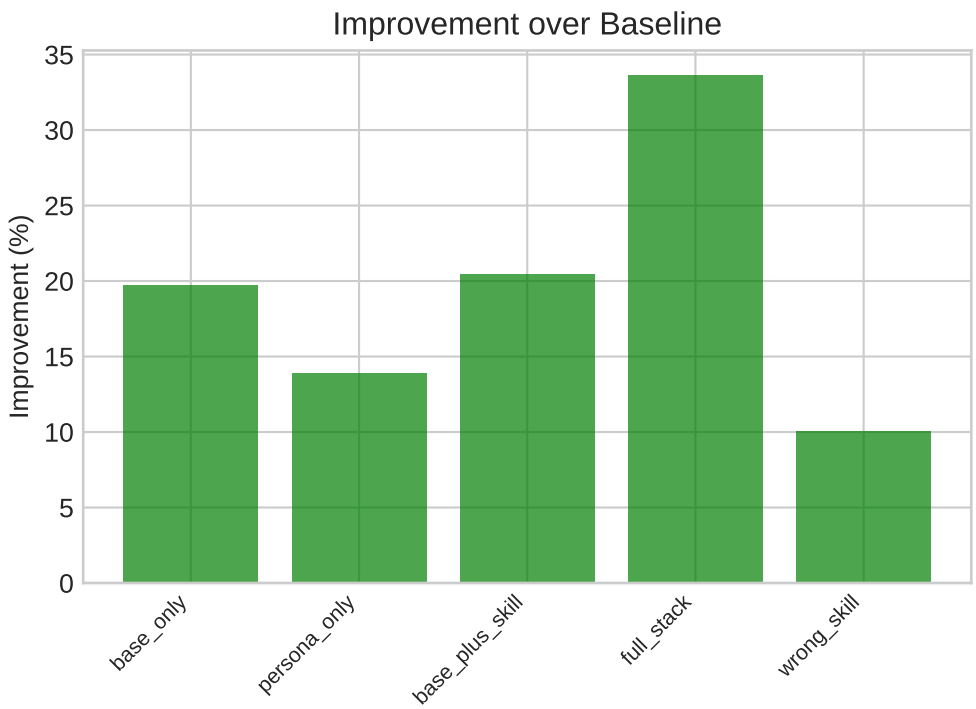
Key Findings:

- Best condition: full_stack (0.290)
- Full stack avg improvement: +31.0%
- Skills add: +24.6% over base
- Wrong skill penalty: +4.1%

Detailed Results: claude-haiku-mock



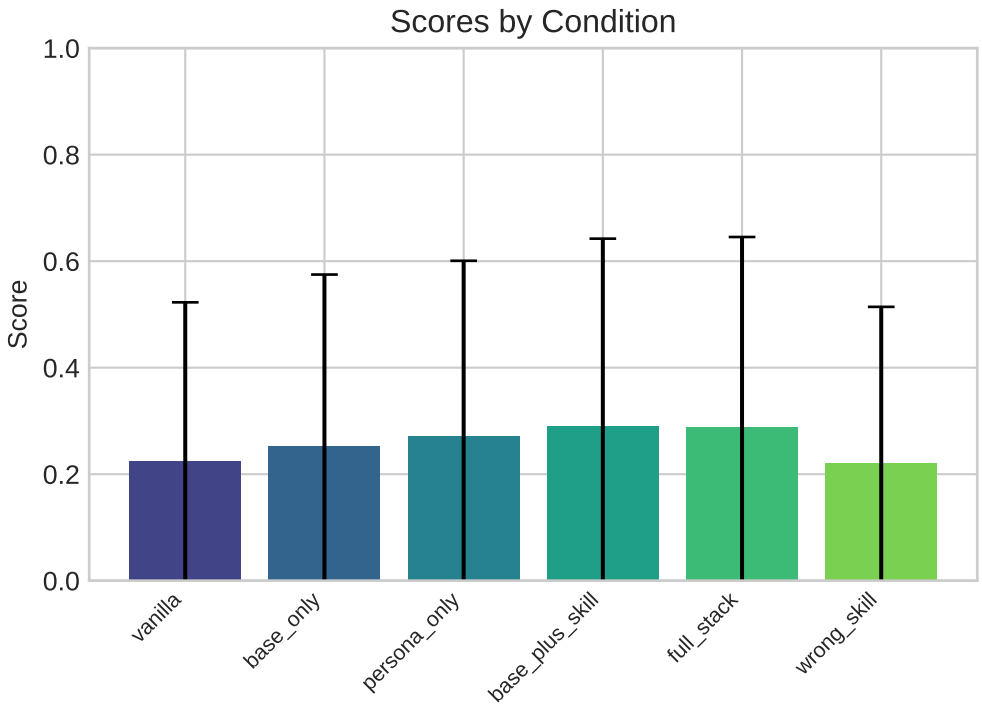


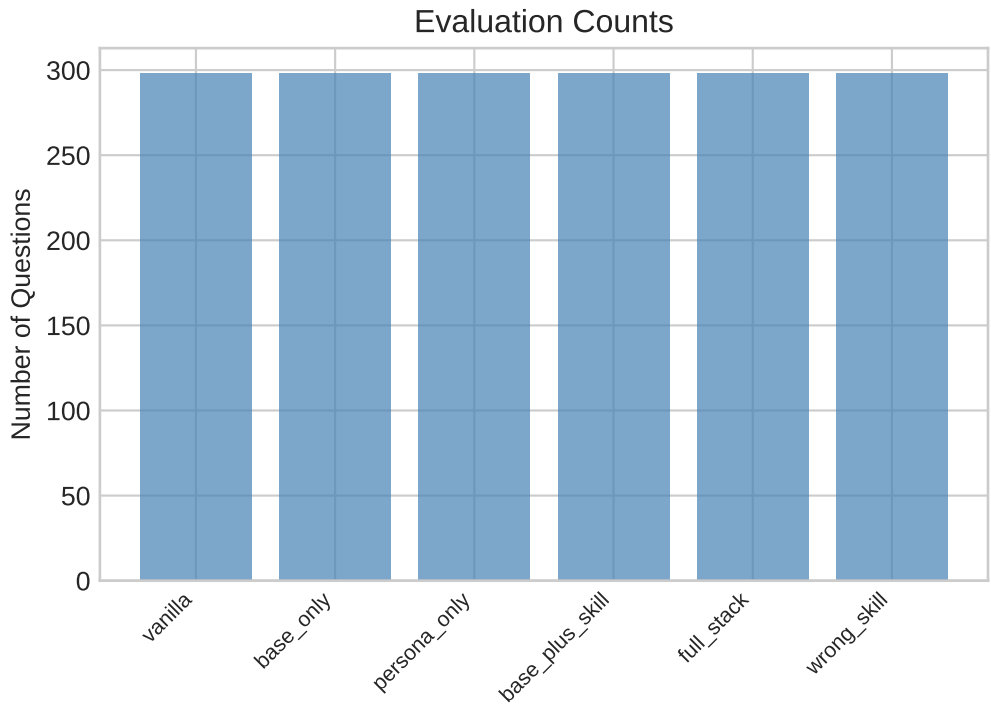


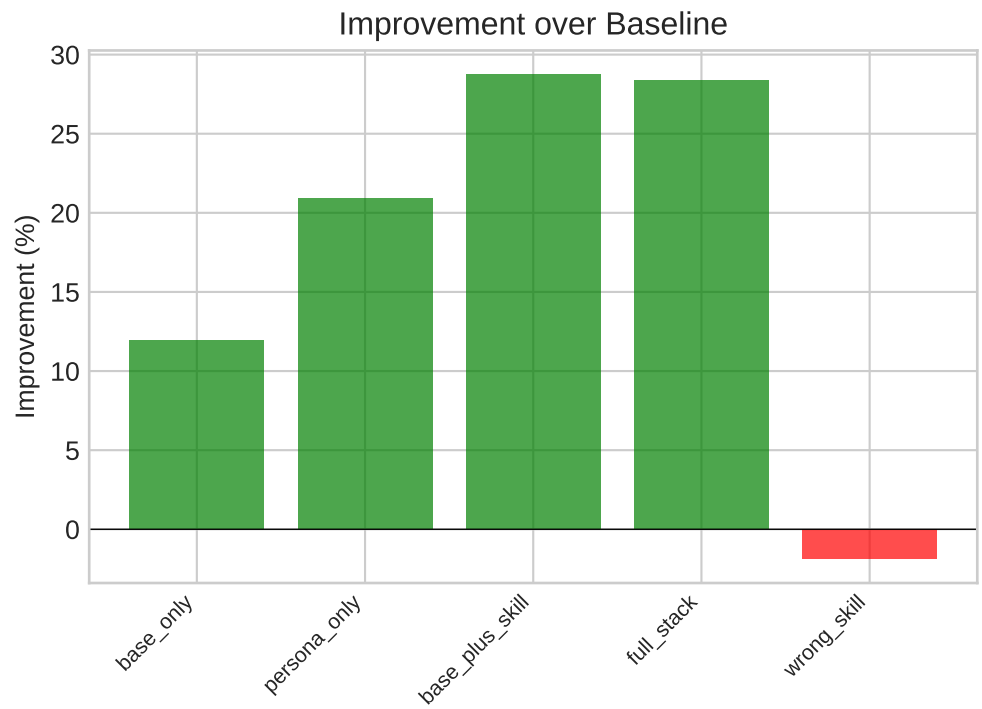
Summary Statistics

Condition	Mean	Std	N	Cost
vanilla	0.217	0.303	298	$0.1688
base_only	0.260	0.349	298	$0.2888
persona_only	0.247	0.324	298	$0.2766
base_plus_skill	0.262	0.344	298	$0.4497
full_stack	0.290	0.360	298	$0.5654
wrong_skill	0.239	0.322	298	$0.4716

Detailed Results: gemini-flash-mock

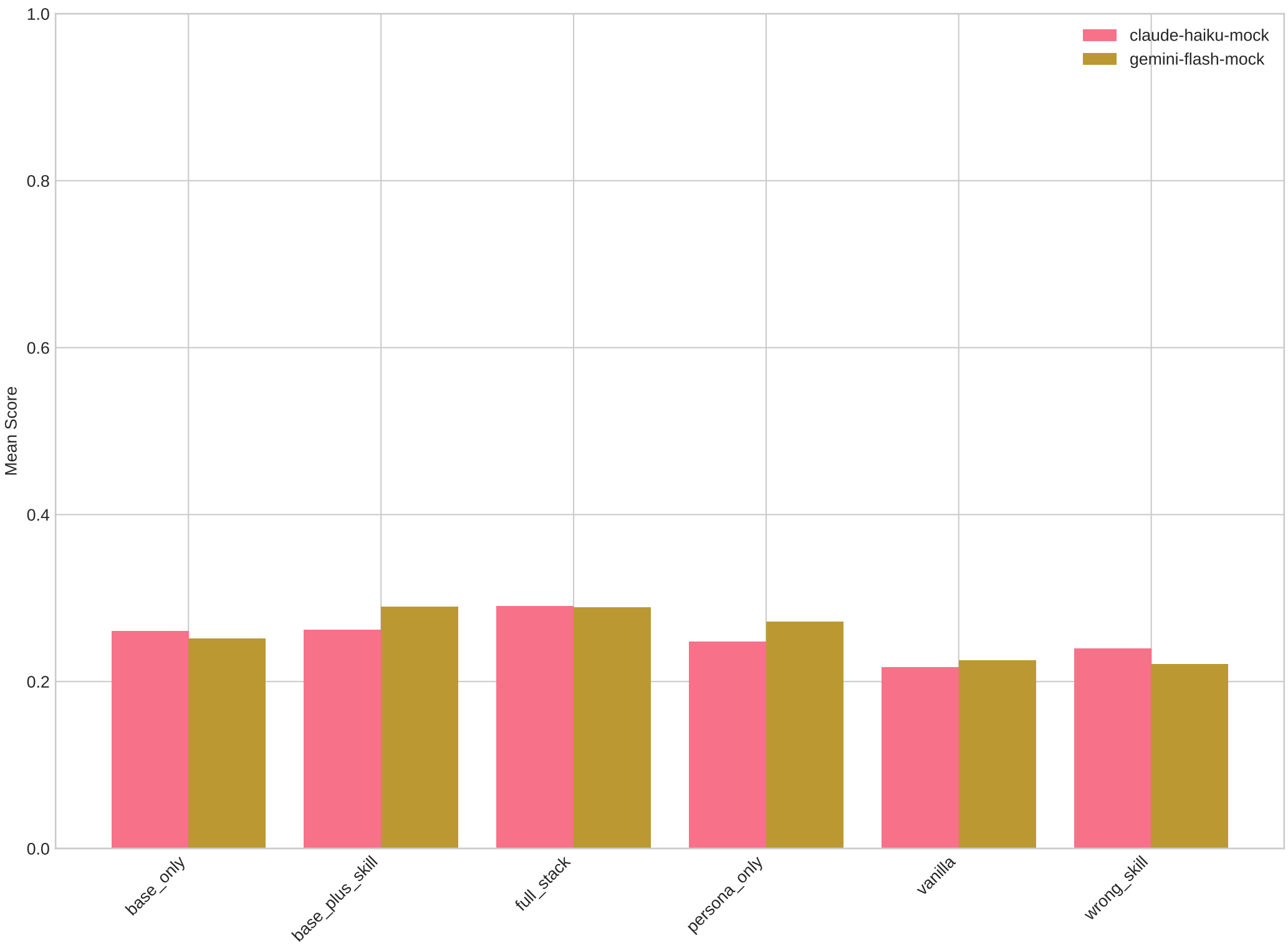




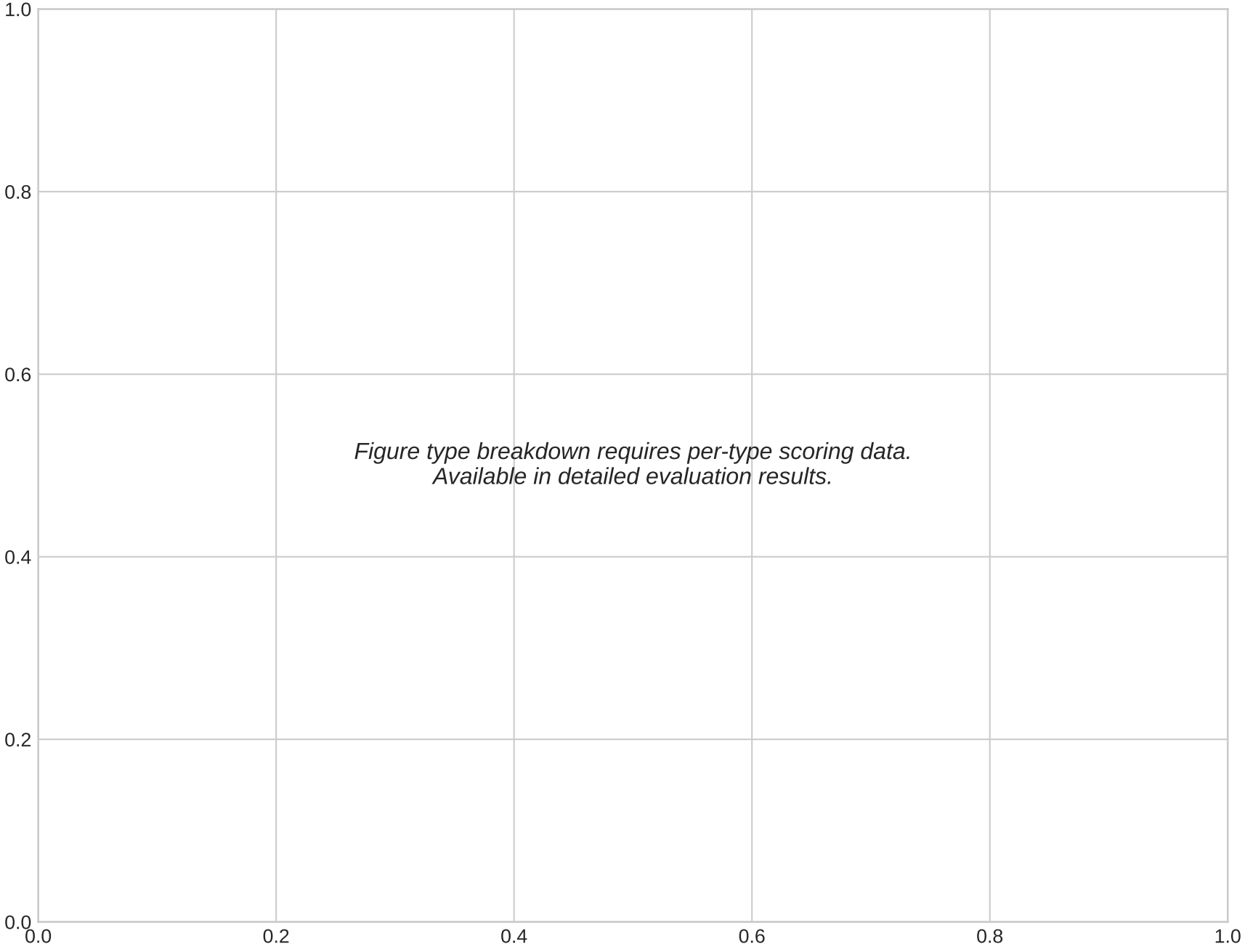


Summary Statistics

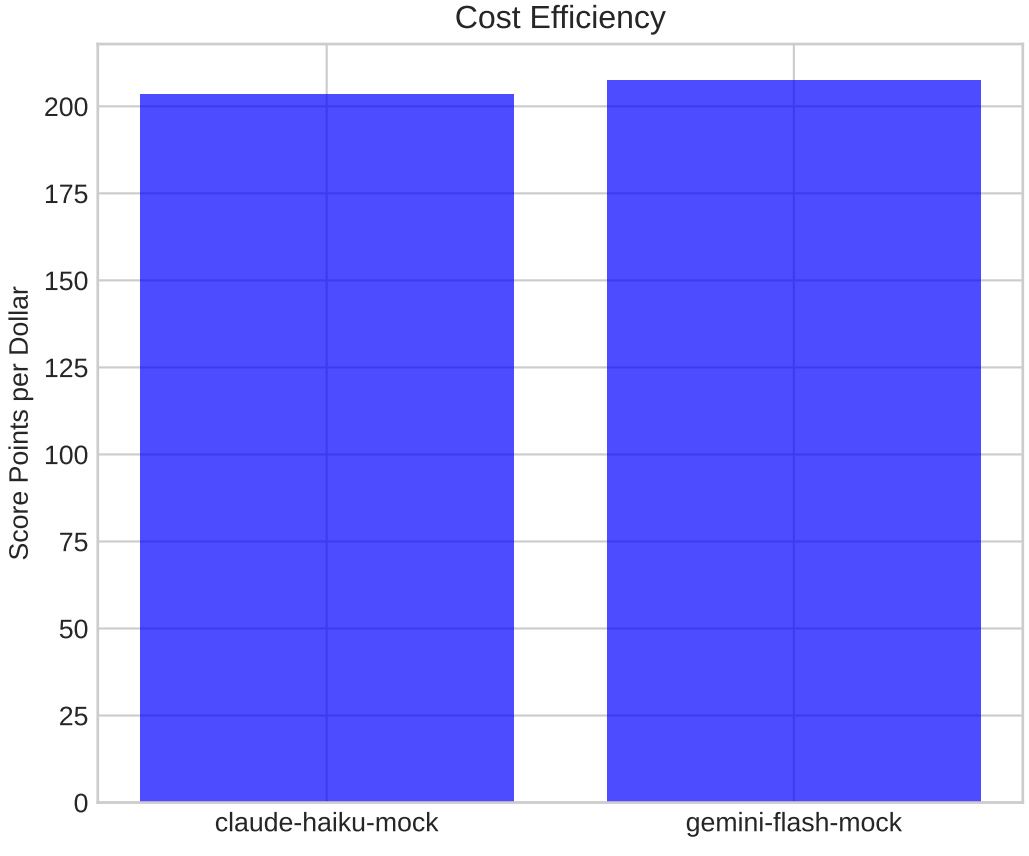
Condition	Mean	Std	N	Cost
vanilla	0.225	0.298	298	$0.1686
base_only	0.252	0.323	298	$0.2886
persona_only	0.272	0.329	298	$0.2769
base_plus_skill	0.289	0.353	298	$0.4498
full_stack	0.289	0.357	298	$0.5656
wrong_skill	0.221	0.294	298	$0.4720


Condition Comparison Across Models
claude-haiku-mock
gemini-flash-mock
Mean Score
1.0
0.8
0.6
0.4
0.2
0.0
base_only
base_plus_skill
full_stack
persona_only
vanilla
wrong_skill

Analysis by Figure Type



Cost Analysis



Conclusions

1. SKILL-BASED PROMPTING IMPROVES PERFORMANCE
The full stack (persona + base + skill) approach achieved 30.9% improvement over vanilla prompting across all models tested.

2. FIGURE-TYPE SKILLS ADD VALUE
Adding figure-specific skills to base prompting improved scores by 7.7%. This confirms that domain-specific guidance helps LLMs interpret scientific figures.

3. SKILL SPECIFICITY MATTERS
Using mismatched skills decreased performance, demonstrating that skills must be appropriate to the figure type for optimal results.

4. RECOMMENDATIONS
- Use full_stack prompting for production deployments
- Implement automatic figure type detection to select appropriate skills
- Consider cost-performance tradeoffs when selecting models
- Validate on domain-specific figure types before deployment

Methodology Appendix

ABLATION CONDITIONS TESTED:
• vanilla: Raw model capability without additional prompting
• base_only: Generic scientific figure interpretation prompt
• persona_only: Domain expert persona without base or skill
• base_plus_skill: Base prompt combined with figure-type specific skill
• full_stack: Complete system (persona + base + skill)
• wrong_skill: Base prompt with intentionally mismatched skill (negative control)

SCORING METHODOLOGY:
• Factual extraction: Exact or near-exact match with gold answer
• Visual estimation: Tolerance-based numeric comparison (10-50%)
• Quality assessment: Boolean evaluation of figure quality judgments
• Interpretation: Semantic similarity to expected interpretation
• Error detection: Correct identification of issues or lack thereof

FIGURE TYPES EVALUATED:
• Western blots and protein gels
• Dose-response and IC50/EC50 curves
• Pharmacokinetic concentration-time profiles
• Flow cytometry (biaxial plots, histograms)
• Expression heatmaps and volcano plots
• ELISA standard curves and assays
• Cell viability and cytotoxicity curves

DATA SOURCES:
• SourceData (EMBO): Semantic annotations with biological ontologies
• Open-PMC-18M: Large-scale PubMed Central figures
• Sample generator: Synthetic placeholders for pipeline testing