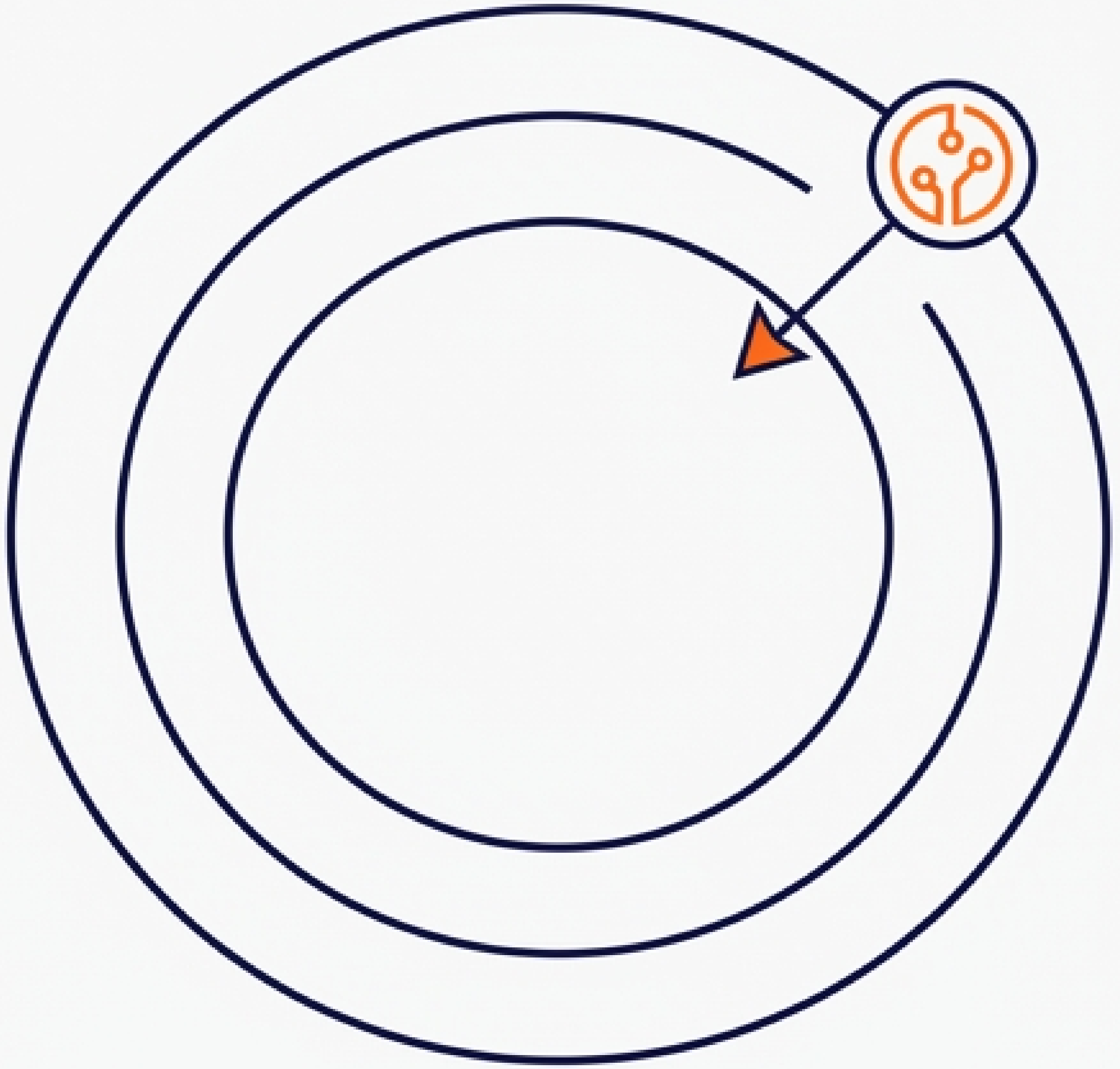
Swiss Technical Editorial



Coupling Planning with Tool-Grounded Checks

A Framework for Reliable Agent Systems: From 3.3% to 99.3% Success Rates

Research Independent | Technical Briefing

Executive Summary

Principled Integration of Tool Feedback Solves Agent Reliability.

96.0 pt

Performance Lift (Percentage Points)

Integrating tool outputs (unit tests, compilers) into the planning loop boosts success rates from a baseline of 3.3% to 99.3%.

The Mechanism

Success depends on tuning two critical variables:

- **Scoring Function:** How we judge the progress of a plan.
- **Termination Criterion:** The logic that decides when to stop.

ANOVA Significance: $F = 4892.9$, $p < 10^{-6}$

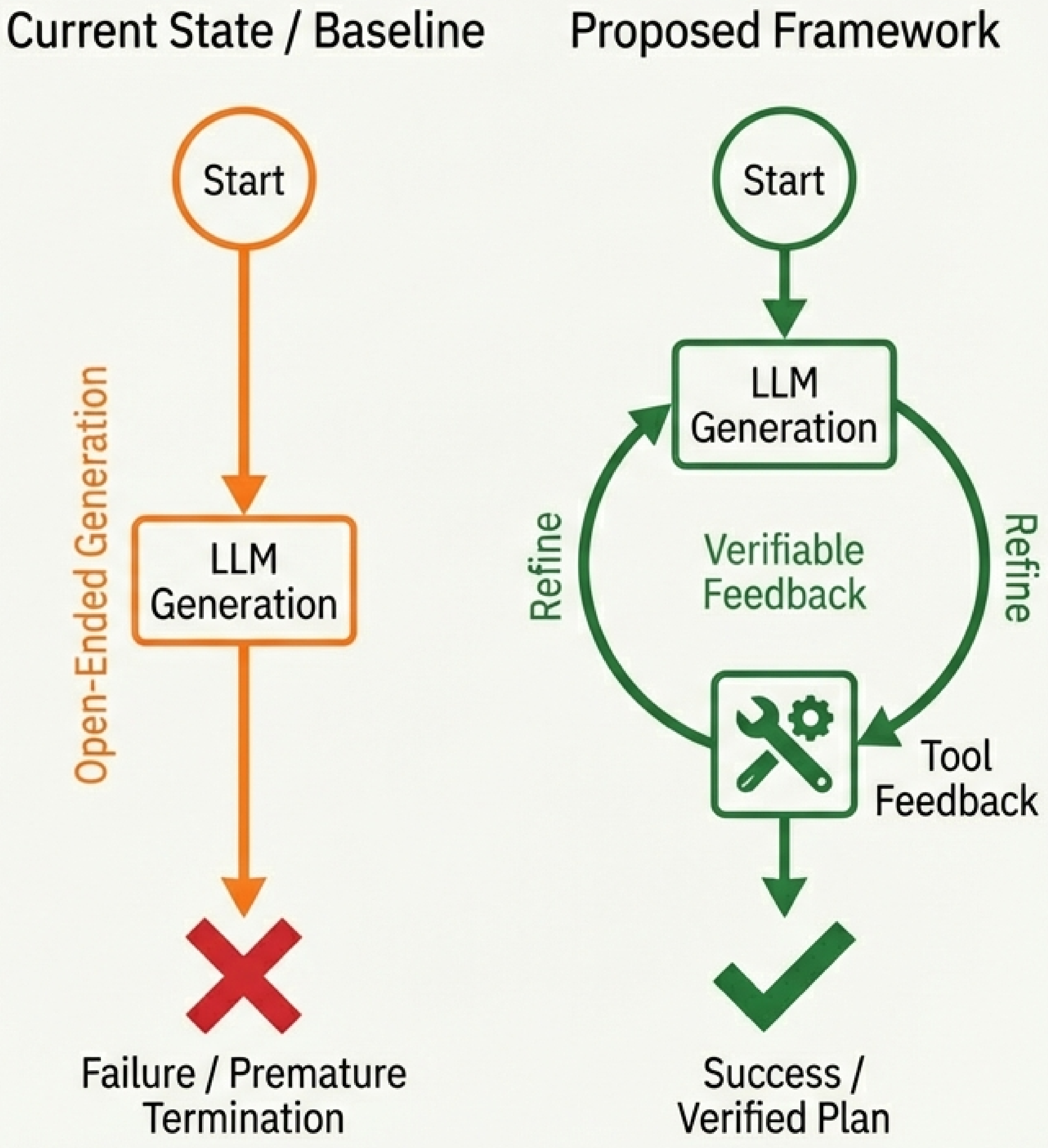
The Trade-off

A distinct Pareto frontier exists between cost and quality:

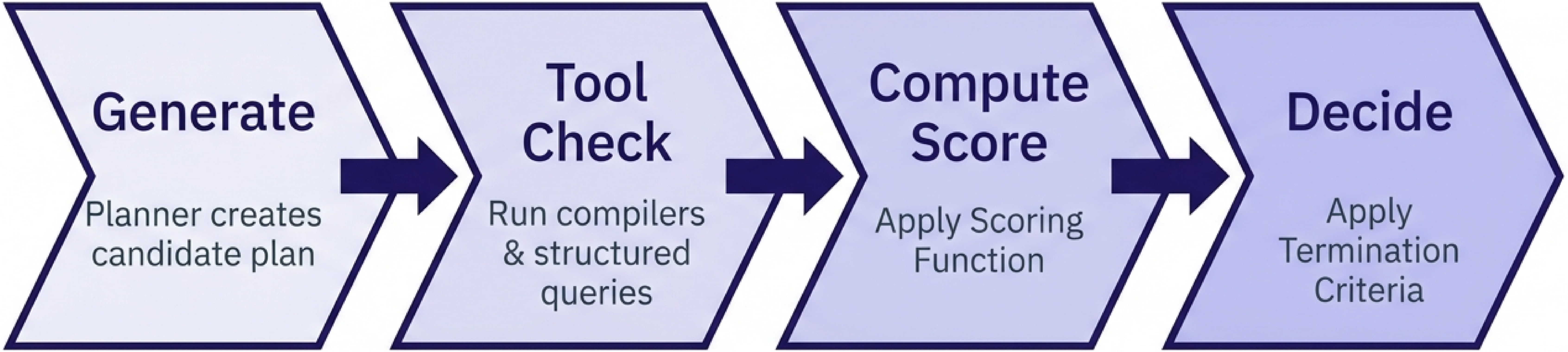
- **Max Quality:** Bayesian Scoring + Patience Termination (99.3% Success).
- **Max Efficiency:** Confidence-based Termination (0.00293 success/compute).

The Reliability Gap in Search-Based Planning

Large Language Models (LLMs) excel at generating candidates but struggle with self-correction without external grounding. Without tool feedback, agent planning success is negligible (3.3%). Agents tend to hallucinate correctness or terminate prematurely.



System Methodology: The Tool-Coupled Planning Loop



Generate

Planner creates candidate plan

Tool Check

Run compilers & structured queries

Compute Score

Apply Scoring Function

Decide

Apply Termination Criteria

Experimental Scope: 100 tasks per trial | 30 trials total | 4 tool types

Variable 1: Scoring Functions (The Judge)

How the agent evaluates quality based on feedback.

Weighted

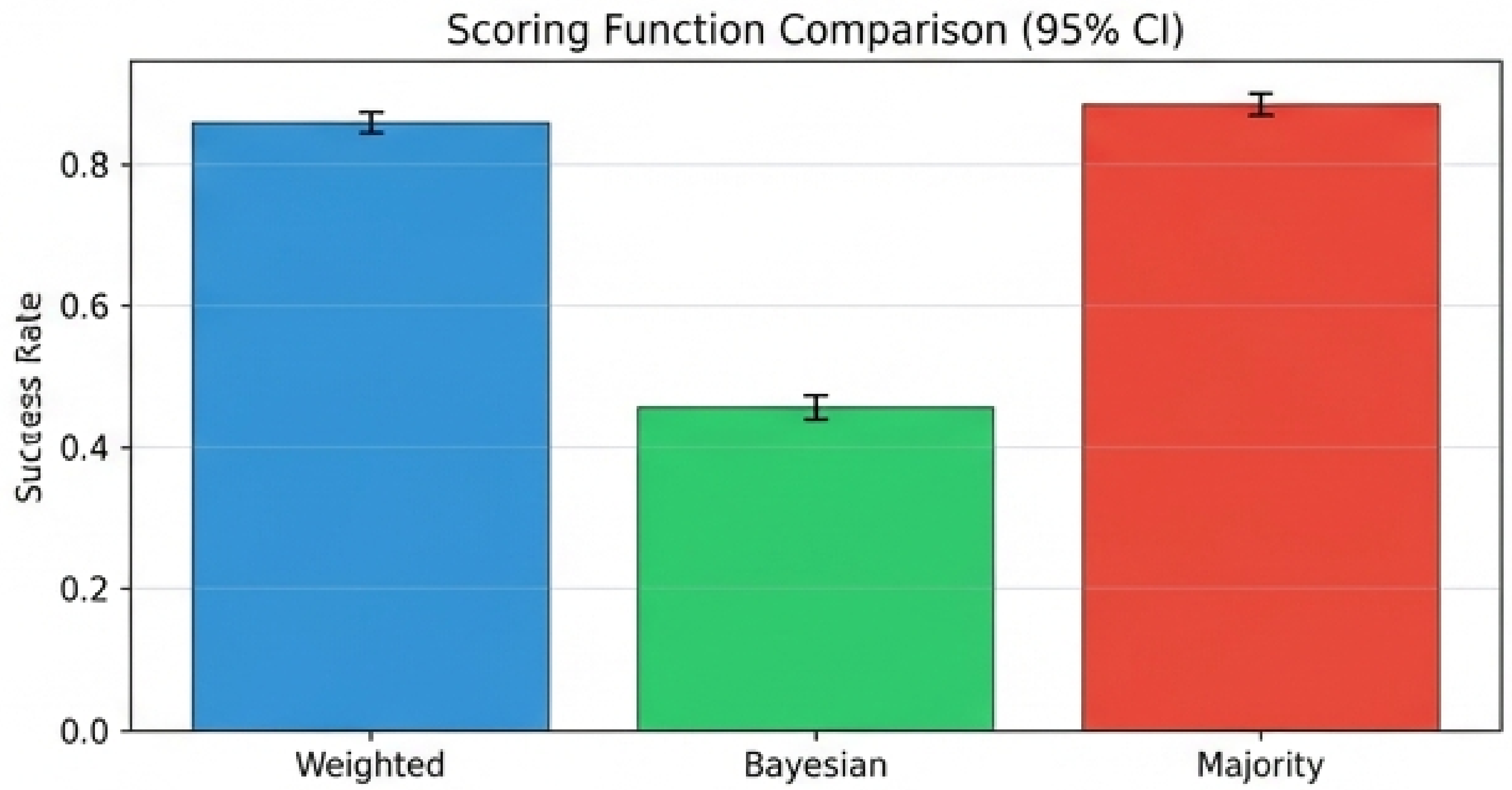
Linear combination. Weight w=0.4 for tool feedback. Optimized for speed.

Bayesian

Sequential posterior update using tool confidences as likelihoods. Lower variance.

Majority

Voting consensus. Average of plan score and tool vote fraction. High raw rate.



Variable 2: Termination Criteria (The Stop Mechanism)

Patience (The Explorer)

Rule: Stop after 5 iterations without > 0.01 improvement.

Outcome: Maximizes raw success (0.993).

Confidence (The Sprinter)

Rule: Stop when the combined score exceeds 0.85.

Outcome: Best compute efficiency (0.00293 success/compute).

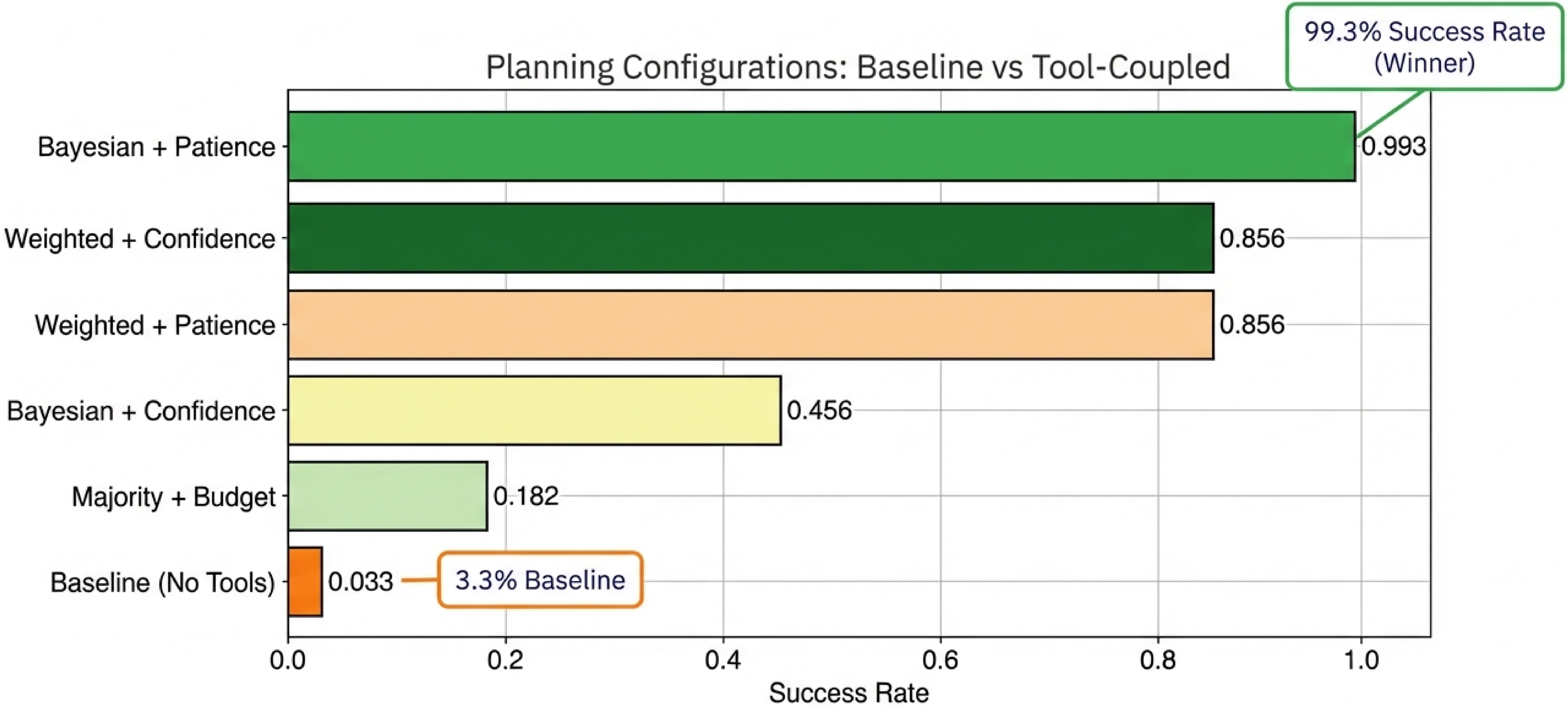
Budget (The Accountant)

Rule: Stop when compute cost exceeds a pre-set limit.

Outcome: Lowest success (0.202). Cuts off reasoning prematurely.

Configuration Results: Achieving a 96-Point Lift

Comparison of success rates across all configurations vs. baseline.



99.3% Success Rate (Winner)

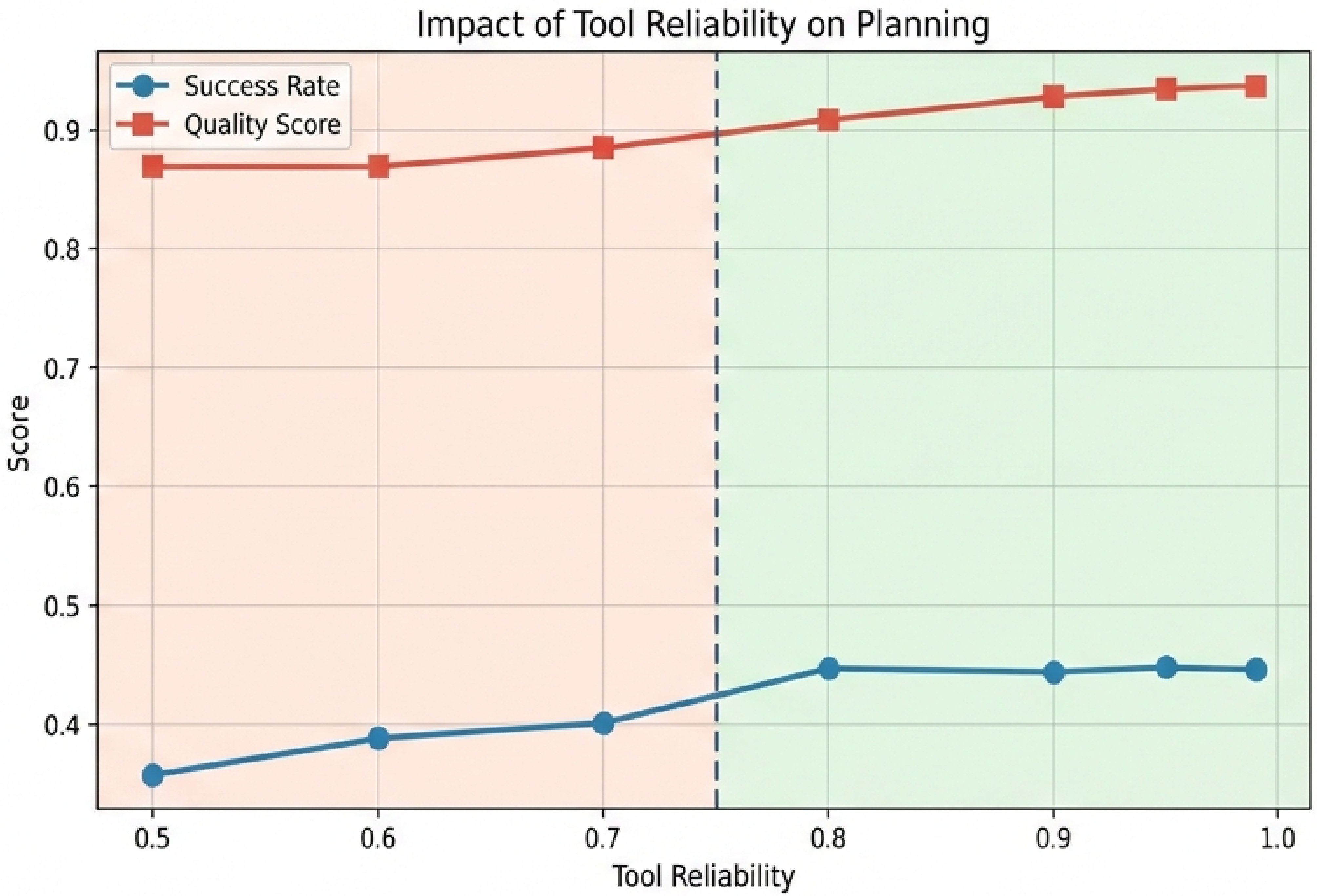
3.3% Baseline

The Reliability Threshold: When Do Tools Help?

The Danger Zone: Below 70% reliability, tool integration is risky. Noisy feedback confuses the planner, degrading quality.

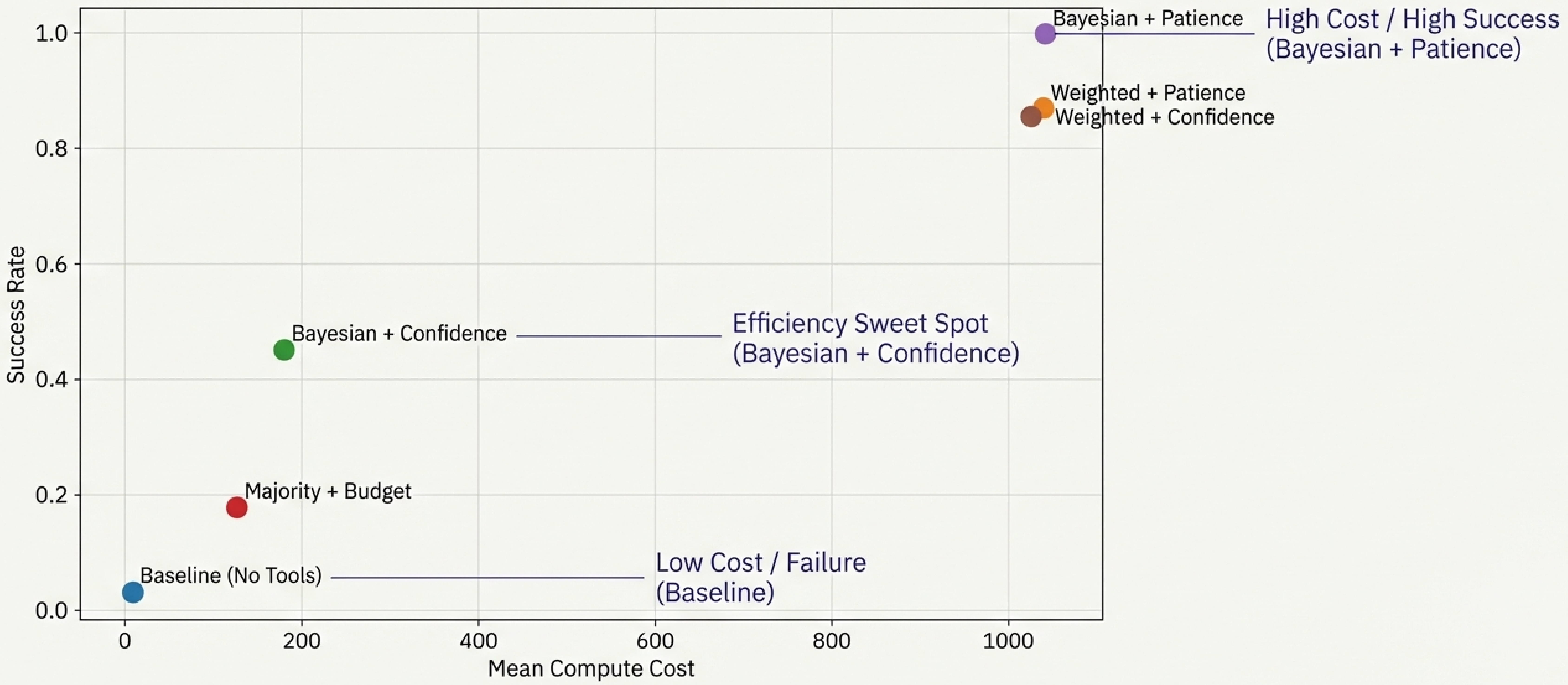
The Safe Zone: Above 70%, success rates climb steadily.

Takeaway: Validate tools (unit tests, verifiers) to ensure >70% accuracy before integration.



The Compute-Quality Pareto Frontier

Strategic choice between efficiency and raw power.



High Cost / High Success (Bayesian + Patience)

Efficiency Sweet Spot (Bayesian + Confidence)

Low Cost / Failure (Baseline)

Swiss Technical Editorial

Efficiency Metrics Breakdown

Termination Strategy	Success Rate	Compute Cost	Efficiency Score
Patience	0.993	1094	0.000908
Confidence	0.454	155	0.002930
Budget	0.202	113	0.001793

> *"Confidence-based termination offers the best compute efficiency... while patience-based termination maximizes raw success.”*

Swiss Technical Editorial

Strategic Configuration Guide

Scenario A: Mission Critical

Offline / Code Gen / Medical

Recommendation:

Bayesian Scoring + Patience Termination

Result:

>99% Success

Scenario B: Real-Time

Chatbots / Live Assistance

Recommendation:

Weighted Scoring + Confidence Termination

Result:

~85% Success at 1/7th cost

Scenario C: Low Reliability

Tool Accuracy < 70%

Recommendation:

Do Not Integrate Tools

Result:

Focus on tool grounding first

Swiss Technical Editorial

Conclusion: The Era of Verifiable Agents

- Agent planning without checks is unreliable (**3.3%**).
- Coupling planning with tool-grounded checks solves this (**99.3%**).
- Iterative refinement (**Patience**) beats raw speed.

Reliable agents will not be built on larger models alone, but on better verification loops.

Data based on ‘Coupling Planning with Tool-Grounded Checks’ (2026).