Prompt Tornado

AI Workflow Evaluation Report

Prompt Tornado — multi-model workflow planner & router · Internal evaluation, May 2026

Prompt Tornado turns a single prompt into a fully executed multi-step AI workflow across leading model providers. Every release is evaluated on two axes: **planning accuracy** (does it decompose and route the workflow correctly?) and **output quality** (is each step's result actually good?), scored 1–10 by an independent LLM judge against task-specific rubrics.

181	8.43 / 10	88.2%	172 / 181
Task types evaluated	Avg quality score (judged)	Quality pass rate	Routing checks passed

Methodology

Planning evaluation: the workflow planner was evaluated across 200 compound prompts representing real-world AI workflows — achieving **98% planning accuracy** (valid schema, correct step ordering, no hallucinated tasks) with 100% deterministic output (identical prompts produce identical plans).

Quality evaluation: 181 registry task types across 17 categories were executed and scored 1–10 by an independent LLM judge against per-task rubrics. Media-generation categories (image, audio, video) are validated structurally rather than judge-scored. Deploys are automatically blocked on quality regressions.

Results by category

Category	Cases	Avg quality	Routing
Text Generation	31	8.97	28/31
Specialized Domains	18	8.83	17/18
Code Generation	17	8.41	16/17
Data & Analysis	14	8.31	13/14
Agentic / Automation	11	6.82	11/11
Research	10	7.70	10/10
Summarization	9	9.11	9/9
Image Generation	9	—	9/9
Reasoning & Planning	9	9.00	8/9
Question Answering	8	8.00	8/8
Content Editing	8	8.62	8/8

Audio Generation	7	—	7/7
Vision / Multimodal	7	8.00	7/7
Translation	6	6.83	6/6
Video Generation	6	—	6/6
Structured Output	6	9.33	5/6
Personalization	5	8.25	4/5
All task types	**181**	**8.43**	**172/181**

Evaluation harness, rubrics, and CI quality gate are part of the Prompt Tornado platform. Figures reflect the quality_v1 baseline (2026-05-11).