Macrodeep Research

Grading the Ungradable — Rubric Methodology for GTM-Bench's Generative Categories

Evaluation · Jul 2026 · 13 min read

Macrodeep Research Team

Abstract

Fifth in the GTM-Bench research series. The objective categories (ground truth, ORCHESTRATE) have deterministic oracles. Two categories don't: COMPOSE and SUMMARIZE. This piece is about how they're graded, why they're deliberately kept to a minority of the score, and how the benchmark stays trustworthy despite containing subjective tasks at all. No model scores — this is methodology, and specifically the methodology for the part that's hardest to defend.

Contributions

- COMPOSE and SUMMARIZE have no deterministic oracle — the design question is containment, not perfection
- Minority weight plus a fully objective-only score keep the subjective categories from being load-bearing
- Pairwise comparison against a held-out reference, decomposed into rubric axes, beats absolute open scoring
- The judge model is held separate from any model under evaluation to remove self-preference bias

Starting from the weakness

Every prior article in this series has argued for GTM-Bench's rigor from a position of strength: deterministic grading, hidden oracles, reproducible harness. This one starts from a weakness, on purpose, because the honest way to handle a soft spot is to name it and show your work.

Two of the seven categories have no deterministic oracle. “Write a good cold email” (COMPOSE) and “summarize this call” (SUMMARIZE) are real GTM tasks — arguably the most visible ones — but there is no test suite for a good email and no exact-match answer for a good summary. Two experienced reps will disagree about the same draft, and both can be right.

This is the exact place where benchmarks go soft. Lean on an LLM judge for a big share of the score and you inherit the judge's biases, and worse, you hand models an easy exploit: write to the judge's preferences rather than to the real objective. A benchmark that lets fluent, judge-pleasing output inflate the headline number is measuring sycophancy, not capability.

So the design question for COMPOSE and SUMMARIZE is not “how do we grade these perfectly” — you can't. It is “how do we extract some real signal from them without letting them corrupt the parts of the benchmark that are trustworthy.” The answer has three parts: contain their influence, structure the judgment, and isolate the judge.

Containment: minority weight and the objective-only escape hatch

The first and most important defense is architectural, not clever. The generative categories are held to a minority of the total score, and the benchmark reports a separate objective-only number that excludes them entirely.

This matters more than any refinement to the rubric. It means a reader who distrusts LLM-judged grading — a completely reasonable stance — can ignore COMPOSE and SUMMARIZE altogether and still have a meaningful, fully deterministic comparison from the objective-only score. The subjective categories are prevented, by construction, from being load-bearing. They add color; they cannot carry the ranking.

The generative categories are allowed to say something about generative quality, and are structurally forbidden from propping up the headline. If they were removed tomorrow, the benchmark's core claim would be unchanged.

Structure: rubric-based pairwise judging, not open scoring

Within that contained budget, the goal is to make the judgment as structured and repeatable as a subjective judgment can be. Two design choices do most of the work.

Pairwise, not absolute. Asking a judge “score this email from 1 to 10” produces noisy, drifting numbers — absolute quality scores are notoriously unstable across runs and prompts. GTM-Bench instead uses pairwise comparison against a held-out reference: is the model's output better than, worse than, or comparable to a reference answer for this instance? Relative judgments are far more stable than absolute ones, and they anchor the judgment to a concrete comparison rather than an abstract scale.

Rubric-bound, not open-ended. The judge is not asked for a vibe. It is given explicit axes to evaluate against, and only those. For COMPOSE, the axes are things like relevance to the stated context, specificity (does it use the provided account details or generically fill), and call-to-action clarity. For SUMMARIZE, the axes are faithfulness (nothing asserted that the source doesn't support) and coverage (nothing decision-critical omitted). The judgment is decomposed into these axes rather than rolled into one holistic score, which both improves consistency and makes a verdict inspectable — you can see why one output beat another, not just that it did.

Faithfulness carries extra weight in SUMMARIZE for a specific reason: a summary that invents a commitment or misstates a next step is not a stylistic miss, it is a factual error with downstream consequences. The rubric treats hallucinated content as a hard fault, not a minor deduction, because that is how it behaves in real GTM use.

Isolation: the judge is not a contestant

A subtle failure mode in judge-based grading is self-preference: a model tends to rate its own style highly. If the judge model is also a model under test, or shares a lineage with one, the grading tilts.

GTM-Bench's rule is that the judge model is held separate from any model under evaluation. A model being scored does not judge its own outputs, and the judge is fixed and disclosed for a given evaluation run so the grading surface is constant across all contestants. Everyone is graded by the same external judge, and no contestant grades itself. This doesn't eliminate judge bias — nothing does — but it removes the most direct conflict of interest and makes whatever bias remains uniform across the field rather than tilted toward one entrant.

Where feasible, the judgment is also taken across more than one judge configuration and checked for agreement, so an instance where the judgment is fragile (judges disagree) is visible as low-confidence rather than reported with false precision.

What the rubric grading honestly is, and isn't

Stated without varnish, because this is the category where overclaiming would be easiest and most damaging: rubric-based pairwise judging with an isolated judge produces a directional, contained signal about generative quality. It can tell you that one model's outbound copy is, under a stated rubric and an external judge, generally preferred to another's against held-out references. That is real and useful information.

It is not a deterministic measurement, and the benchmark never presents it as one. It carries residual judge bias, it depends on rubric design choices that reasonable people could contest, and it is the softest number GTM-Bench produces. This is exactly why it is minority-weighted and quarantined from the objective-only score. The honest posture is: here is a structured, isolated, contained attempt to say something about the ungradable parts of GTM work — read it as directional, and if you don't trust it, the objective-only number is right there and owes nothing to it.

Why include them at all

A fair question: if these categories are the weak point, why not drop them and ship a purely objective benchmark? Because outbound copy and call summarization are not fringe tasks in go-to-market work — for many teams they are the most visible thing a GTM model does. A benchmark that measured only

enrichment, qualification, and orchestration while ignoring the generative surface entirely would be cleaner but less complete; it would be silent on capabilities buyers actively care about.

The design bet is that it is better to include these tasks with honest, contained, clearly-labeled subjective grading than to pretend they don't exist. The alternative to a directional signal here is not a better signal — it is no signal, plus a blind spot.

So GTM-Bench includes them, grades them as rigorously as subjective tasks allow, weights them so they cannot distort the trustworthy core, and labels exactly what the resulting number is worth. That combination — include, contain, disclose — is the whole methodology.

The series, and where the soft edges go

This completes the taxonomy coverage: the objective categories rest on deterministic oracles and constructed ground truth, and the generative categories rest on contained, isolated, rubric-bound judgment that is deliberately prevented from being load-bearing. The benchmark's trustworthy core is the objective-only score; the generative categories add breadth without being allowed to compromise that core.

The leaderboard remains a separate release, run under the documented harness, published only when the numbers are real and reproducible — with the objective-only number as the figure to trust, and the generative categories reported alongside it for exactly what they are.

References

1. Zheng et al. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. 2023.

2. Dubois et al. AlpacaFarm: A Simulation Framework for Methods that Learn from Human Feedback. 2023.

3. Wang et al. Large Language Models are not Fair Evaluators. 2023.