Macrodeep Research

What “Correct” Means — Constructing Ground Truth for GTM-Bench's Objective Categories

Evaluation · Jul 2026 · 15 min read

Macrodeep Research Team

Abstract

Fourth in a series on GTM-Bench. The design note gave the taxonomy, the ORCHESTRATE deep-dive covered agentic grading, and the contamination note covered keeping scores honest. This piece goes under all of them, to the question they each depend on: how is the ground truth built, and what makes it trustworthy enough to grade against? No model scores — this is about the oracle itself.

Contributions

- Each objective category — ENRICH, QUALIFY, TAM, ROUTE — derives ground truth differently, with a different failure mode
- Correctness is sourced from records, outcomes, criteria, and consensus — never from the author's say-so
- Where sources conflict or practitioners genuinely disagree, the instance is dropped, not adjudicated by fiat
- For every category, the definition of “correct” is published as part of the spec

The load-bearing assumption

Every claim GTM-Bench makes rests on one thing being true: that the oracle is right. A deterministic harness, a clean held-out split, an un-gameable assertion language — all of it is worthless if the answer the harness grades against is wrong. Garbage ground truth produces a rigorous-looking measurement of nothing.

This is the least discussed part of most benchmarks and the most important. It is easy to describe a grading harness; it is hard to defend the answers the harness trusts. The four objective categories — ENRICH, QUALIFY, TAM, ROUTE — each derive their ground truth differently, and each has a different failure mode. This piece takes them one at a time, because “correct” does not mean the same thing in all four.

ENRICH: correctness by canonical record

For entity resolution and enrichment, “correct” means matching a canonical record — the true legal name, headquarters country, employee band, industry code for a given entity. The ground truth is a

resolved, verified company or person record, and grading is field-by-field exact match.

The construction challenge is that canonical records are not free-floating truths; they are assembled, and assembly introduces error. Multi-source corroboration — a field enters ground truth only when independent sources agree; a single-source value is held out or flagged, not promoted to oracle. Recency bounds — each ground-truth record carries an as-of date, and instances are constructed so the answer is correct as of a stated point, not “correct forever.” Normalization before comparison — field-exact-match is only fair if both sides are normalized the same way (“Ltd” vs “Limited,” country names vs ISO codes), so the normalization rules are part of the spec, not left to the grader's discretion.

The residual risk, stated plainly: canonical records are as good as their sources, and no source is perfect. ENRICH ground truth is therefore treated as high-confidence but not infallible, and instances where sources genuinely conflict are excluded rather than adjudicated by fiat.

QUALIFY: correctness by realized outcome

For ICP fit and tiering, ground truth comes from something a synthetic benchmark can't fake: what actually happened. An account labeled “good fit, Tier A” is labeled that because accounts like it closed; a “poor fit” label rests on accounts like it that didn't. Ground truth is derived from real closed-won and closed-lost outcomes, and the task grades as classification.

Outcome-derived labels are powerful because they are grounded in reality rather than opinion — but they carry a specific, honest hazard: outcomes are noisy. A deal closes for reasons that have nothing to do with fit; a good-fit account is lost for reasons orthogonal to it. If each label is a single deal, the benchmark measures luck as much as capability.

The response: aggregate, don't anecdote — labels are derived from cohorts of similar accounts and their outcome rates, not individual deals, which averages out the per-deal noise. Confidence-weighted inclusion — segments with thin or contradictory evidence are excluded rather than assigned a coerced label. Explicit label definition — what “Tier A” means (the outcome-rate thresholds behind each tier) is defined in the spec, so the classification target is a stated rule, not a vibe.

QUALIFY ground truth encodes real signal and real noise, and the benchmark claims only the former as measurable. It does not pretend outcome labels are a perfect proxy for fit.

TAM: correctness by curated target set

For lookalike and total-addressable-market construction, the model expands seed accounts into a set, and “correct” means matching a curated target set — graded as precision and recall at k. The hard part is that a target set is a judgment about a boundary, and boundaries are arguable. Where does a market

end? If the target set is one person's opinion of the boundary, the benchmark measures agreement with that person, not capability.

The defenses push toward defensible boundaries. Criteria-defined membership — a target set is built from explicit inclusion criteria (sector, size, geography, signal thresholds) applied consistently, not hand-picked, so the boundary is reproducible and inspectable. Precision/recall rather than exact-set — grading at k tolerates the inherent fuzziness of market edges, rewarding a model for getting the core right without catastrophic penalty for a defensible edge case. Seed–target independence — seeds and targets are constructed so the expansion is non-trivial and cannot be reached by surface pattern-matching on the seeds alone.

Residual risk: market boundaries are genuinely contestable, and TAM ground truth reflects one criteria-defined boundary among possible ones. The mitigation is that the criteria are published, so a disagreement is a disagreement with a stated rule, not with a hidden opinion.

ROUTE: correctness by decision structure

For next-action and sequencing logic, the model sees a scenario and picks the next action; “correct” is a defined branch of a decision structure. The risk here is the inverse of TAM's: it is tempting to make routing ground truth too crisp, encoding one team's playbook as universal truth. A defensible next action in one motion is wrong in another.

The response: consensus decision structures — the branches encode widely-agreed routing logic, the actions most experienced practitioners would converge on for a given state, not one org's idiosyncratic sequence. State-complete scenarios — each scenario carries enough state that the correct action is actually determined by it; under-specified scenarios are rejected in construction. Single-correct where defensible, excluded where not — ROUTE only includes instances with a defensible single best action; genuinely multi-valid situations are not in the objective set.

The thread through all four

The four categories construct ground truth differently — canonical record, realized outcome, curated target set, decision structure — but the same discipline runs through each. Derive, don't decree — ground truth comes from something external (a registry, an outcome, a criteria-defined set, a consensus structure), not from the author's say-so; this is what keeps a lab-authored benchmark from being circular. Exclude the unresolvable — where sources conflict, evidence is thin, or competent practitioners genuinely disagree, the instance is dropped, not adjudicated by fiat; a smaller trustworthy set beats a larger contestable one. Publish the rule — for every category, the definition of “correct” (normalization scheme, tier thresholds, inclusion criteria, decision logic) is part of the spec, so a reader who disagrees is disagreeing with a stated rule they can see.

Why the oracle is the whole game

It is easy to be impressed by grading harnesses, sandboxes, and held-out splits. But all of that machinery grades against the oracle, and if the oracle is wrong, the machinery just measures wrongness precisely. The credibility of GTM-Bench lives or dies on the answers being right and their construction being inspectable.

A benchmark is a claim about what "correct" means in a domain. GTM-Bench's claim is that in go-to-market work, correctness can be sourced from records, outcomes, criteria, and consensus rather than from opinion — and that where it can't, the honest move is to not ask the question.

When the leaderboard comes, it will grade against these oracles, constructed this way, with the rules published. The number will mean exactly as much as the ground truth beneath it — which is why the ground truth came first.

The series, so far

Four pieces now define GTM-Bench without reporting a single score: the design note (taxonomy and scoring), the ORCHESTRATE deep-dive (agentic grading), the contamination note (keeping scores honest over time), and this one (constructing the ground truth). That was deliberate. The methodology is public before the leaderboard, so that when frontier models are run under this harness, every question a skeptic could ask — how is it graded, how is gaming prevented, why is the answer correct — already has a documented answer. The leaderboard is a separate release, and it goes out only when the numbers are real and reproducible by someone outside.

References

1. Jimenez et al. SWE-bench: Can Language Models Resolve Real-World GitHub Issues? 2024.

2. Rein et al. GPQA: A Graduate-Level Google-Proof Q&A Benchmark. 2023.

3. Northcutt et al. Pervasive Label Errors in Test Sets Destabilize ML Benchmarks. 2021.