Complex-IF and Beyond: Expert Rubrics for RLVR

Sushant Mehta*, **Liudas Panavas, Eleanor Fleming, Paul Mains, Edwin Chen**
Surge AI

Abstract

As LLM capabilities advance rapidly, the evaluation methods used to assess them increasingly lag behind. Traditional benchmarks rely on programmatic verification of narrow, surface-level constraints, yet real-world instruction following and agentic tasks demand assessment of nuanced, context-dependent behaviors that resist simple scripted checks. We present a systematic analysis of *expert-curated rubric-based evaluation* as an alternative paradigm, drawing on empirical evidence from two domains: complex instruction following and enterprise agentic tasks. We first articulate five design principles for constructing high-quality rubrics, including *Maximum Viable Atomicity*, *intent-aware criterion design*, and *iterative LLM-judge calibration*. To validate these principles, we introduce COMPLEX-IF, a new expert-curated instruction-following dataset in which each prompt is paired with 10–40 atomic rubric criteria. We demonstrate that these expert rubrics are not only better evaluation instruments but also highly effective training signals: training on approximately 1,000 COMPLEX-IF examples yields +15.5 pp improvement for a 4B-parameter model and +12.2 pp for a 235B-parameter model on instruction following, while single-epoch RL training on a rubric-graded enterprise environment produces gains that transfer to out-of-distribution benchmarks the model was never trained on (+4.5 pp BFCL, +7.4 pp τ^2-Bench, +6.8 pp Toolathlon). Our findings establish that expert-authored rubrics improve both the measurement and the development of frontier LLM capabilities, serving as effective evaluation and RL training signals.

1 Introduction

Traditional benchmarks are increasingly saturated, with frontier models topping out on tests where meaningful capability differences can no longer be distinguished (Eriksson et al., 2025). Data contamination can further erode confidence in reported scores. More fundamentally, there is a growing misalignment between what benchmarks measure and what real-world deployment requires: namely, the ability to follow complex, layered instructions; to maintain behavioral constraints across extended interactions; and to execute multi-step professional workflows reliably.

IFEval (Zhou et al., 2023b), one of the most widely cited instruction-following benchmarks, illustrates this gap vividly. Its 25 verifiable constraint types (word counts, forbidden characters, formatting rules) can be checked programmatically, but this programmatic verifiability comes at the cost of validity. A model can produce incoherent nonsense and still pass, so long as it avoids commas and the letter "c." The benchmark shaped itself around the evaluation method rather than around the construct it claims to measure.

This paper examines an alternative paradigm: *expert-curated rubric-based evaluation*, in which domain experts decompose task success into atomic, verifiable criteria that capture both explicit requirements and pragmatic user intent. We analyze this approach across two complementary domains: **complex instruction following**, using COMPLEX-IF, a new expert-curated dataset we introduce in this work, alongside the independently developed AdvancedIF benchmark (He et al., 2025); and **agentic task execution**, using the CoreCraft enterprise simulation (Mehta et al., 2026; Ritchie et al., 2026).

Our paper makes three contributions. First, we articulate five design principles for expert rubric construction that improve construct validity relative to programmatic verification (§3). Second, we introduce COMPLEX-IF, an expert-curated

*Correspondence to: sushantmehta@surgehq.ai

Proceedings of the Fifth Workshop on Generation, Evaluation and Metrics (GEM), pages 668–677

instruction-following dataset embodying these principles, with approximately 1,000 prompts each paired with 10–40 atomic rubric criteria. Third, we present empirical evidence demonstrating that expert rubrics designed according to these principles are highly effective as reward signals for reinforcement learning (§5): in both instruction following and agentic task execution, rubric-based RL training on modest amounts of data produces substantial, transferable improvements, suggesting that expert rubrics serve not only as more valid evaluation instruments but also as highly effective training signals for reinforcement learning.

2 Background and Related Work

From Programmatic to Rubric-Based Evaluation. Instruction-following evaluation has progressed through several stages. IFEval (Zhou et al., 2023b) established programmatic verification with 25 constraint types. FollowBench (Jiang et al., 2024) revealed that model performance degrades as constraint complexity increases. InFoBench (Qin et al., 2024) proposed decomposing complex instructions into binary verification questions. ComplexBench (Wen et al., 2024) introduced a hierarchical taxonomy of constraint composition. AdvancedIF (He et al., 2025) moved to expert-written rubrics for 1,600+ prompts, finding that even frontier models achieve only approximately 75% when assessed against these richer criteria. In a complementary medical setting, HealthBench (Arora et al., 2025) pairs 5,000 physician–patient conversations with 48,562 physician-written rubric criteria across multiple behavioral axes (accuracy, communication, context awareness). Our COMPLEX-IF dataset, introduced in this paper, shares the expert-authored rubric design of these recent efforts but is built from the start to serve a second purpose: the same rubrics that evaluate frontier models also function as data-efficient reward signals for reinforcement learning. The broader progression reflects a general trend: as evaluation criteria become more expressive, they better distinguish model capabilities, but they also become harder to automate reliably. Our results suggest that this investment in expressiveness pays a second dividend: rubrics rich enough to evaluate frontier models are also rich enough to train them.

LLM-as-a-Judge. Rubric-based evaluation typically relies on LLM judges to assess criterion satisfaction. A growing body of work examines the reliability of this approach. Gu et al. (2024) provide a comprehensive survey, while Yamauchi et al. (2025) find that evaluation criteria are more important than chain-of-thought reasoning for reliability. Schroeder and Wood-Doughty (2024) demonstrate limitations of single-shot LLM evaluations, and Chehbouni et al. (2025) argue from a measurement-theoretic perspective that LLM-judge validity remains undertested. Rao and Callison-Burch (2026) consolidate many of these design choices into Autorubric, a unified open-source framework for rubric-based LLM evaluation that operationalizes per-criterion atomic evaluation, ensemble judging, and psychometric reliability metrics. These concerns motivate careful rubric design that reduces evaluator ambiguity.

Agentic Evaluation. Agent benchmarks have evolved from simplified web interfaces toward realistic, execution-based environments (Zhou et al., 2024). Customer service evaluation is addressed by τ-bench (Yao et al., 2024) and its successor τ^2-bench (Barres et al., 2025). Toolathlon (Li et al., 2025) benchmarks agents on 108 diverse, long-horizon tasks. A survey of 306 practitioners found that 68% of deployed agents execute ten or fewer steps before human intervention (Pan et al., 2025), underscoring the gap between benchmark performance and deployment readiness.

Rubrics as Training Signals. Reinforcement Learning from Verifiable Rewards (RLVR), popularized by DeepSeek-R1 (DeepSeek-AI, 2025) and developed in the open by Tülu 3 (Lambert et al., 2024), has been extended to instruction following through rubric-based reward signals. RIFL (He et al., 2025) uses a finetuned rubric verifier to provide rewards for RL training. VerIF (Peng et al., 2025) combines rule-based and LLM-based verification. RLCF (Viswanathan et al., 2025) extracts instruction-specific checklists. ToolRL (Qian et al., 2025) demonstrates that GRPO-based training with rubric rewards enables tool-use generalization.

Concurrent Rubric-Based Work. Several recent and concurrent efforts use rubrics for either evaluation or RL training, and COMPLEX-IF occupies a distinctive position in this landscape along two axes: *rubric authorship* (human-expert vs. LLM-generated) and *purpose* (evaluation only vs. RL training signal). RubricRAG (Dhole et al., 2026) addresses the cost of human authoring by retrieving domain knowledge to gen-

erate query-specific rubrics at inference time. Relative to these, COMPLEX-IF contributes (i) *expert-authored* rubrics, deliberately chosen over synthetic generation to capture pragmatic intent that LLM-generated rubrics tend to miss; (ii) coverage of *instruction-following with high constraint density* (10–40 criteria per prompt), complementary to the medical, scientific, and research-report focuses of prior work; and (iii) explicit *dual-purpose* design: the same rubrics serve as evaluation instruments and as RL reward signals. The five design principles in §3 represent our attempt to consolidate lessons from this line of work into actionable guidance for expert rubric authoring.

3 Design Principles for Expert Rubrics

We begin by articulating five principles for expert rubric construction. These principles are motivated not only by evaluation quality but also by their downstream consequences for RL training: rubrics that more faithfully capture task success produce more informative reward signals, enabling more efficient and transferable learning.

3.1 Principle 1: Maximum Viable Atomicity

Standard rubric design prescribes that criteria should be atomic, each testing one thing. However, naive atomicity can reduce evaluative utility. Consider a prompt asking for the notes in a C^7 chord (C, E, G, B♭). Maximally atomic criteria testing each note independently would assign 75% correctness to a response identifying C, E, G, and B♮, which is a completely different chord (C major seventh).

The principle of *Maximum Viable Atomicity* states that each criterion should reflect the prompt's smallest *meaningful* unit, even when further decomposition is technically possible. This ensures that rubric scores correlate with genuine response quality rather than with superficial partial credit. For RL training, this principle is especially important: overly atomized criteria can reward fundamentally wrong responses that happen to satisfy a majority of sub-criteria, producing misleading gradient signals.

3.2 Principle 2: Intent-Aware Criterion Design

Criteria should reflect the user's pragmatic intent rather than only the literal language of the prompt. Before writing criteria, rubric authors review multiple model responses to the same prompt and articulate the user's intent in their own words, accounting for both the "what" and the "why" of the request.

For example, a user may ask for classes to improve their Spanish while also mentioning that they are already reading economics articles in Spanish. A rubric that takes the phrase "improve my Spanish" at face value might reward beginner-level support, such as basic vocabulary lists, introductory grammar courses, or simple graded readers. Although this superficially matches the stated goal, it ignores the contextual signal that the user is already operating at a more advanced, domain-specific level. In an RL setting, this does more than incorrectly score one response: it actively teaches the model to privilege surface phrasing over pragmatic intent, reinforcing generic help even when the prompt contains evidence that a more calibrated response is needed.

Intent-aware criteria instead reward responses that infer the user's actual proficiency and goals from context, such as recommending advanced Spanish coursework, economics-focused reading practice, or discussion-based classes conducted in Spanish. As a training signal, intent-aware criteria teach models to attend to contextual cues rather than surface-level instruction parsing.

3.3 Principle 3: Three-Category Criterion Taxonomy

Rather than treating all criteria as equivalent, the framework organizes them into three categories reflecting distinct relationships between criteria and user experience. Throughout, we use the following internal nomenclature; readers unfamiliar with these terms can read them as "standard", "bonus", and "penalty-only" criteria respectively.

Primary Intent criteria are derived directly from the prompt and represent requirements the user would expect to be met. These constitute the majority of most rubrics and serve as the principal source of reward signal during training.

Extra Credit criteria (bonus-only) capture elements not requested but likely to enhance the user's experience (e.g., including inflation-adjusted figures alongside historical prices). They carry no penalty when unfulfilled but earn additional reward when satisfied.

Dodged Bullet criteria (penalty-only) check that the response avoids propagating likely misconceptions the user may be unaware of. They incur a penalty when violated but carry no reward when

fulfilled.

This asymmetric weighting provides a more nuanced reward signal than uniform criterion scoring, encouraging models to satisfy explicit requirements, reach for excellence, and avoid subtle but consequential errors. We describe how this taxonomy maps to the actual RL reward function in §5.3.

3.4 Principle 4: Iterative LLM-Judge Calibration

Since rubrics are evaluated by LLM judges (both during evaluation and during RL training), criterion language must be interpretable by both human evaluators and the judge model. Each criterion undergoes an iterative verification process: the rubric author drafts a criterion and assesses a reference response; the criterion is then evaluated by an LLM verifier; disagreements are diagnosed and resolved through revised criterion language; and the author modifies the reference response to elicit the opposite judgment and verifies that the verifier tracks the change.

This process surfaces subtle ambiguities. For instance, a criterion requiring a poem to "avoid alliteration" may cause a verifier to flag "his heart" as alliterative. Revising to specify "poetic alliteration (repetition of initial stressed consonant sounds)" resolves the disagreement while preserving intent. For RL applications, this calibration step is critical: ambiguous criteria introduce noise into the reward signal, which can lead to reward hacking or unstable training dynamics.

3.5 Principle 5: Domain-Grounded Task Complexity

For agentic tasks, rubrics can decompose success across multiple dimensions that reflect authentic professional workflows. In the CoreCraft environment (Mehta et al., 2026), rubric criteria span four categories: *completeness* (did the agent address all required aspects?), *correctness* (are factual claims accurate given the world state?), *constraint satisfaction* (are business rules and policies correctly applied?), and *format compliance* (does the output follow required structure?). As training signal, this multi-dimensional decomposition provides dense, criterion-level feedback that enables the RL optimizer to identify and correct specific failure modes rather than treating task failure as a monolithic signal.

4 The COMPLEX-IF Dataset

To validate the design principles above and to provide a concrete instantiation of expert rubric-based evaluation, we introduce COMPLEX-IF, an expert-curated instruction-following dataset designed to serve as both a challenging evaluation benchmark for frontier models and an effective training dataset for RLVR.

COMPLEX-IF consists of approximately 1,000 prompts, each paired with 10 to 40 atomic rubric criteria authored by domain experts following the design principles described in §3. Every prompt mirrors authentic professional use cases with high instruction density, including layered constraints, negative instructions, conditional logic, and context-dependent expectations. Tasks are grounded in real-world scenarios such as scheduling under combinatorial constraints, iterative document editing, and multi-requirement content generation.

The dataset consists exclusively of single-turn prompts, each authored to exercise a high density of layered constraints within a single user instruction. Rubric criteria are organized using the three-category taxonomy (Primary Intent, Extra Credit, Dodged Bullet) and undergo the iterative LLM-judge calibration process described in Principle 4.

Data construction. Prompts were authored from scratch by a workforce of domain experts spanning several task categories. Each prompt passed through a three-stage workflow: (i) draft authoring by a domain expert, including a reference response used to anchor rubric writing; (ii) rubric authoring against the reference, including assignment to one of the three taxonomy categories; and (iii) independent review by a second expert, with adjudication by a senior reviewer when authors disagreed. Difficulty was calibrated through a pilot evaluation against a frontier model.

Difficulty profile. Critically, COMPLEX-IF is designed to occupy the capability frontier: no frontier model evaluated exceeds 17% perfect task completion (where all rubric criteria must be satisfied). This difficulty profile is not artificially engineered through adversarial tricks but arises naturally from the realistic complexity of the prompts, making it well-suited for both evaluation and RL training. The continuous gradient between partial and full success, provided by the high rubric density, produces a rich reward landscape for reinforcement

Model	Perfect Task %
GPT-5.1	16.55
Gemini 3 Pro	14.83
Claude Sonnet 4.5	12.41
Kimi K2 Thinking	8.28
DeepSeek v3.2	4.48
Qwen3-235B	4.14

Table 1: Frontier model performance on COMPLEX-IF (single-turn evaluation set). Even the strongest frontier model perfectly satisfies all rubric criteria on only 16.55% of tasks, revealing substantial headroom for rubric-based RL training. Reported numbers reflect frontier models available at the time of COMPLEX-IF evaluation; models released subsequently (e.g., GPT-5.2, Claude Opus 4.6) are not included here but appear in Table 2.

Model	Task Pass %
GPT-5.2 (xHigh Reasoning)	42.6
Claude Opus 4.6 (Adaptive)	30.8
GPT-5.2 (High Reasoning)	29.7
Gemini 3.1 Pro	27.2
DeepSeek v3.2 Thinking	24.1
Claude Opus 4.6	22.1
Claude Sonnet 4.6	16.4

Table 2: Frontier model performance on CoreCraft (Mehta et al., 2026). Even the best model solves fewer than 43% of tasks, establishing substantial headroom for training.

learning.

5 Expert Rubrics as RL Training Signals: Empirical Evidence

Having established what high-quality rubrics look like, we now turn to the central empirical question: *are expert-curated rubrics effective as reward signals for reinforcement learning?* We present evidence across both instruction following and agentic task execution, showing that rubric-based RL training produces substantial, transferable improvements from modest amounts of data.

5.1 Frontier Capability Gaps Motivate Training

Dense expert rubrics reveal trainable capability gaps that coarser metrics hide. By evaluating each response against many atomic criteria, these rubrics show not only whether a model failed a task, but which requirements it failed to satisfy.

Table 1 shows that even the strongest frontier model perfectly satisfies all rubric criteria on only 16.55% of tasks; weaker frontier models fall below 5%. This low perfect-task rate is consistent with the multiplicative difficulty of satisfying many constraints simultaneously: a task with 20 rubric criteria requires every criterion to be satisfied, and even small per-criterion failure probabilities compound rapidly. Crucially, this rubric density creates a continuous gradient between partially and fully correct responses, providing a richer reward landscape for RL than binary pass/fail metrics. While COMPLEX-IF itself is single-turn, prior work documents an average 39% degradation when frontier models are evaluated on multi-turn variants of standard tasks (Laban et al., 2025), suggesting that the constraint-tracking gap we measure here may understate the headroom that remains in conversational settings.

Table 2 shows a similar picture for agentic tasks: even with maximum reasoning effort, frontier models solve fewer than half of CoreCraft tasks.[1] Trajectory analysis reveals recurring failure patterns (poor search strategy, failure to paginate through incomplete results, incomplete exploration of available tools), each diagnosable through specific rubric criteria (Ritchie et al., 2026). This diagnostic granularity is precisely what makes rubrics valuable as training signals: the RL optimizer receives targeted feedback about *which aspects* of task execution need improvement.

5.2 The Hierarchy of Agentic Capabilities

High quality expert rubrics can also make agentic failures easier to interpret and target. In CoreCraft, rubric-level analysis supports the Hierarchy of Agentic Capabilities (Ritchie et al., 2026), an empirically derived framework that organizes common agent failures into five levels. By analyzing trajectories of nine frontier models on 150 CoreCraft tasks, the authors observed that failures cluster around distinct competence levels rather than occurring randomly. The hierarchy comprises:

Level 1: Tool Use. Whether a model can reliably invoke tools with correctly formatted arguments. Models at this level fail at passing the correct data type to a parameter, or confuse field names (e.g., passing "gold" into a `customer_id` field). Models that cannot reliably use tools are not agents; they are chatbots with tool access.

[1] Tables 1 and 2 reflect snapshots of frontier model availability at different evaluation windows. GPT-5.2 and Claude Opus 4.6 were released after the COMPLEX-IF evaluation in Table 1 was completed; based on their CoreCraft scores, we expect their COMPLEX-IF performance to be somewhat higher than reported here but to remain well below 50% perfect task rate.

Level 2: Planning and Goal Formation. Decomposing multi-step tasks into sub-objectives and executing them in order. Weaker models skip steps or forget sub-objectives (e.g., searching only for "fulfilled" orders when "paid" and "pending" are also specified).

Level 3: Adaptability. Adjusting when the plan meets reality. Searching for "Vortex Labs" returns no results because the system stores the brand as "VortexLabs" (no space). Less adaptable models accept empty results; more adaptable models iterate with different search parameters.

Level 4: Groundedness. Staying tethered to the task context. Failures include hallucinating identifiers, using incorrect dates despite explicit context, or misattributing data fields in the final response.

Level 5: Common-Sense Reasoning. Reasoning sensibly about situations not explicitly covered by instructions. Even GPT-5 failed to infer that a support ticket mentioning "the package showed up a few hours ago" describes a return, not a cancellation.

This hierarchy has an implication for RL training design: rubric criteria can target failures at multiple levels of agentic capability, including tool invocation correctness, plan completeness, recovery from unexpected results, factual consistency, and inferential reasoning. Because expert rubrics decompose task success across these capability dimensions, rubric-based RL provides more localized training signal than task-level success alone. This structured reward signal is one reason rubric-based training proves effective, as we demonstrate next.

5.3 Training Methodology: Rubrics as RL Rewards

Both the instruction-following and agentic training pipelines employ RLVR, using expert rubric criteria as the verifiable reward signal. Concretely, given a rubric $C = C_{\text{PI}} \cup C_{\text{EC}} \cup C_{\text{DB}}$ partitioned into the three taxonomy categories from §3.3, and per-criterion satisfaction judgments $s_c \in \{0, 1\}$ provided by an LLM judge, the reward is

$$r = \underbrace{\frac{1}{|C_{\text{PI}}|} \sum_{c \in C_{\text{PI}}} s_c}_{\text{Primary Intent}} + \underbrace{\alpha \frac{1}{|C_{\text{EC}}|} \sum_{c \in C_{\text{EC}}} s_c}_{\text{Extra Credit bonus}} - \underbrace{\beta \frac{1}{|C_{\text{DB}}|} \sum_{c \in C_{\text{DB}}} (1 - s_c)}_{\text{Dodged Bullet penalty}} \tag{1}$$

where $\alpha, \beta \geq 0$ scale the bonus and penalty terms, and any term with an empty criterion set is defined as zero. With $\alpha = \beta = 0$ the reward reduces to the fraction of Primary Intent criteria satisfied; the asymmetric structure encodes the three-category semantics from §3.3: Extra Credit criteria add reward when satisfied but never penalize, and Dodged Bullet criteria penalize violations but never reward satisfaction. This formulation provides a dense learning signal: unlike binary task success, it gives credit for partially satisfying the Primary Intent criteria while allowing Extra Credit and Dodged Bullet criteria to shape the reward around response quality and avoidable errors.

Judge model. Per-criterion satisfaction judgments are produced by GPT-5-mini during COMPLEX-IF training and CoreCraft training. AdvancedIF transfer numbers (Table 4) use the judge specified by He et al. (2025); BFCL, τ^2-Bench, and Toolathlon use their respective official grading protocols.

Training configurations. For instruction following, we train on COMPLEX-IF with LoRA fine-tuning using approximately 900 single-turn tasks from the training split, each with 10–40 expert-authored criteria. For agentic tasks, CoreCraft (Mehta et al., 2026) trains using Group Relative Policy Optimization (GRPO) (Shao et al., 2024) with adaptive clipping (Yu et al., 2025), generating 16 rollouts per prompt that interact with stateful Docker containers running the enterprise simulation.

Reward hacking. We monitored training trajectories for two common forms of reward hacking: (i) responses that verbally satisfy criteria without substantive content, and (ii) responses that exploit known judge biases (e.g., verbosity preference).

Model	Base	Trained	Δ
Qwen3-4B	57.9%	73.4%	+15.5 pp
Qwen3-235B	73.9%	86.1%	+12.2 pp

Table 3: Mean per-criterion pass rate (the fraction of rubric criteria satisfied, averaged across tasks) on the holdout COMPLEX-IF evaluation set after training on ∼900 expert-curated examples via RLVR. Note that this is a different summary statistic than the perfect-task rate reported in Table 1: a model can satisfy a high fraction of individual criteria without satisfying *all* criteria on any given task.

AdvancedIF Category	Base	Trained	Δ
Qwen3-4B			
Single Turn	34.1%	40.1%	+6.0 pp
System Steerability	22.5%	34.9%	+12.4 pp
Multi-Turn Context	28.8%	35.9%	+7.1 pp
Overall	28.2%	36.6%	+8.5 pp
Qwen3-235B			
Single Turn	64.9%	66.2%	+1.2 pp
System Steerability	46.9%	51.7%	+4.7 pp
Multi-Turn Context	49.3%	51.8%	+2.5 pp
Overall	52.4%	55.3%	+2.9 pp

Table 4: Transfer to Meta's AdvancedIF benchmark (He et al., 2025) (task pass rate). Training uses *only single-turn* COMPLEX-IF data, yet multi-turn and system steerability improve substantially.

Qualitative inspection of paired pre/post-training rollouts on a held-out subsample showed no systematic degradation in response quality alongside the reward improvement, and the iterative judge-calibration step (Principle 4) is designed in part to harden criterion phrasing against verbal trickery. The use of a judge model distinct from the policy model further limits shared-representation exploits.

5.4 Results: Generalization on Instruction Following

Table 3 presents the core result: training on approximately 900 expert-curated rubric examples produces +15.5 pp improvement for Qwen3-4B and +12.2 pp for Qwen3-235B on mean per-criterion pass rate. After training, Qwen3-235B reaches 86.1% mean per-criterion pass rate, and the trained 4B model (73.4%) approaches the baseline performance of its 50×-larger counterpart (Qwen3-235B baseline: 73.9%) on the same metric, demonstrating that rubric-dense training can substantially close the capability gap between smaller and frontier models. We caution that the reported numbers reflect single-seed training runs; we have not yet quantified seed-to-seed variance, and treat the magnitude rather than the precise value of each Δ as the more reliable signal.

Benchmark	Base	Trained	Δ
CoreCraft (held-out)	25.4%	36.8%	+11.4 pp
BFCL Parallel	91.0%	95.5%	+4.5 pp
τ^2-Bench Retail	68.7%	76.1%	+7.4 pp
Toolathlon (Pass@1)	18.8%	25.6%	+6.8 pp

Table 5: GLM 4.6 after one epoch of GRPO on CoreCraft with rubric-based rewards (Mehta et al., 2026). The last three rows are out-of-distribution benchmarks the model was never trained on. GLM 4.6 was the latest available model at the start of training; the frontier evaluations in Table 2 includes GLM-5, released subsequently.

Table 4 demonstrates that these gains transfer to Meta's AdvancedIF benchmark, designed independently with different rubric authors. Notably, COMPLEX-IF training consists exclusively of single-turn tasks, yet both models improve on multi-turn evaluation: +7.1 pp on carried context and +12.4 pp on system steerability for Qwen3-4B. This cross-format transfer is somewhat surprising given that single-turn training data does not directly model conversational dynamics; we hypothesize that constraint-tracking competencies learned from satisfying 10–40 simultaneous criteria generalize to retaining and respecting instructions across conversation turns, with system steerability showing the largest gains because system prompts function as persistent constraints directly analogous to the dense constraint sets in COMPLEX-IF training data. A controlled investigation of the transfer mechanism remains future work.

5.5 Results: Generalization on Agentic Tasks

Table 5 presents results from training GLM 4.6 (357B parameters, 32B active) on CoreCraft with rubric-based rewards. After a single epoch, the model improves by 11.4 pp on held-out CoreCraft tasks, a gain exceeding the capability gap between Claude Sonnet 4.5 and Claude Opus 4.5 (+7.05 pp) (Mehta et al., 2026).

More importantly, these gains transfer to out-of-distribution benchmarks: +4.5 pp on BFCL Parallel function calling, +7.4 pp on τ^2-Bench Retail customer service, and +6.8 pp on Toolathlon long-horizon tool use. The Toolathlon result is particularly notable because its tasks (Kubernetes management, Canvas grading, database synchronization) differ substantially from CoreCraft's customer support domain. The model's Pass^3 (the fraction of

tasks solved on all three independent runs, a reliability variant of pass@k) nearly doubles from 9.3% to 17.6%, indicating that rubric-based training improves not just peak capability but also reliability (Mehta et al., 2026).

Qualitative analysis of paired trajectories reveals three categories of learned competencies: improved multi-step workflow execution (correct task decomposition and sequencing), better constraint handling (accurate temporal filtering and policy application), and higher response quality (structured, professional communication). These competencies align with the multi-dimensional decomposition described in Principle 5 (§3), confirming that decomposing task success across multiple rubric dimensions provides training signal that develops each competency independently.

5.6 Discussion: Why Expert Rubrics Are Effective Training Signals

Across both domains, three properties of expert rubrics explain their training effectiveness, each tied directly to the design principles from §3.

Dense reward from rubric granularity. With 10–40 criteria per prompt, the model receives detailed feedback about which specific aspects of task completion succeeded or failed (Principles 1 and 5). This enables more precise credit assignment than binary task-level rewards or holistic preference judgments. A response satisfying 28 of 30 criteria receives a meaningfully different reward from one satisfying 15, guiding the optimizer toward targeted improvements.

Optimal task difficulty. Expert rubric design helps calibrate task difficulty: annotators can build prompts, evaluate frontier model responses against the rubric, and then iterate on the prompt to target an appropriate difficulty range. By designing tasks where no model exceeds a 17% perfect-task rate on COMPLEX-IF and no model exceeds 43% on CoreCraft, the training distribution occupies a region where learning signal is most informative. Tasks that are too easy or too hard provide minimal gradient; expert-authored prompts and rubrics allow task difficulty to be adjusted toward the productive middle ground.

Data efficiency from expert curation. Expert curation can concentrate training signal in a small number of high-fidelity examples. Training on COMPLEX-IF achieves an 8.45% overall improvement on AdvancedIF with approximately 1,000 expert-curated examples, despite AdvancedIF being an independently authored benchmark. This gain is larger than the 6.7% improvement reported by RIFL (He et al., 2025) on AdvancedIF, its in-distribution evaluation setting, using manually written prompts paired with synthetically generated rubrics. Because the base models and training pipelines differ, this comparison is suggestive rather than controlled, but it is consistent with the view that expert rubrics can provide unusually dense and generalizable supervision. This data efficiency also echoes the Superficial Alignment Hypothesis (Zhou et al., 2023a): small amounts of high-quality data can suffice when the training distribution sits at the capability frontier.

The efficiency advantage plausibly arises because expert curation captures pragmatic and categorical distinctions that are difficult to control in synthetic rubric generation (Principles 2, 3, and 4). Recent work on automatic rubric generation finds a substantial gap between human-annotated and model-generated rubrics, indicating that even state-of-the-art models struggle to autonomously specify valid evaluation criteria (Zhang et al., 2026). In our setting, these distinctions include inferring pragmatic user intent, separating Primary Intent from Extra Credit, and calibrating criteria for reliable judge interpretation.

These properties also explain the observed cross-format and cross-domain transfer. The constraint-tracking, workflow decomposition, and quality standards learned from rubric-dense training are general competencies, not environment-specific heuristics.

6 Broader Discussion

Construct Validity. Expert rubrics improve construct validity along three dimensions. *Content validity*: rubrics extend the evaluable space to include semantic correctness and pragmatic appropriateness (e.g., AdvancedIF's system steerability requires evaluating whether a fitness assistant remembers a user's ACL injury (He et al., 2025)). *Predictive validity*: models scoring well on decomposed rubrics show improvements on out-of-distribution benchmarks (Mehta et al., 2026), while IFEval scores do not predict performance on more realistic tasks (Pyatkin et al., 2025). *Discriminant validity*: the wide spread in perfect-task rates among frontier models on COMPLEX-IF (from 4.14% to 16.55%, a roughly 4× spread between the weakest and strongest model) creates space in which

meaningful capability differences become visible, in contrast to saturated benchmarks where frontier models cluster within a few percentage points of each other.

Infrastructure Investment. Expert rubric creation costs more, but our results demonstrate that this investment yields returns in *both* measurement quality and training efficiency. Hybrid approaches that combine expert-authored seed rubrics with synthetic scaling (Dhole et al., 2026) may offer a middle ground; the finetuned RIFL verifier achieves 0.728 F1 agreement with humans, compared to 0.515 for a vanilla LLM judge (He et al., 2025), suggesting that even small amounts of expert supervision substantially raise the ceiling on automated rubric quality.

7 Conclusion

We have presented expert-curated rubric-based evaluation as a paradigm that improves both the measurement and the training of large language models. We articulated five design principles—Maximum Viable Atomicity, Intent-Aware Criterion Design, the three-category Primary Intent / Extra Credit / Dodged Bullet taxonomy, Iterative LLM-Judge Calibration, and Domain-Grounded Task Complexity—and instantiated them in COMPLEX-IF, a 1,000-prompt instruction-following dataset with 10–40 atomic criteria per prompt. Training with rubric-based rewards produces substantial in-distribution gains and transfers to out-of-distribution benchmarks designed by independent teams in both instruction-following and agentic settings. We release COMPLEX-IF publicly to support research on rubric-based evaluation and training. Open questions—including controlled isolation of which design principles drive transfer, broader use of asymmetric reward shaping, and integration with synthetic rubric generation pipelines—remain natural next steps.

Limitations

We highlight several limitations of the work as a guide for future research and interpretation.

Single-seed training results. The headline training numbers in Tables 3, 4, and 5 reflect single training runs without seed-variance estimates. We expect the magnitudes (Δ on the order of 5–15 pp) to be robust to seed choice, but precise values should be interpreted with caution. Multi-seed replications are an important next step.

No controlled isolation of design principles. We articulate five design principles and present evidence that the resulting rubrics produce strong training signals, but we do not run controlled ablations that isolate the contribution of each principle. Designing such ablations is a clear next step.

Ethics Statement

Rubrics use. Rubrics that grade compliance with arbitrary instructions can in principle be used to train models toward harmful objectives. COMPLEX-IF prompts target professional and benign tasks, but the design principles in §3 are domain-agnostic. We encourage downstream users to apply standard safety filtering before training on derived data and to maintain held-out safety-relevant rubrics that are never used as RL rewards.

References

Rahul K. Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñonero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, Johannes Heidecke, and Karan Singhal. HealthBench: Evaluating large language models towards improved human health. 2025. arXiv preprint arXiv:2505.08775.

Victor Barres, Honghua Dong, Soham Ray, Xujie Si, and Karthik Narasimhan. τ^2-Bench: Evaluating conversational agents in a dual-control environment. 2025. arXiv preprint arXiv:2506.07982.

Khaoula Chehbouni, Mohammed Haddou, Jackie Chi Kit Cheung, and Golnoosh Farnadi. Neither valid nor reliable? investigating the use of LLMs as judges. 2025. In NeurIPS Position Paper Track. arXiv preprint arXiv:2508.18076.

DeepSeek-AI. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *Nature*, 645:633–638, 2025.

Kaustubh D. Dhole et al. RubricRAG: Towards interpretable and reliable LLM evaluation via domain knowledge retrieval for rubric generation. 2026. arXiv preprint arXiv:2603.20882.

Maria Eriksson, Erasmo Purificato, Arman Noroozian, João Vinagre, Guillaume Chaslot, Emilia Gómez, and David Fernández-Llorca. Can we trust AI benchmarks? an interdisciplinary review of current issues in AI evaluation. 2025. arXiv preprint arXiv:2502.06559.

Jiawei Gu, Xuhui Jiang, et al. A survey on LLM-as-a-judge. 2024. arXiv preprint arXiv:2411.15594.

Yun He, Wenzhe Li, Hejia Zhang, Songlin Li, Karishma Mandyam, Sopan Khosla, Yuanhao Xiong, Nanshu

Wang, Selina Peng, Beibin Li, et al. AdvancedIF: Rubric-based benchmarking and reinforcement learning for advancing LLM instruction following. 2025. arXiv preprint arXiv:2511.10507.

Yuxin Jiang, Yufei Wang, Xingshan Zeng, Wanjun Zhong, Liangyou Li, Fei Mi, Lifeng Shang, Xin Jiang, Qun Liu, and Wei Wang. FollowBench: A multi-level fine-grained constraints following benchmark for large language models. 2024. In Proceedings of ACL. arXiv preprint arXiv:2310.20410.

Philippe Laban, Hiroaki Hayashi, Yingbo Zhou, and Jennifer Neville. LLMs get lost in multi-turn conversation. 2025. arXiv preprint arXiv:2505.06120.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, et al. Tülu 3: Pushing frontiers in open language model post-training. 2024. arXiv preprint arXiv:2411.15124.

Junlong Li, Wenshuo Zhao, Jian Zhao, Weihao Zeng, Haoze Wu, Xiaochen Wang, Rui Ge, et al. The tool decathlon: Benchmarking language agents for diverse, realistic, and long-horizon task execution. 2025. In ICLR. arXiv preprint arXiv:2510.25726.

Sushant Mehta, Logan Ritchie, Suhaas Garre, Ian Niebres, Nick Heiner, and Edwin Chen. EnterpriseBench CoreCraft: Training generalizable agents on high-fidelity RL environments. 2026. arXiv preprint arXiv:2602.16179.

Melissa Z. Pan, Negar Arabzadeh, Riccardo Cogo, Yuxuan Zhu, Alexander Xiong, et al. Measuring agents in production. 2025. arXiv preprint arXiv:2512.04123.

Hao Peng, Yunjia Qi, Xiaozhi Wang, Bin Xu, Lei Hou, and Juanzi Li. VerIF: Verification engineering for reinforcement learning in instruction following. 2025. In Proceedings of EMNLP. arXiv preprint arXiv:2506.09942.

Valentina Pyatkin, Saumya Malik, Victoria Graf, Hamish Ivison, Shengyi Huang, Pradeep Dasigi, Nathan Lambert, and Hannaneh Hajishirzi. Generalizing verifiable instruction following. 2025. arXiv preprint arXiv:2507.02833.

Cheng Qian, Emre Can Acikgoz, Qi He, Hongru Wang, Xiusi Chen, Dilek Hakkani-Tür, Gokhan Tur, and Heng Ji. ToolRL: Reward is all tool learning needs. 2025. In Advances in NeurIPS. arXiv preprint arXiv:2504.13958.

Yiwei Qin, Kaiqiang Song, Yebowen Hu, Wenlin Yao, Sangwoo Cho, Xiaoyang Wang, Xuansheng Wu, Fei Liu, Pengfei Liu, and Dong Yu. InFoBench: Evaluating instruction following ability in large language models. 2024. In Findings of ACL. arXiv preprint arXiv:2401.03601.

Delip Rao and Chris Callison-Burch. Autorubric: Unifying rubric-based LLM evaluation. 2026. arXiv preprint arXiv:2603.00077.

Logan Ritchie, Sushant Mehta, Nick Heiner, Michelle Yu, and Edwin Chen. The hierarchy of agentic capabilities: Evaluating frontier models on realistic enterprise environments. 2026. arXiv preprint arXiv:2601.09032.

Kayla Schroeder and Zach Wood-Doughty. Can you trust LLM judgments? reliability of LLM-as-a-judge. 2024. arXiv preprint arXiv:2412.12509.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, et al. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. 2024. arXiv preprint arXiv:2402.03300.

Vijay Viswanathan, Yanchao Sun, Shuang Ma, Xiang Kong, Meng Cao, Graham Neubig, and Tongshuang Wu. Checklists are better than reward models for aligning language models. 2025. In Advances in NeurIPS. arXiv preprint arXiv:2507.18624.

Bosi Wen, Pei Ke, Xiaotao Gu, Lindong Wu, Hao Huang, Jinfeng Zhou, et al. Benchmarking complex instruction-following with multiple constraints composition. 2024. In NeurIPS Datasets and Benchmarks. arXiv preprint arXiv:2407.03978.

Yusuke Yamauchi, Taro Yano, and Masafumi Oyamada. An empirical study of LLM-as-a-judge: How design choices impact evaluation reliability. 2025. arXiv preprint arXiv:2506.13639.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. τ-bench: A benchmark for tool-agent-user interaction in real-world domains. 2024. arXiv preprint arXiv:2406.12045.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, et al. DAPO: An open-source LLM reinforcement learning system at scale. 2025. arXiv preprint arXiv:2503.14476.

Qiyuan Zhang, Junyi Zhou, Yufei Wang, Fuyuan Lyu, Yidong Ming, Can Xu, Qingfeng Sun, Kai Zheng, Peng Kang, Xue Liu, and Chen Ma. RubricBench: Aligning model-generated rubrics with human standards. 2026. arXiv preprint arXiv:2603.01562.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. LIMA: Less is more for alignment. In *Advances in NeurIPS*, 2023a.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models. 2023b. arXiv preprint arXiv:2311.07911.

Shuyan Zhou, Frank F. Xu, Hao Zhu, et al. WebArena: A realistic web environment for building autonomous agents. In *ICLR*, 2024.