# RUBRICFLY: RUBRIC-GUIDED ADAPTIVE EVALUATION AND JUDGE CALIBRATION FOR MULTILINGUAL TRANSLATION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

As large language models improve at translation, static benchmarks are losing diagnostic value: overall scores separate models less and reveal little about which capabilities remain weak. Sustaining that value requires tests that target specific capabilities, such as meaning preservation, register, and cultural adaptation, with question-specific criteria, and that are renewed as models improve. Yet translation benchmarks are mostly authored and updated by hand, and automatic generation for translation either optimizes difficulty alone or yields a fixed pool of items that is not renewed from observed model behavior. Model failures could guide renewal, but only if the Judge reliably separates genuine failures from valid translations; rejecting a valid rendering misstates a model's weaknesses and misdirects the next round of tests. We present RubricFly, the first LLM-driven translation evaluation framework that combines automated question and rubric generation with feedback-guided benchmark renewal and human-guided Judge rule learning. Shared testpoints specify the capability under test and the response requirements that guide both construction and evaluation. A production loop turns model failures and coverage gaps into new questions and rubrics, and a Judge loop distills human corrections into reusable decision rules without updating model parameters. We build *fly-corpus*, a 500-question benchmark covering bidirectional translation and cross-lingual rewriting between Chinese and English, French, Japanese, Spanish, and Korean. Under a common GPT-5.5 Medium judging protocol, the 47 questions introduced by renewal average 43.32 across ten models, against 80.95 for the saturated questions they replace; full-feedback replacements are harder than those generated without model feedback (43.06 vs. 51.25) and separate models more than difficulty-only ones. Rule-guided rechecking reduces essential-item errors from 307 to 291 under three-draw voting, more than rechecking without rules, although response-level errors are unchanged. Together, these results show how capability diagnosis can drive test renewal and how human corrections can improve the judgments on which renewal depends.

## 1 INTRODUCTION

As language models improve, static, general-purpose translation benchmarks become less informative about what separates models and what they still get wrong (Kocmi et al., 2025; Zouhar et al., 2026b). Harder questions alone do not restore this signal: a benchmark must also cover a broad range of translation capabilities and apply criteria fine-grained enough to separate meaning preservation from appropriateness for the intended audience, register, and communicative purpose (House, 2015; Sun et al., 2026). Keeping a benchmark informative therefore means turning the weaknesses it reveals into new questions and new criteria.

Existing approaches cover parts of this cycle but offer limited support for renewing translation tests automatically with explicit capability targets and question-specific rubrics (Table 1). LTB pairs targeted challenges with verification rules through manual authoring and review, whereas MT-Breaker automates the search for difficult source texts but does not construct rubric items for them (Zouhar et al., 2026b;a). Error taxonomies and constraint checks make failures explicit (Kocmi & Federmann, 2023; Sun et al., 2026), yet a failed requirement does not say how to build a new question that probes



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

the same capability in another scenario or fills a coverage gap. Judge errors compound the problem: rejecting a valid translation misrepresents a model's weaknesses and misdirects the next round of generation (Fu & Liu, 2025).

We present RubricFly, the first LLM-driven translation evaluation framework that combines automated question and rubric generation with feedback-guided renewal and human-guided Judge rule learning. Its central unit is the *testpoint*, which specifies a target capability, a source phenomenon that elicits it, and the observable requirements a response must meet within a given task and communicative scenario. Because a question and its rubric derive from the same testpoint, a failed requirement points directly to what the next round should test. Rather than treating the benchmark as a fixed artifact and the Judge as a fixed instrument, RubricFly updates both: a production loop turns failures and coverage gaps into new questions and rubrics, and a Judge loop distills human corrections into reusable decision rules while the model parameters, rubric text, and scoring stay fixed.

We instantiate the framework in *fly-corpus*, a 500-question benchmark of translation and cross-lingual rewriting across five Chinese-centered language pairs, audited by experts and accompanied by a separate set of blinded human reference labels. Three findings emerge. First, task-level scores reveal differences that overall rankings hide: Qwen-3.8 and GLM differ by only 0.31 points overall but by 16.77 on style-adaptive translation. Second, renewal adds challenge, and how it is targeted matters: under a common GPT-5.5 Medium protocol, the 47 replacement questions average 43.32 against 80.95 for the saturated questions they replace, and full-feedback replacements are harder than those generated without model feedback and separate models more than difficulty-only ones. Third, human corrections can be reused: rule-guided rechecking reduces essential-item errors by 5.21% on questions excluded from rule induction, the largest reduction among the rechecking arms, and case analysis shows two corrected judgments alongside the rules routed to them.

## 2 Related Work

Fine-grained assessment, test renewal, and judge adaptation have each been studied, but they are rarely combined in translation evaluation (Table 1).

**Fine-grained assessment.** Beyond shared test sets and protocols such as FLORES-101 and WMT (Goyal et al., 2022; Kocmi et al., 2023), recent work explains why a translation fails. GEMBA-MQM and RUBRIC-MQM locate error spans under a common error taxonomy (Kocmi & Federmann, 2023; Kim, 2025), while LTB and IFMTBench attach checks to individual items, as expert-written verification rules or as generated hard and soft constraints (Zouhar et al., 2026b; Sun et al., 2026); outside translation, InFoBench, TICK, WildBench, and HealthBench likewise score responses against item-level checklists or rubrics (Qin et al., 2024; Cook et al., 2024; Lin et al., 2024; Arora et al., 2025). Such checks make failures explicit, but they are designed for a fixed test set: a failed check does not specify how to build the next question. RubricFly derives each question and its rubric from the same testpoint, which names the target competence, the eliciting source phenomenon, and the required behavior, so a failed rubric item maps directly to a generation target; the same scheme covers both translation and cross-lingual rewriting.

Table 1: Comparison with representative translation benchmarks.

| Work | Trans. | Crosslingual Rewrite | Auto generation | Itemized rubrics | Criterion design | Test renewal | Judge self-evolution |
|---|---|---|---|---|---|---|---|
| FLORES-101 (Goyal et al., 2022) | ✓ | – | – | – | N/A | – | – |
| WMT23 (main) (Kocmi et al., 2023) | ✓ | – | – | – | N/A | – | – |
| LTB (Zouhar et al., 2026b) | ✓ | – | – | ✓ | Verification checks | Human updates | – |
| MT-Breaker (Zouhar et al., 2026a) | ✓ | – | ✓ | – | N/A | Source refinement | – |
| IFMTBench (Sun et al., 2026) | ✓ | – | ✓ | ✓ | Hard + soft constraints | – | – |
| **RubricFly (ours)** | ✓ | ✓ | ✓ | ✓ | Essential + quality items | **Failures + gaps** | ✓ |

**Test renewal.** A benchmark stays informative only if it is renewed, and translation benchmarks renew in different ways: LTB through continued manual authoring and review (Zouhar et al., 2026b), HardMTBench by selecting difficult examples from an existing corpus (Li et al., 2026b), and MT-Breaker by refining source texts with model feedback, without rubrics for the refined texts (Zouhar

et al., 2026a). IFMTBench generates tasks and rubric items together but does not renew its released items from observed model behavior (Sun et al., 2026). In translation, renewal is thus manual, bounded by an existing pool, or driven by difficulty alone. Outside translation, Dynabench collects adversarial examples with humans and models in the loop (Kiela et al., 2021), DyVal generates evaluation samples dynamically (Zhu et al., 2024), and AutoBencher and BenchBuilder construct benchmarks that optimize difficulty and model separation (Li et al., 2025; 2024); none of them derives item-level rubrics from capability targets. RubricFly renews questions and rubrics jointly from diagnosed failures and coverage gaps, targeting the requirements on which models differ rather than difficulty alone.

**Judge adaptation.** Judge adaptation has been explored through explicit evaluation criteria, prompt refinement, and feedback-based memory. Prometheus and Auto-Rubric make evaluation criteria explicit (Kim et al., 2024; Xie et al., 2026); GEPA and ACE refine prompts or accumulated context from feedback (Agrawal et al., 2026; Zhang et al., 2026); and MemAlign stores judging principles distilled from human feedback (The Mosaic Research Team, 2026). GrowLoop and CoReflect go further, revising conversational tests together with their rubrics (Lin et al., 2026; Li et al., 2026a), and EvalGen finds that evaluation criteria drift as graders inspect outputs (Shankar et al., 2024). When tests, criteria, and judge change together, however, a shift in scores cannot be attributed to any one of them. RubricFly keeps the two updates apart: the production loop renews questions and rubrics, and the Judge loop learns decision rules while the rubric text stays fixed, so gains in agreement reflect how requirements are interpreted rather than what they require. Table 1 compares translation benchmarks only; the general-domain systems above are therefore not listed.

## 3 RubricFly: Benchmark Construction and Evaluation

RubricFly traces every score to a target capability through shared testpoints, then routes the verdicts into test renewal or Judge rule learning (Figure 1).

### 3.1 Capability-guided Test Design

We separate *tasks*, which specify the deliverable and its use setting, from *competencies*, which name the abilities under test. A task label alone conflates the two: a legal translation can fail through a wrong number, an inappropriate register, or a changed obligation, and each failure points to a different ability. Describing each question along separate axes lets results be sliced by any combination, such as strategic decisions in legal texts. The task families are ordered by a principle of functionalist translation theory, that the purpose of a text determines how far its rendering may depart from the source (Reiss & Vermeer, 2014; Nord, 2018). The four translation families move from source-bound rendering (faithful, professional) to target-oriented adaptation (style-adaptive, cultural), and the four cross-lingual rewriting families differ in what the output changes: quality, form, style, or content. Twenty-six generation routes instantiate the eight families (Table 2).

Each competence family operationalizes an established concept from translation studies: equivalence underlies semantic fidelity (Nida, 1964; Baker, 2018), translation shifts underlie structural transfer (Catford, 1965; Vinay & Darbelnet, 1995), register analysis underlies register command (House, 2015), domestication and foreignization underlie strategic decision (Venuti, 1995), and translation norms underlie norm compliance (Toury, 1995). The nine families and 27 subcompetencies are each defined by observable behavior, and they recur across routes; explaining medical findings to patients, for example, calls for both register command and purpose alignment. Appendix A.2 gives the source of every family.

A *testpoint* makes a target capability testable. It records the source phenomenon, scenario, intended operation, and the observable requirement; numerical fidelity, for example, can be tested in a financial report or an everyday notice. The testpoint guides the construction of both the question and its rubric, so a failed requirement can be traced back to a generation target.

### 3.2 Automated Question and Rubric Construction

Construction turns a design target, a combination of task, scenario, language pair, and competence, into a question and rubric in two stages linked by a testpoint: Stage 1 gathers source material and decides what to test, and Stage 2 writes the question and then its rubric. The model assignments

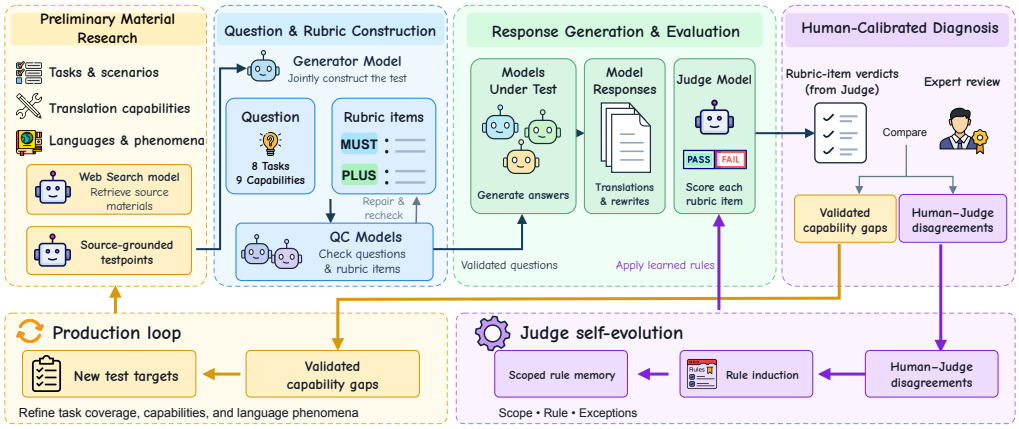


Figure 1: RubricFly framework. Shared testpoints link each question to its rubric. Production feedback renews both (gold), and human corrections update the rule memory of a frozen Judge (purple). The diagram shows the full protocol, including validation stages; our studies examine released production traces and one Judge update per fold.

Table 2: Design axes of RubricFly. Appendix A gives definitions and counts; rubric roles are assigned per requirement.

| Design axis | Categories |
|---|---|
| **Task coverage**<br>8 families, 26 routes | *Translation:* faithful, professional, style-adaptive, cultural adaptation.<br>*Cross-lingual rewriting:* polishing, format, style, creative. |
| **Competence targets**<br>9 families,<br>27 subcompetencies | Semantic fidelity; structural transfer; register command; cultural mediation; strategic decision; purpose alignment; creative recreation; discourse management; norm compliance. |
| **Rubric dimensions**<br>7 dimensions | MUST by default: instruction following, content correctness, completeness, safety and compliance.<br>Mainly PLUS: expression and format, relevance, consistency; MUST when prescribed. |

below are those used to generate the 47 replacement questions; Appendix A.4 lists the assignments for all construction runs.

**Stage 1: Preliminary material research.** We use Gemini 3.1 Pro to turn a combination of task, scenario, language pair, and competence target into a research strategy and search queries, and a search-enabled Doubao Seed 2.0 Pro (high) to retrieve background and source material. DeepSeek V4 Pro then synthesizes the retrieved evidence into source material and extracts testpoints, each specifying the linguistic challenge, a typical error, and the requirement a response should meet.

**Stage 2: Question and rubric construction.** We use DeepSeek V4 Pro to turn each testpoint into a question consisting of a source passage and an instruction that fixes the target language, audience, and requested output. Given the completed question, the same model writes the rubric, tying each item to the passage or the instruction. Each rubric item states one independently judgeable requirement and carries a dimension, a MUST or PLUS role, and a weight. Writing the rubric after the question keeps scoring anchored to the actual task while admitting different valid translations.

**Dimensions.** A dimension records what an item assesses. The seven dimensions are shared by translation and rewriting, so failures can be compared across task families. Content correctness, completeness, expression and format, and consistency together cover the MQM issue types of accuracy, fluency, terminology, style, and locale convention (Freitag et al., 2021); instruction following, relevance, and safety and compliance record requirements set by the task rather than by the source. In initial construction, the rubric prompt also emphasized the primary dimension of each labeled competence, such as content correctness for semantic fidelity (Appendix A.2).

**Roles.** A role decides how a verdict counts. A MUST item states a requirement that a usable response cannot miss. Requirements of instruction following, content correctness, completeness, and safety and compliance are therefore MUST by default, whereas expression and format, relevance, and consistency mostly describe quality beyond this minimum and are PLUS (Table 2). When a question explicitly prescribes a format, genre, register, or rhetorical effect, however, the corresponding items

also become MUST, because a mismatch would make the deliverable unusable: the requested notice format, for example, is the MUST item R2 in Figure 2 (Appendix A.5).

**Quality control.** Generated rubrics are reviewed before use rather than trusted as produced. The construction pipeline includes a preliminary automated check in which models from families other than the generator's review each rubric item and flag those requiring revision (Panickssery et al., 2024). For fly-corpus, multilingual experts audited the 8,683 rubric items of the 500 questions in the initial benchmark (Section 3.4), covering wording, dimension, role, and judgment; before release, a Judge-boundary audit of the benchmark, including the replacements, revised six problematic items and rejudged all 60 affected responses (Appendix A.4).

### 3.3 RUBRIC-LEVEL EVALUATION AND HUMAN CALIBRATION

**Stage 3: Response generation and evaluation.** The rubric items written in Stage 2 are also the units of evaluation. After the evaluated models answer each question, GPT-5.5 Medium judges each response against one item at a time, receiving the complete question and response and returning a pass/fail verdict with a rationale. Judging at the granularity of the rubric lets every score be traced to the items that failed.

**Scoring.** A translation that turns "5%" into "50%" misrepresents its source however fluent its wording. Strictly Hybrid scoring therefore awards weighted credit only when every MUST item passes: the gate keeps secondary strengths from compensating for essential failures, and the weighted component still distinguishes quality among acceptable responses (Figure 2). The design shares the hard-before-soft principle of IFMTBench (Sun et al., 2026), and each verdict doubles as diagnostic feedback.

For response $y$, let $z_i(y) \in \{0, 1\}$ be the verdict on item $i$, $w_i$ its weight, and $M$ the MUST set:

$$S(y) = \begin{cases} 0, & \exists i \in M : z_i(y) = 0, \\ \dfrac{\sum_i w_i z_i(y)}{\sum_i w_i}, & \text{otherwise.} \end{cases} \tag{1}$$

Roles set the gate independently of weights. Default MUST/PLUS weights of 1/0.5 give an additional-quality item half the credit of an essential one; production evaluation keeps the stored weights after revision, whereas the Judge study fixes the defaults. We report $100S(y)$ averaged equally over questions. A zero means that at least one MUST item failed, so one disputed essential verdict decides the score; MUST items are judged more consistently than PLUS items (Appendix C.3), and Appendix C.2 analyzes sensitivity to aggregation.

**Stage 4: Human-calibrated diagnosis.** A failed rubric item can reflect a response error, an unsuitable requirement, or a Judge mistake. Expert audits and a separate blinded round separate these cases (Section 3.4): valid failures guide renewal, and Human–Judge disagreements supply evidence for correction rules (Section 4).

### 3.4 THE FLY-CORPUS DATASET

*fly-corpus* contains 500 questions covering bidirectional translation and cross-lingual rewriting between Chinese and English, French, Japanese, Spanish, and Korean. Each question pairs an instruction and a source passage with a question-specific rubric; together with the scoring protocol, these form our evaluation benchmark.

**Data sources and curation.** The automated pipeline generated candidate questions from web-retrieved or model-synthesized material; an archived snapshot contains 1,000 candidates, 200 per language pair. Difficulty- and discrimination-oriented selection yielded an initial benchmark of 100 questions per language pair. The production loop then replaced 47 questions on which all ten panel models passed every MUST item in the selection-stage evaluation. The release contains 233 translation and 267 rewriting questions and 8,643 rubric items (4,522 MUST, 4,121 PLUS). Language and task-family labels cover all questions; finer capability and phenomenon labels are verified for the replacements. Of the released questions, 398 build on retrieved and 102 on synthesized material; Appendix A.6 details selection and provenance.

**Human review.** Thirty-nine multilingual experts, recruited through a translation company under written qualification requirements, took part in two annotation rounds (Appendix B.1).

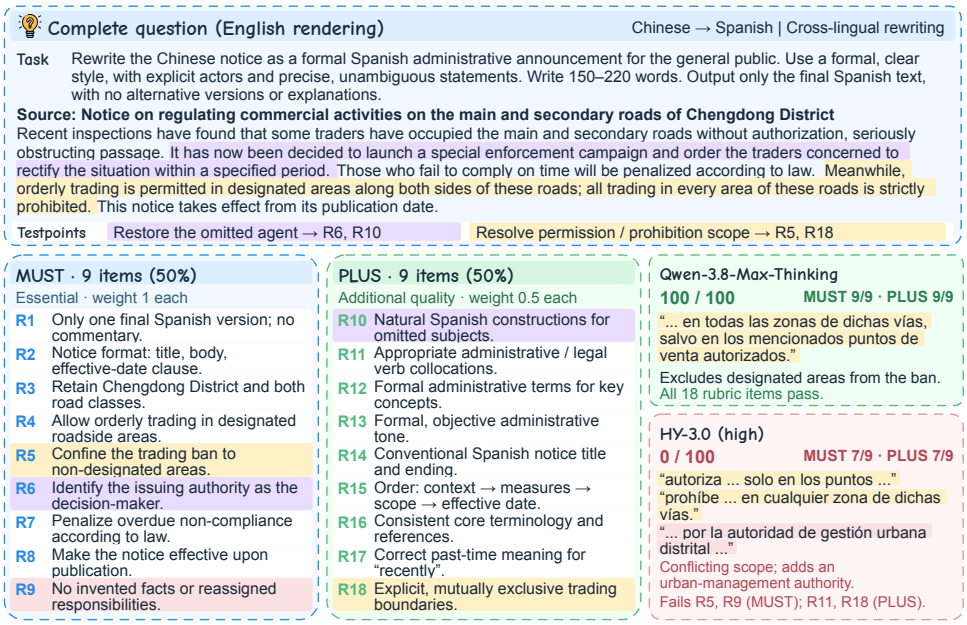


Figure 2: A complete question (English rendering), all 18 rubric items (abbreviated), and Spanish output excerpts. Shares count items; scores are single-question Strictly Hybrid results from stored GPT-5.5 Medium judgments. Appendix A.3 traces production.

*Round 1: production audit.* Thirty-five experts, each with at least five years of professional experience, reviewed one response to each of 500 earlier-version questions, covering 8,683 rubric-item records with the automatic verdicts and rationales visible. They assessed rubric wording, dimensions, roles, and judgment quality (Appendix B.2).

*Round 2: blinded reference annotation.* 23 experts, 19 of whom had also taken part in Round 1, labeled six anonymized responses to each of 100 questions, 20 per language pair, yielding 10,440 rubric-item–response positions, with model identities and automatic judgments hidden. Twenty questions were independently double-annotated, with disagreements resolved by expert review; the remaining 80 were singly annotated. After a final quality check, the annotation consistency rate was 98.14% of the 2,046 double-annotated judgments (Cohen's $\kappa = 0.897$).

## 4 Dual Feedback Loops

Stage 4 separates two kinds of evidence (Figure 1): failures on valid requirements, which show what models cannot yet do, and Human–Judge disagreements, which show where the Judge misreads a requirement. RubricFly routes each kind to its own loop. The production loop turns validated failures into new questions and rubrics without changing the Judge; the Judge loop turns disagreements into decision rules without changing the rubric text. Keeping the two updates apart lets the effect of each be measured on its own (Section 5). Our studies pair each loop with a different Judge: the production loop with the benchmark Judge, and the Judge loop with a frozen open-weight model, whose fixed parameters can be verified and whose multi-draw control arms remain affordable.

### 4.1 Feedback-guided Question and Rubric Renewal

The production loop turns failed rubric items and coverage gaps into explicit generation targets. An aggregate score shows that a model struggles but not what to test next; item-level failures supply that missing target, so renewal follows observed weaknesses while extending the capabilities under test.

Within a fixed development pool, we group failures along the design axes of Section 3.1, namely task, competence, and linguistic phenomenon, and compare the recorded coverage against the intended taxonomy. Scores of a development panel indicate how challenging a question is, and cross-model variation indicates whether it separates models. We use Gemini 3.1 Pro to turn these summaries into specifications for new testpoints, stating the source characteristics, required response operations, and

requirements. Model identities and full answers are withheld so that the targets describe reusable failure patterns rather than individual systems. Stages 1 and 2 (Section 3.2) then instantiate the targets as new questions and rubrics.

Because a question and its rubric are renewed together, the intended challenge becomes observable in the next evaluation. For the Spanish notice in Figure 2, content and consistency failures, together with a structural-transfer coverage gap, lead to a source with omitted agents and conflicting permission boundaries; separate rubric items then test whether a response resolves each problem. After quality assessment and model evaluation, selected questions enter the benchmark, where their item-level results inform the next round. Appendix A.3 traces this example.

### 4.2 JUDGE RULE LEARNING

The Judge loop turns human corrections into reusable decision rules while keeping the model parameters, rubric text, roles, and scoring fixed. Correcting a single verdict fixes one response; a rule that captures why the verdict was wrong can also fix future judgments of other valid translations. Judge learning here means updating this external rule memory.

We group Human–Judge disagreements by task, rubric dimension, role, and language-related scope. Qwen3.6-27B, which also serves as the Judge in this study, uses corrected false rejections and contrastive supporting evidence to propose rule cards. Each card states when the rule applies, what evidence warrants acceptance, and which exceptions restore the original decision boundary. Cards omit full answers and case identifiers, so they describe judging principles rather than individual cases.

At evaluation time, the same Judge uses the cards to recheck initial FAIL verdicts within their scope; all other verdicts are kept. Rechecking targets FAIL verdicts because the production loop acts on failures: a rejected valid translation would send renewal after a weakness the model does not have. For a sarcasm requirement, for example, a card asks whether the irony is recognizable to the target reader, which admits a valid nonliteral rendering while excluding neutral contexts. Rechecking thus corrects only where a rule applies and leaves the evaluation requirement unchanged. Section 5.3 compares the learned rules with rechecking controls.

## 5 EXPERIMENTS

We evaluate RubricFly along three axes: the diagnostic resolution of *fly-corpus* (Section 5.1), the effect of feedback-guided renewal on difficulty and model separation (Section 5.2), and the effect of learned judging rules on agreement with human labels (Sections 5.3 and 5.4).

### 5.1 DIAGNOSTIC RESOLUTION OF FLY-CORPUS

**Setup.** We evaluate ten general-purpose and three translation-specialized models. GPT-5.5 Medium judges every response against its fixed rubric one item at a time, and we report Strictly Hybrid scores on a 0–100 scale. Table 3 restricts the comparison to the 494 questions for which every model has a nonempty response and complete judgments; Appendix C.1 gives evaluation details and separate translation and rewriting comparisons.

**Task-level scores separate models that overall scores tie.** Qwen-3.8 and GLM differ by 0.31 points overall, yet Qwen-3.8 leads by 16.77 on style-adaptive translation and GLM by 10.55 on cultural adaptation (Table 3). Translation-specialized models show a sharper split: Qwen-MT-Plus scores 49.60 on professional translation but 18.54 on style rewriting, and HY-MT2-7B scores 40.30 and 15.54. Their strength in professional translation does not carry over to cross-lingual rewriting.

**Experts accept rubric wording far more often than diagnostic labels.** The production audit scores wording, dimension, and role separately. Experts accepted 99.32% of rubric texts but only 77.31% of dimension assignments and 96.19% of MUST/PLUS roles: a well-formed requirement can still carry the wrong diagnostic label or aggregation role. The release applies every dimension and role correction suggested for retained questions. Reviewers also accepted 99.44% of the visible judge verdicts and rationales on valid items; because these verdicts were visible, this rate may be inflated by anchoring and is not used to measure Judge accuracy (Appendix B.2). Against blinded expert labels, the benchmark Judge agrees on 93.4% of MUST items, above the 89.6% at which two experts agree with each other (Appendix C.3).

Table 3: Strictly Hybrid scores on the same 494 valid fly-corpus questions (0–100). Models are ranked within each group; bold marks column maxima. Judge: GPT-5.5 Medium.

| | **Overall** | **Language pair** | | | | | **Translation** | | | | **Cross-lingual rewriting** | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Model | | Zh–En | Zh–Es | Zh–Fr | Zh–Ja | Zh–Ko | Faithful | Prof. | Style | Culture | Polish. | Format | Style | Creative |
| *General-purpose models* | | | | | | | | | | | | | | |
| GPT-5.6-Sol(max) | **63.7** | **69.4** | **57.7** | **72.2** | **64.0** | **55.1** | **69.3** | **66.2** | 70.6 | **66.6** | 61.7 | **60.2** | **62.9** | **55.6** |
| Qwen-3.8-Max-Thinking | 52.9 | 56.9 | 55.9 | 55.9 | 55.2 | 40.8 | 56.2 | 57.4 | **72.0** | 44.9 | **62.7** | 37.4 | 50.1 | 45.8 |
| GLM-5.3(max) | 52.6 | 58.9 | 54.6 | 53.0 | 55.1 | 41.7 | 56.1 | 52.7 | 55.2 | 55.4 | 56.7 | 45.4 | 49.0 | 51.5 |
| Kimi-K3(max) | 49.2 | 55.2 | 49.0 | 51.2 | 44.3 | 46.8 | 53.8 | 54.5 | 46.0 | 43.5 | 59.6 | 35.7 | 50.3 | 50.2 |
| Claude-Opus-5(max) | 47.7 | 52.9 | 49.9 | 51.1 | 49.2 | 35.5 | 58.1 | 57.2 | 57.9 | 39.9 | 49.7 | 43.6 | 38.0 | 41.3 |
| Grok-4.6(xhigh) | 47.4 | 51.1 | 53.1 | 45.4 | 43.6 | 44.0 | 51.4 | 46.4 | 43.7 | 36.2 | 57.6 | 49.2 | 48.7 | 46.8 |
| Doubao-Seed-2.1-Pro(high) | 47.1 | 50.0 | 46.6 | 55.2 | 47.9 | 35.8 | 58.4 | 44.3 | 63.0 | 40.2 | 53.8 | 40.1 | 46.6 | 37.1 |
| HY-3.0(high) | 45.9 | 54.6 | 51.0 | 45.0 | 38.6 | 40.6 | 51.3 | 57.7 | 52.7 | 29.3 | 44.2 | 46.1 | 38.5 | 48.4 |
| Muse-Spark-1.2(xhigh) | 43.8 | 47.8 | 48.6 | 42.1 | 41.9 | 39.0 | 51.5 | 51.0 | 47.8 | 35.4 | 50.6 | 35.2 | 40.1 | 41.2 |
| Gemini-3.7-Flash(high) | 40.6 | 44.9 | 49.5 | 41.4 | 31.2 | 36.1 | 49.7 | 49.8 | 55.4 | 47.5 | 47.4 | 27.2 | 29.4 | 24.2 |
| *Translation-specialized models* | | | | | | | | | | | | | | |
| Qwen-MT-Plus | 31.8 | 38.6 | 41.9 | 33.1 | 25.5 | 20.1 | 32.9 | 49.6 | 32.4 | 33.7 | 27.9 | 33.6 | 18.5 | 26.2 |
| HY-MT2-7B | 25.1 | 29.2 | 36.2 | 19.2 | 18.6 | 22.8 | 20.0 | 40.3 | 34.5 | 27.3 | 23.3 | 15.9 | 15.5 | 22.6 |
| HY-MT2-1.8B | 13.8 | 17.6 | 17.4 | 14.2 | 13.5 | 6.6 | 4.2 | 19.7 | 14.0 | 15.0 | 12.9 | 20.6 | 4.5 | 16.1 |

### 5.2 Effects of feedback-guided renewal

**Setup.** We compare the 47 questions removed during renewal with the 47 questions that replaced them, and then compare the three strategies that generated the replacements. *Static balancing* follows task and capability coverage targets without model feedback; *difficulty only* targets low development-round scores; and *full feedback* additionally uses diagnosed failures and coverage gaps. All comparisons use the same ten-model panel and GPT-5.5 Medium judging protocol. Renewal is adaptive by design, but its generation feedback did not come from this panel: a separate development round of 60 questions, answered by GPT-5.5, Kimi K2.6, and Doubao Seed 2.0 Pro, was summarized by language and task family without model identities, responses, Judge rationales, or rubric text. We report the mean score (lower is harder), the mean item SD across models (higher separates models more), and the number of all-zero questions, which contribute no separation.

**Renewal replaces saturated questions with much harder ones.** All ten panel models passed every MUST item of the removed questions in the selection-stage evaluation. Under the common GPT-5.5 Medium protocol, their replacements average 43.32, against 80.95 for the questions they replace. Because the removed questions were saturated by construction, the fairer baseline is static balancing, which generates questions without model feedback: its replacements average 51.25, close to the pre-renewal benchmark mean of 52.19, against 43.06 for full feedback. Over the full benchmark, the mean falls from 52.19 to 48.91: the replacements lower it by 3.54 points, and rubric revisions on retained questions offset 0.26 points (Appendix D.2).

Table 4: Replacement questions from three generation strategies.

| Strategy | $n$ | Mean score | Mean item SD ↑ | All-zero items ↓ |
|---|---|---|---|---|
| Static balancing | 15 | 51.25 | 26.54 | 2 |
| Difficulty only | 16 | 36.15 | 22.59 | 5 |
| Full feedback | 16 | 43.06 | **35.16** | **0** |

**Full-feedback replacements are the most discriminative.** Static balancing produces the easiest replacements, with a mean of 51.25. Difficulty-only generation produces the hardest, with a mean of 36.15, but five of its 16 questions receive zero from every model (Table 4). Full feedback lies between the two in difficulty, at 43.06, yet has the largest mean item SD, 35.16 against 22.59 for difficulty only, and no all-zero questions. Targeting the requirements on which models differ, rather than difficulty alone, is thus associated with questions that are both challenging and discriminative.The ordering does not hinge on the models that supplied feedback. On the seven panel models from families absent from the development round, static balancing, difficulty only, and full feedback average 50.05, 33.25, and 43.56, with mean item SDs of 22.28, 20.71, and 34.97; the three translation-specialized models, which took part in neither generation nor selection, also score lowest on full-feedback replacements (13.60, against 20.49 for difficulty only and 25.11 for static balancing; Table 21). Selection filled the

removed questions' language-pair slots without using panel scores, but allocation was not randomized and each group has only 15–16 questions, so these differences are descriptive (Appendix D).

### 5.3 Reusable Judge corrections

**Setup.** Qwen3.6-27B serves as both rule proposer and Judge with frozen parameters. The study uses 100 questions, 20 per language pair, each with six anonymized responses and blinded human reference labels. Four-fold cross-fitting induces rules from 60 questions and evaluates them on the remaining 20 in each fold, starting from an empty rule memory; rules learned from all 80 development questions are then evaluated on a separate 20-question block. We compare learned rule cards with direct judging and three rechecking controls: seed resampling, which re-asks without rules; a scoped placebo, which keeps each card's scope label but replaces its rule text with length-matched filler; and a blank placebo, which also blanks the label. All rechecking arms revisit the same scope-selected initial FAIL verdicts, and every arm, including direct judging, uses three-seed majority voting (Appendix E).

Table 5: Essential-item agreement after three-draw voting on the same human-labeled items. Net gains are relative to direct judging. Full correction counts appear in Appendix E.4.

| Judge arm | MUST errors ↓ | Error reduction (%) ↑ | Net correct ↑ |
|---|---|---|---|
| Direct | 307 | – | – |
| Cards | **291** | **5.21** | **+16** |
| Resampling | 301 | 1.95 | +6 |
| Scoped placebo | 299 | 2.61 | +8 |
| Blank placebo | 302 | 1.63 | +5 |

**Learned rules give the largest net correction.** Rule-guided rechecking reduces essential-item errors from 307 to 291, a 5.21% reduction relative to direct judging and the largest net gain among the evaluated arms, at $+16$ decisions (Table 5). The controls also improve on direct judging, by $+5$ to $+8$ decisions, so part of the gain comes from rechecking itself, and the margins of the cards over these controls are small (Appendix E.4).

Rechecking targets false rejections by design: cards reduce responses wrongly failed at the gate from 68 to 54 (55–58 for the controls). Response-level errors nevertheless stay at 150, because a repaired rejection can open the gate of a response whose genuine failure the Judge missed on another item (Appendix E.4). Under three-draw voting, cards also raise the Spearman correlation between Judge and human rankings of the six answer models from 0.371 to 0.600, whereas resampling (0.314) and the blank placebo (0.086) lower it; with six models and no gain under single-draw evaluation, we treat this ranking result as descriptive.

### 5.4 Analysis of corrected judgments

Tables 28 and 29 in Appendix E.6 trace two corrected judgments and their routed cards. In both cases, humans accepted a response that the initial Judge rejected; the routed cards came from other questions. In Correction A, the response spells out all five areas instead of a set phrase; the routed terminology card accepts a rendering that names every component the rubric requires, and rechecking turns a two-to-one FAIL vote into three PASS draws. In Correction B, a length cap leads the response to drop "last time" and the subject; no routed card concerns participant reference, but the closest boundary accepts omissions forced by length limits, and two of three rechecks agree with humans.

## 6 Discussion and Conclusion

We present RubricFly, the first LLM-driven translation evaluation framework that couples automated question and rubric generation with feedback-guided renewal and human-guided Judge rule learning. Shared testpoints turn an aggregate score into a diagnosis of what to test next and how to judge it. On 13 models, fly-corpus shows uneven progress that overall rankings hide, including a 16.77-point gap on style-adaptive translation between models 0.31 points apart overall. Renewal replaced saturated questions with harder, more discriminative ones when generation used model feedback rather than difficulty alone. Rules distilled from human corrections reduced essential-item errors, though only slightly more than rechecking without rules. These one-step studies leave untested whether the learned rules would change the benchmark Judge or the next renewal targets; future work will run both loops jointly with a shared Judge, over multiple rounds, and beyond Chinese-centered pairs.

### AI USE STATEMENT

Generative AI was used to produce candidate materials, testpoints, questions, rubrics, model responses, Judge rationales, and external calibration rules. It also assisted with hypothesis refinement, experimental design, implementation and debugging, data cleaning and reformatting, qualitative analysis, interpretation, literature discovery and summarization, scientific figure creation, and manuscript drafting, editing, and translation. The human authors defined the evaluation constructs and acceptance criteria and are responsible for verifying statistics, artifact provenance, citations, analysis code, and the final manuscript. Human reference labels anchor Judge calibration through independent blind annotation and expert review of disagreements. The authors take responsibility for the final content, claims, code, and released artifacts.

### ETHICS STATEMENT

Human reference judgments were collected with model identities and automatic verdicts hidden; the production audit showed verdicts to reviewers by design. Public artifacts will omit annotator identities, credentials, and private source material. Release preparation includes checking source permissions and reviewing examples for sensitive information and cultural stereotypes.

### REPRODUCIBILITY STATEMENT

Sections 3–5 specify the benchmark, construction and feedback procedures, cohorts, scoring, and comparisons; the appendix adds production and annotation details, induction settings, split results, paired analyses, and propagation traces. The anonymized supplementary package will bind each result to versioned data, prompts, splits, scoring code, and provenance manifests, excluding private source material, annotator identities, credentials, and non-releasable outputs.

## REFERENCES

Lakshya A Agrawal, Shangyin Tan, Dilara Soylu, Noah Ziems, Rishi Khare, Krista Opsahl-Ong, Arnav Singhvi, Herumb Shandilya, Michael J Ryan, Meng Jiang, Christopher Potts, Koushik Sen, Alex Dimakis, Ion Stoica, Dan Klein, Matei Zaharia, and Omar Khattab. Gepa: Reflective prompt evolution can outperform reinforcement learning. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 8479–8565, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/0e9e708b6f48e14fd0ac29e167413f76-Paper-Conference.pdf.

Rahul K. Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñonero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, et al. HealthBench: Evaluating large language models towards improved human health. *arXiv preprint arXiv:2505.08775*, 2025. URL https://arxiv.org/abs/2505.08775.

Mona Baker. *In Other Words: A Coursebook on Translation*. Routledge, 3rd edition, 2018.

J. C. Catford. *A Linguistic Theory of Translation: An Essay in Applied Linguistics*. Oxford University Press, London, 1965.

Jonathan Cook, Tim Rocktäschel, Jakob Foerster, Dennis Aumiller, and Alex Wang. TICKing all the boxes: Generated checklists improve LLM evaluation and generation. *arXiv preprint arXiv:2410.03608*, 2024. URL https://arxiv.org/abs/2410.03608.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474, 2021. doi: 10.1162/tacl_a_00437.

Xiyan Fu and Wei Liu. How reliable is multilingual LLM-as-a-judge? In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 11040–11053, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.587. URL https://aclanthology.org/2025.findings-emnlp.587/.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538, 2022. doi: 10.1162/tacl_a_00474. URL https://aclanthology.org/2022.tacl-1.30/.

Juliane House. *Translation Quality Assessment: Past and Present*. Routledge, 2015.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. Dynabench: Rethinking benchmarking in NLP. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4110–4124, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.324. URL https://aclanthology.org/2021.naacl-main.324/.

Ahrii Kim. RUBRIC-MQM : Span-level LLM-as-judge in machine translation for high-end models. In Georg Rehm and Yunyao Li (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pp. 147–165, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-288-6. doi: 10.18653/v1/2025.acl-industry.12. URL https://aclanthology.org/2025.acl-industry.12/.

Seungone Kim, Jay Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Ryan S. Shin, Sungdong Kim, James Thorne, and Minjoon Seo. Prometheus: Inducing fine-grained evaluation capability in language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/803485352e61e3ebf41221e4776c9fd4-Abstract-Conference.html.

Tom Kocmi and Christian Federmann. GEMBA-MQM: Detecting translation quality error spans with GPT-4. In Philipp Koehn, Barry Haddow, Tom Kocmi, and Christof Monz (eds.), *Proceedings of the Eighth Conference on Machine Translation*, pp. 768–775, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.wmt-1.64. URL https://aclanthology.org/2023.wmt-1.64/.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Philipp Koehn, Benjamin Marie, Christof Monz, Makoto Morishita, Kenton Murray, Masaaki Nagata, Toshiaki Nakazawa, Martin Popel, Maja Popović, Mariya Shmatova, and Jun Suzuki. Findings of the 2023 conference on machine translation (WMT23): LLMs are here but not quite there yet. In Philipp Koehn, Barry Haddow, Tom Kocmi, and Christof Monz (eds.), *Proceedings of the Eighth Conference on Machine Translation*, pp. 1–42, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.wmt-1.1. URL https://aclanthology.org/2023.wmt-1.1/.

Tom Kocmi, Ekaterina Artemova, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Howard Lakougna, Jessica Lundin, Christof Monz, Kenton Murray, Masaaki Nagata, Stefano Perrella, Lorenzo Proietti, Martin Popel, Maja Popović, Parker Riley, Mariya Shmatova, Steinthór Steingrímsson, Lisa Yankovskaya, and Vilém Zouhar. Findings of the WMT25 general machine translation shared task: Time to stop evaluating on easy test sets. In Barry Haddow, Tom Kocmi, Philipp Koehn, and Christof Monz (eds.), *Proceedings of the Tenth Conference on Machine Translation*, pp. 355–413, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-341-8. doi: 10.18653/v1/2025.wmt-1.22. URL https://aclanthology.org/2025.wmt-1.22/.

André Lefevere. *Translation, Rewriting, and the Manipulation of Literary Fame*. Routledge, London, 1992.

Tianle Li, Wei-Lin Chiang, Evan Frick, Lisa Dunlap, Tianhao Wu, Banghua Zhu, Joseph E. Gonzalez, and Ion Stoica. From crowdsourced data to high-quality benchmarks: Arena-Hard and Bench-Builder pipeline. *arXiv preprint arXiv:2406.11939*, 2024. URL https://arxiv.org/abs/2406.11939.

Xiang Lisa Li, Farzaan Kaiyom, Evan Zheran Liu, Yifan Mai, Percy Liang, and Tatsunori Hashimoto. AutoBencher: Towards declarative benchmark construction. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/eb216114f3eaad22506fd1bc7bbff0ca-Abstract-Conference.html.

Yunzhe Li, Richie Yueqi Feng, Tianxin Wei, and Chin-Chia Hsu. CoReflect: A reflective co-evolution framework for improving conversational evaluation. *arXiv preprint arXiv:2601.12208*, 2026a. URL https://arxiv.org/abs/2601.12208v2. Version 2, August 31, 2026; accepted to Findings of EMNLP 2026.

Zheng Li, Mao Zheng, Mingyang Song, and Tianxiang Fei. HardMTBench: Stress-testing Chinese–English translation on knowledge-intensive domains. *arXiv preprint arXiv:2605.28315*, 2026b. URL https://arxiv.org/abs/2605.28315.

Bill Yuchen Lin, Yuntian Deng, Khyathi Chandu, Faeze Brahman, Abhilasha Ravichander, Valentina Pyatkin, Nouha Dziri, Ronan Le Bras, et al. WildBench: Benchmarking LLMs with challenging tasks from real users in the wild. *arXiv preprint arXiv:2406.04770*, 2024. URL https://arxiv.org/abs/2406.04770.

Yihang Lin, Yunze Gao, Zeyang Lin, Dongbo Li, Kun Peng, and Yue Liu. GrowLoop: Self-evolving conversation evaluation seeded by human. *arXiv preprint arXiv:2605.28882*, 2026. URL https://arxiv.org/abs/2605.28882v2. Version 2, June 10, 2026.

Peter Newmark. *A Textbook of Translation*. Prentice-Hall International, New York, 1988.

Eugene A. Nida. *Toward a Science of Translating: With Special Reference to Principles and Procedures Involved in Bible Translating*. E. J. Brill, Leiden, 1964.

Christiane Nord. *Translating as a Purposeful Activity: Functionalist Approaches Explained*. Routledge, London, 2nd edition, 2018. doi: 10.4324/9781351189354.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. LLM evaluators recognize and favor their own generations. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/7f1f0218e45f5414c79c0679633e47bc-Abstract-Conference.html.

Yiwei Qin, Kaiqiang Song, Yebowen Hu, Wenlin Yao, Sangwoo Cho, Xiaoyang Wang, Xuansheng Wu, Fei Liu, et al. InFoBench: Evaluating instruction following ability in large language models. *arXiv preprint arXiv:2401.03601*, 2024. URL https://arxiv.org/abs/2401.03601.

Katharina Reiss and Hans J. Vermeer. *Towards a General Theory of Translational Action: Skopos Theory Explained*. Routledge, London, 2014. doi: 10.4324/9781315759715.

Shreya Shankar, J. D. Zamfirescu-Pereira, Björn Hartmann, Aditya G. Parameswaran, and Ian Arawjo. Who validates the validators? aligning LLM-assisted evaluation of LLM outputs with human preferences. *arXiv preprint arXiv:2404.12272*, 2024. URL https://arxiv.org/abs/2404.12272.

Mingrui Sun, Mao Zheng, Zheng Li, and Mingyang Song. IFMTBench: A comprehensive benchmark for multilingual translation instruction following. *arXiv preprint arXiv:2605.28218*, 2026. URL https://arxiv.org/abs/2605.28218.

The Mosaic Research Team. MemAlign: Building better LLM judges from human feedback with scalable memory. MLflow research blog, February 3, 2026. URL https://mlflow.org/blog/memalign.

Gideon Toury. *Descriptive Translation Studies and Beyond*. John Benjamins, Amsterdam, 1995.

Lawrence Venuti. *The Translator's Invisibility: A History of Translation*. Routledge, 1995.

Jean-Paul Vinay and Jean Darbelnet. *Comparative Stylistics of French and English: A Methodology for Translation*, volume 11 of *Benjamins Translation Library*. John Benjamins, Amsterdam, 1995. doi: 10.1075/btl.11.

Lipeng Xie, Sen Huang, Zhuo Zhang, Anni Zou, Yunpeng Zhai, Dingchao Ren, Kezun Zhang, Haoyuan Hu, Boyin Liu, Haoran Chen, Zhaoyang Liu, and Bolin Ding. Auto-rubric: Learning from implicit weights to explicit rubrics for reward modeling, 2026. URL https://arxiv.org/abs/2510.17314.

Qizheng Zhang, Changran Hu, Shubhangi Upasani, Boyuan Ma, Fenglu Hong, Vamsidhar Kamanuru, Jay Rainton, Chen Wu, Mengmeng Ji, Hanchen Li, Urmish Thakker, James Y Zou, and Kunle Olukotun. Agentic context engineering: Evolving contexts for self-improving language models. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 86069–86100, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/8a94ff6f922d995d7d3f4ebf4143e442-Paper-Conference.pdf.

Kaijie Zhu, Jiaao Chen, Jindong Wang, Neil Zhenqiang Gong, Diyi Yang, and Xing Xie. Dyval: Dynamic evaluation of large language models for reasoning tasks, 2024. URL https://arxiv.org/abs/2309.17167.

Vilém Zouhar, Wenda Xu, Parker Riley, Juraj Juraska, Mara Finkelstein, Markus Freitag, and Daniel Deutsch. Generating difficult-to-translate texts. In Pinzhen Chen, Vilém Zouhar, Hanxu Hu, Simran Khanuja, Wenhao Zhu, Barry Haddow, Alexandra Birch, Alham Fikri Aji, Rico Sennrich, and Sara Hooker (eds.), *Proceedings of the First Workshop on Multilingual Multicultural Evaluation*, pp. 204–219, Rabat, Morocco, March 2026a. Association for Computational Linguistics. ISBN 979-8-89176-368-5. doi: 10.18653/v1/2026.mme-main.14. URL https://aclanthology.org/2026.mme-main.14/.

Vilém Zouhar, Niyati Bafna, Mukund Choudhary, Maike Züfle, Sara Rajaee, Pinzhen Chen, Jannis Vamvas, Sara Papi, Ona de Gibert, Bhavitvya Malik, Eliya Habba, Orfeas Menis Mastromichalakis, Patrícia Schmidtová, Michelle Wastl, Sheriff Issaka, Leshem Choshen, Stella Biderman, Antonis Anastasopoulos, Jan Niehues, Rico Sennrich, Mrinmaya Sachan, Ondřej Bojar, Kenton Murray, Jörg Tiedemann, Alham Fikri Aji, Philipp Koehn, Christof Monz, Alexandra Birch, Sowmya Vajjala, Chalamalasetti Kranti, Cristina España-Bonet, Nobin Sarwar, David Kaczér, Shunta Asano, Malik Marmonier, Daban Q. Jaff, Vaisakhi Mishra, Hend Al-Khalifa, Gabriele Sarti, Sourajit Saha, Nils Rehlinger, Juan Daniel Cuervo Villa, Jonathan Tonglet, Saugata Purkayastha, Dominik Macháček, Jagannathan Ramanujam, Heejin Do, Zuzana Nadova, Fred Philippy, Fabian Retkowski, Maria Lymperaiou, Silvia Casola, Hanna Yukhymenko, Shubhashis Roy Dipta, Sangwon Ryu, Andrés Jerez, Ron Keinan, Shuaib Shuaib Yusuf, Avantica Vempati, Maria Carmen Staiano, Sukannya Purkayastha, Adrian Cosma, Vitalii Babenko, Erivan Inan, Aviral Nigam, Wafa Aissa, Fatima Haouari, Venkata Prasanth Kumar Gummadi, Mehdi Jafarzadeh, Valentin Scourneau, Lukas Edman, Kaiser Sun, Shaomu Tan, Mohammad Sadegh Gholizadeh, Johannes-Rudolf David, Dipankar Srirag, Javier García Gilabert, Ruta Binkyte, Manar Ali, Ana-Maria Bucur, Sabry E. Farrag, Youssef Saber, Yihong Liu, Jean Maillard, Cojocaru Nicoleta, Xiaochuang Yuan, Sina Ahmadi, Philipp Mondorf, Kaustubh Dhole, Roman Wixinger, Shenbin Qian, Manuel Tuor, Sergey Troshin, Jonathan Yahav, Fida Mohammad Thoker, Amir Arsalan Rezapour, Lance Calvin Lim Gamboa, Manon Reusens, Kätriin Kukk, Koel Dutta Chowdhury, Giuseppe Gallipoli, Christian Hoang, Shaswati Saha, Seth Aycock, Jan Kocoń, Bo Chen, Linh Vu, Vatsal Venkatkrishna, Arafat Ahsan, Luan Thanh Nguyen, Hassan Soliman, Daryna Dementieva, Theresia Veronika Rampisela, Ngoc Quynh Tram Do, Marius Huber, Kazuki Egashira, Azmine Toushik Wasi, Vladislav Poritski, Mike Zhang, Deep Shah, Paul Gavrikov, Luis Frentzen Salim, David Africa, R. Damanhuri, Bello Umar Bello, Anumit Garg, Gengyu Rao, Pawan Sasanka Ammanamanchi, Kamile Dementaviciute, Andrianos Michail, L D M S Sai Teja, Dawei Zhu, Yi Fan, Wei Liu, Farhan Farsi, Elias Herranen, Sankalan Pal Chowdhury, Karen Sanchez, Farzad Shami, Ashok Urlana, Zimu Wang, Tomasz Limisiewicz, Priyaranjan Pattnayak, Marii Ojastu, Hongbin Na, Emilian Radoi, Chenyi Zhao, Carlos Hinojosa, Andrea Gregor de Varda, Zaid Alyafeai, Reem Alzahrani, Nehal Kathrotia, Alex Flückiger, Ulysses Sekai Tully Carr, Jimson Paulo Layacan, Guy Kaplan, Ritwik Tiwari, Rishit Dagli, Oksana Volchek, Isaac R Caswell, Bowen Yi, Blanka Kövér, Amir Hossein Yari, Aicha Chorana, Zhengxiang Wang, Selja Keränen, Samuel Simko, Joy Olusanya, Jenny

Chim, Enzo Doyen, Vivek Harsha Lakkamaneni, Sophia Conrad, Pouya Sadeghi, Panayiotis Panayiotou, Luis Lara, Jannatul Nayem, Eran Yahav, Debanshu Das, Antonia Karamolegkou, Anmol Goel, Aishik Mandal, Tommaso Cerruti, Raoyuan Zhao, Mykola Haltiuk, Thura Aung, Naser Almousa, Amir Hossein Kargaran, Rachel Bawden, Qiaoyuan Zheng, Mateusz Lango, Beni Egressy, Fidel Rodríguez Velásquez, Natchapon Jongwiriyanurak, Minh Ngoc Do, Marco Gaido, Lena Libon, Dzmitry Kuzmin, Badal Nyalang, Antoine Taroni, Andrei Niculae, Abdulaziz Nura Kani, Rushikesh Zawar, Marek Šuppa, Beatrice Savoldi, Andreas Simons, Rayyan Merchant, Ilai Yaron Levy, Francesco Pinto, Ziyi Yang, Yolanda Xavier, Samuel Frontull, Muhammad Ravi Shulthan Habibi, Kenneth Enevoldsen, Harris Abdul Majid, Francesca Padovani, Tim Graf, Tatiana Bielakova, Sharifa Djurabaeva, Shaoxiong Ji, Raia Abu Ahmad, Pavel Stepachev, Jirui Qi, Ayush Sunil Munot, Alireza Pakniat, Ayla Rigouts Terryn, Yuxing Lu, Yurii Paniv, Xiyan Fu, Tosin Adewumi, Sunisth Kumar, Stéphane J. P. S. Thunus, Shree Harsha Bokkahalli Satish, Shayan Bali, Prakhar Gupta, Papa Abdou Karim Karou Diallo, Matija Akrap, Marko Culjak, Kristýna Onderková, Joseph Attieh, Esrael Teferi Tensay, Elisabeth Fittschen, Benoît Sagot, Jingwei Ni, and Yu Fan. Last translation benchmark, 2026b. URL https://arxiv.org/abs/2609.04173.

# A Benchmark Design and Construction

## A.1 Tasks and scenarios

RubricFly separates what a response must deliver from the abilities needed to produce it. Task families organize eight translation and rewriting objectives, while 26 generation routes specify forms such as legal translation, compression, and audience adaptation (Table 6). Material attributes describe domain, text function, register, and discourse structure; competence and phenomenon labels specify the ability and linguistic challenge. Their combinations support targeted coverage: a legal notice and an everyday dialogue can test the same discourse ability, while two business texts can test numerical fidelity and audience adaptation separately.

**Why separate axes.** An earlier version of the taxonomy organized questions only by generation route, which fused domain, task form, and ability: legal texts and strategic decisions belonged to different routes, so results could not show how well a model makes strategic decisions in legal texts. Separating the axes makes such combinations addressable. The task families are ordered by how far a rendering may depart from its source, following functionalist translation theory (Reiss & Vermeer, 2014; Nord, 2018): faithful and professional translation stay close to the source, whereas style-adaptive translation and cultural adaptation adjust it to the target reader and culture. The rewriting families are distinguished by what the output changes: its quality (polishing, bilingual review, variant selection), form (compression, expansion, genre), style (formality, audience, tone), or content (copywriting, scenario transfer, multi-source integration). The text-function attribute follows Reiss's distinction between informative, expressive, and operative texts (Reiss & Vermeer, 2014).

Table 6: Task families, generation routes, and question counts in fly-corpus.

| Task family | $n$ | Generation routes |
|---|---|---|
| T1 Faithful translation | 45 | General texts; spoken dialogue; long-form discourse |
| T2 Professional translation | 67 | Legal; medical/life sciences; technology/IT; business/finance |
| T3 Style-adaptive translation | 57 | Register shifts; idioms/proverbs; literary/creative translation |
| T4 Cultural adaptation | 64 | Cultural concepts; metaphor/pun/humor; transcreation; localization |
| R1 Cross-lingual polishing | 65 | Basic polishing; bilingual review; variant selection/fusion |
| R2 Cross-lingual format rewriting | 62 | Compression; expansion; genre transformation |
| R3 Cross-lingual style rewriting | 64 | Formality shifts; audience adaptation; tone/emotion |
| R4 Cross-lingual creative rewriting | 76 | Copywriting; cross-cultural scenario transfer; multi-source integration |

## A.2 Competence targets

We operationalize translation theory as nine competence families, each with three subcompetencies (Table 7). Each family corresponds to a concept from translation studies. Semantic fidelity (B1) operationalizes equivalence (Nida, 1964); structural transfer (B2) operationalizes translation shifts such as transposition and modulation (Catford, 1965; Vinay & Darbelnet, 1995); register command (B3) follows register analysis in translation quality assessment (House, 2015). Cultural mediation (B4) concerns culture-specific meaning (Nida, 1964; Newmark, 1988), and strategic decision (B5) the choice between domesticating and foreignizing strategies (Venuti, 1995). Purpose alignment (B6) follows skopos theory (Reiss & Vermeer, 2014; Nord, 2018); creative recreation (B7) treats literary translation as rewriting (Lefevere, 1992); discourse management (B8) concerns thematic and information structure (Baker, 2018); and norm compliance (B9) concerns translation norms (Toury, 1995).

These author-defined operational categories support multi-label diagnosis. A medical explanation, for example, can preserve numerical facts, maintain an appropriate register, and adapt its presentation for patients; these correspond to different requirements within one task.

**Link to rubric dimensions.** Competence labels also steer rubric construction. Each competence has a primary and a secondary rubric dimension (Table 7). In initial construction, the rubric prompt received these dimensions for the question's competence labels, together with the dimensions emphasized for its route, and asked that the primary dimension account for at least 40% of the items. The prompt thus directs a semantic-fidelity question toward content correctness and a creative-recreation question toward expression and format.

Table 7: Translation competencies, their second-level subcompetencies, the concepts they operationalize, and their rubric focus (primary; secondary dimension).

| ID | Family | Second-level subcompetencies | Concept | Rubric focus |
|---|---|---|---|---|
| B1 | Semantic fidelity | Terminological precision; numerical precision; logical fidelity | Equivalence (Nida, 1964) | Content correctness; completeness |
| B2 | Structural transfer | Syntactic restructuring; part-of-speech shift; viewpoint/modulation shift | Translation shifts (Catford, 1965; Vinay & Darbelnet, 1995) | Expression & format; content correctness |
| B3 | Register command | Register recognition; register maintenance; register shift | Register analysis (House, 2015) | Expression & format; relevance |
| B4 | Cultural mediation | Cultural-concept transfer; rhetorical-effect transfer; cultural sensitivity | Cultural equivalence (Nida, 1964; Newmark, 1988) | Relevance; content correctness |
| B5 | Strategic decision | Domestication/foreignization; fidelity/fluency balance; literal/free-translation choice | Domestication and foreignization (Venuti, 1995) | Relevance; expression & format |
| B6 | Purpose alignment | Audience awareness; functional matching; constraint compliance | Skopos theory (Reiss & Vermeer, 2014; Nord, 2018) | Instruction following; relevance |
| B7 | Creative recreation | Style recreation; imagery transfer; aesthetic creation | Translation as rewriting (Lefevere, 1992) | Expression & format; relevance |
| B8 | Discourse management | Cohesion reconstruction; information-flow management; cross-paragraph consistency | Thematic and information structure (Baker, 2018) | Consistency; completeness |
| B9 | Norm compliance | Format conventions; terminology conventions; genre conventions | Translation norms (Toury, 1995) | Expression & format; consistency |

### A.3 TESTPOINT REPRESENTATION

A testpoint connects a source phenomenon to an intended operation and observable success conditions. It specifies task and language targets, competence and phenomenon labels, material attributes, a task contract, source material, and illustrative errors. The task contract separates core information from adaptable presentation details; question construction adds final output constraints.

Table 8 presents the released Chinese-to-Spanish notice used in Figure 2. For this route, development feedback identifies consistency and content failures alongside gaps in R1a, B2a, and SYN-01 coverage. The task requires both recovering the issuing authority as agent and reconciling permission in designated areas with prohibition elsewhere.

Table 8: Structured presentation of the administrative-notice testpoint and resulting question. Field names are standardized and grouped for exposition.

| Field | Example value |
|---|---|
| Task and identity | Cross-lingual polishing, route R1a; Chinese → Spanish; testpoint `R1a_tp01`; released question `CLDEV_ab120_002_f`. |
| Competence and phenomenon | B2a (syntactic restructuring); SYN-01 (omitted subjects). Assessment focus: logical consistency and recovery of the acting subject. |
| Material attributes | Legal domain (`D-legal`); informative function (`TF-info`); formal register (`R-formal`); simple discourse structure (`DS-simple`). Source identifier: `zh_src_001`; complete passage below. |
| Task contract | Deliverable: final translation. Audience: the general public. Style goal: a formal, clear Spanish administrative notice with consistent permissions and explicit agents. |
| Core information | Regulation of trading on Chengdong District's main and secondary roads; orderly trading permitted in designated areas along those roads; trading prohibited outside those areas; lawful penalties for failure to rectify violations within the required period; effect from the publication date. |
| Adaptable details | Vague time wording such as "recently" may be retained or appropriately specified; the title's wording and format may follow Spanish notice conventions. |
| Intended operation | Restore omitted agents and resolve the permission conflict so that permitted and prohibited areas are mutually exclusive. |
| Illustrative errors | "Se permite operar en los puntos designados y se prohíbe cualquier operación en todas las áreas" permits and prohibits activity in the same area. "Los comerciantes han decidido realizar una rectificación especial" incorrectly makes traders the authority ordering enforcement. |
| Resulting question | Formal Spanish administrative notice; general-public audience; explicit agents and unambiguous wording; 150–220 words; final text only, with one version and no explanation. |

**Complete source passage.**

关于城东区主次干道经营秩序整治的通告

> 近期，经查发现，部分商户擅自占用主次干道经营，已严重影响通行。现决定开展专项整治，责令相关商户限期整改。逾期不整改的，将依法予以处罚。同时，允许在主次干道两侧划定的疏导点内有序经营，严禁在主次干道所有区域开展任何形式的经营活动。本通告自发布之日起施行。

The resulting MUST items separately assess permission boundaries, the authority's agency, and unsupported additions. This separates a targeted linguistic challenge from other requirements that can also determine the response score.

### A.4 MATERIAL PREPARATION, CONSTRUCTION, AND QUALITY CONTROL

Material preparation either retrieves background information and source snippets for model synthesis or generates material directly. A source passage supplies the question's factual content; retrieved links identify external material used in its preparation. The construction model converts each testpoint into a question, then writes rubric items against that question. Table 9 summarizes the construction model assignments and the study's quality checks. GPT-5.5 Medium evaluates released responses one rubric item at a time, receiving the complete question and response and returning a verdict with a rationale.

Table 9: Construction roles and quality checks for fly-corpus.

| Component | Model assignment or procedure |
|---|---|
| Initial candidate generation | DeepSeek V4 Pro plans research topics and queries, judges evidence sufficiency, and consolidates material retrieved by Doubao Seed 2.0 Pro (high, search-enabled); in the synthesis mode, material is generated without retrieval (Appendix A.6); DeepSeek V4 Pro writes testpoints; DeepSeek V4 Pro or GPT-5.5 constructs questions and rubrics. |
| 47-replacement generation | Gemini 3.1 Pro plans strategies and queries; Doubao Seed 2.0 Pro (high, search-enabled) retrieves material; DeepSeek V4 Pro synthesizes material and constructs testpoints, questions, and rubrics. |
| Expert production audit | Experts reviewed rubric text, dimensions, roles, and visible Judge decisions in an earlier 500-question snapshot: 8,683 rubric-item records (Section B.2). |

### A.5 RUBRIC DIMENSIONS, ROLES, AND WEIGHTS

A rubric item specifies one judgeable requirement, its dimension, aggregation role, and weight. Its dimension identifies the property assessed; its role determines whether it is an essential requirement (MUST) or additional quality (PLUS). Table 10 reports their realized combinations.

**Role assignment.** Roles follow a minimum-requirement rule shared by the rubric-construction prompts of the initial and replacement questions. A MUST item states a requirement that a usable response cannot miss. Instruction following covers the task, the target language, and every hard constraint stated in the question; content correctness covers facts, terms, numbers, dates, agents, objects, conditions, negation, causality, and legal or technical consequences; completeness covers the information a response may not drop; and safety and compliance excludes fabricated or unsupported additions, altered responsibilities or risk boundaries, and harmful content. Items in these four dimensions are therefore MUST by default: they account for 836 of the 842 instruction-following items, 1,837 of the 2,020 content-correctness items, 842 of the 1,089 completeness items, and 214 of the 230 safety-and-compliance items (Table 10). Following the construction prompts, details that the genre or length allows a response to compress, and refinements beyond the minimum, are PLUS even within these dimensions. Expression and format, relevance, and consistency usually describe quality beyond the minimum, such as idiomatic word choice, natural syntax, or register refinement, and are PLUS; ordinary layout, acceptable synonyms, and minor wording or consistency preferences do not justify a MUST item. When the question prescribes a format, genre, register, or rhetorical effect, however, the corresponding items become MUST. The requested administrative-notice format yields the MUST item R2 in Figure 2, and the instruction to preserve the force of the sports metaphors yields the MUST items R6–R8 in Appendix A.7.1. When the requested genre, such as a formal email or a patient notice, is central to the task, genre fit is MUST if a mismatch would make the deliverable unusable and PLUS if it only affects quality. Consistency items are MUST only when an inconsistency would change an agent, concept, obligation, condition, or deadline. Across the release, 548 of the 3,312 expression-and-format items, 137 of the 709 relevance items, and 108 of the 441 consistency items are MUST. The annotation guidelines draw the same line: a MUST item marks the minimum

for core meaning, hard instructions, and a usable deliverable; a PLUS item separates correct from excellent responses; and formatting counts only where the question requires it (Appendix B.1).

The default weights are 1 for MUST and 0.5 for PLUS. Equal numbers of the two roles give PLUS one third of the weighted credit. Strictly Hybrid scoring additionally requires every MUST item to pass (Equation 1). Production scoring uses the stored weights after revisions; the Judge study fixes role-based weights across its arms.

Table 10: The seven rubric dimensions and their aggregation-role counts across 8,643 rubric items in fly-corpus. Each item has one dimension and one role.

| Dimension | Operational scope | MUST | PLUS |
|---|---|---|---|
| Instruction following | Explicit task, language, length, format, and output constraints | 836 | 6 |
| Content correctness | Factual, semantic, numerical, terminological, and logical accuracy | 1,837 | 183 |
| Completeness | Preservation or realization of required information and subparts | 842 | 247 |
| Expression & format | Target-language naturalness, genre conventions, readability, and formatting | 548 | 2,764 |
| Relevance | Fit to the requested purpose, audience, situation, and communicative function | 137 | 572 |
| Consistency | Terminology, reference, data, register, and style across the response | 108 | 333 |
| Safety & compliance | Unsupported high-risk additions, privacy, stereotyping, and domain-specific boundaries | 214 | 16 |
| Total | | 4,522 | 4,121 |

### A.6 DATASET SELECTION AND COMPOSITION

The candidate snapshot contains 1,000 questions, with 200 per language pair. Language-specific selection formed the initial 500-question benchmark. English and French selection ranked task and linguistic complexity under coverage constraints; Japanese and Korean selection prioritized cross-model score differences and diverse errors. Renewal replaced 47 questions on which every panel model passed all MUST items. The question sets are compared using a common GPT-5.5 Medium protocol (Appendix D.2). The release combines 453 retained questions, 39 selected from an evaluated 60-question pilot, and eight from a nine-question supplement. The strategy comparison is described in Appendix D.

The release contains 233 translation and 267 rewriting questions, with 100 per language pair (Tables 6 and 12). All language and task-family labels are verified; detailed route, competence, and phenomenon mappings are verified for the 47 replacements. Its 8,643 rubric items comprise 4,522 MUST and 4,121 PLUS items (Table 13).

**Material provenance.** Candidate materials were prepared in one of two modes. In the retrieval mode, a search-enabled model retrieved background information and source snippets, a construction model wrote the source material from them, and the retrieved links were kept as provenance. In the synthesis mode, a model generated the material directly without retrieval; this mode produced the initial Zh–En pool and part of the Zh–Fr pool. All 47 replacements use the retrieval mode. As a result, 398 of the 500 released questions (79.6%) rest on retrieved material (Table 11).

Table 11: Source-material mode of the released questions by language pair. Retained questions come from the initial benchmark; replacements come from feedback-guided renewal.

| Material mode | Zh–En | Zh–Es | Zh–Fr | Zh–Ja | Zh–Ko | Total |
|---|---|---|---|---|---|---|
| Retrieval, retained | 0 | 91 | 66 | 94 | 100 | 351 |
| Retrieval, replacement | 18 | 9 | 14 | 6 | 0 | 47 |
| Synthesis, retained | 82 | 0 | 20 | 0 | 0 | 102 |
| Total | 100 | 100 | 100 | 100 | 100 | 500 |

Table 12: Realized task-family composition of fly-corpus. Language pairs are bidirectional.

| Task family | Zh–En | Zh–Es | Zh–Fr | Zh–Ja | Zh–Ko | Total |
|---|---|---|---|---|---|---|
| Faithful translation | 12 | 12 | 8 | 8 | 5 | 45 |
| Professional translation | 14 | 17 | 16 | 8 | 12 | 67 |
| Style-adaptive translation | 9 | 11 | 10 | 14 | 13 | 57 |
| Cultural adaptation | 13 | 10 | 21 | 11 | 9 | 64 |
| Cross-lingual polishing | 12 | 19 | 6 | 12 | 16 | 65 |
| Cross-lingual format rewriting | 13 | 10 | 11 | 17 | 11 | 62 |
| Cross-lingual style rewriting | 16 | 2 | 16 | 16 | 14 | 64 |
| Cross-lingual creative rewriting | 11 | 19 | 12 | 14 | 20 | 76 |
| Total | 100 | 100 | 100 | 100 | 100 | 500 |

Table 13: Rubric-item counts across the 500 questions.

| Role | Total rubric items | Mean/question | Median | Min | Max |
|---|---|---|---|---|---|
| MUST | 4,522 | 9.04 | 9 | 5 | 16 |
| PLUS | 4,121 | 8.24 | 8 | 3 | 11 |
| All | 8,643 | 17.29 | 18 | 12 | 20 |

## A.7 Worked examples

### A.7.1 A complete rubric and score calculation

The following English-to-Chinese rewriting example includes the task, complete source, two model responses, and all 15 rubric items. Task wording and rubric items are rendered in English; verdicts are the stored benchmark Judge decisions.

**Question.** Rewrite the following English political speech in Chinese. Preserve its argumentative logic, make academic language accessible to general readers, and use no more than three sentences per paragraph. Preserve the rhetorical force of the sports metaphors while making them understandable to Chinese readers; explanations or notes may be incorporated where necessary. Output only the final translation, without commentary or multiple versions.

> When I stepped into the Oval Office, the trade war with China was a fourth-and-goal situation, and the previous administration just punted. We're not here to play prevent defense—we're going to march down the field, score a touchdown for American workers, and make sure the other side knows they can't run out the clock on us.

**Response A: Muse-Spark-1.2 (xhigh).**

> 当我走进白宫椭圆形办公室时，对华贸易战已经到了美式橄榄球中“四档攻门”[最后一次进攻机会，站在对方球门前，成败在此一举]的绝境，而上届政府却直接选择了弃踢，把球权拱手让人。我们不是来打保守的“防反战术”的，我们要全线向前推进，为美国工人赢得一次达阵得分，并让对手明白，别想靠拖延时间把我们耗死。

*English gloss:* When I entered the Oval Office, the trade war with China had reached the desperate American-football situation of “fourth-down attack on goal” [the last attacking chance, in front of the opponent's goal, with success or failure at stake], yet the previous administration simply punted and surrendered possession. We are not here for conservative “defensive counterattack” tactics: we will advance on all fronts, win a touchdown for American workers, and tell our opponents not to think they can wear us down by stalling.

**Response B: Gemini-3.7-Flash (high).**

> 当我入主白宫时，对华贸易战正如橄榄球比赛中只差临门一脚的关键决胜时刻，然而上届政府却选择了消极弃踢、放弃进攻。我们绝不是来采取消极防守策略的——我们要全线推进，为美国工人赢得决定性的达阵得分！我们更要让对手清楚地知道，他们休想靠拖延战术来耗光时间。

*English gloss:* When I entered the White House, the trade war with China was like a decisive football moment just one final kick from the goal, yet the previous administration chose to punt passively and abandon the attack. We are certainly not here to defend passively: we will advance on all fronts and score a decisive touchdown for American workers! We will make clear that our opponents cannot exhaust the time through delaying tactics.

**Interpreting dimensions and roles.** Expression and format covers both essential metaphor requirements (R6–R8) and additional refinements (R9–R11 and R13–R15). For A, the Judge rejects only R12 because explanation strategies are mixed. For B, it rejects only R6: that phrase lacks the fourth-down or must-score-a-touchdown image. B's later touchdown phrase translates a different source clause. Table 14 includes every verdict used below.

Table 14: Complete rubric and stored verdicts for the example. IF: instruction following; CC: content correctness; EF: expression and format; CO: consistency. Verdicts are 1 (satisfied) or 0 (not satisfied). A and B refer to the full responses reproduced in this section.

| ID | Rubric item (English rendering; acceptance examples condensed) | Dim. | $w_i$ | A | B |
|---|---|---|---|---|---|
| **Essential requirements (MUST)** | | | | | |
| R1 | Output only the final Chinese translation: no introduction, translation commentary, version comparisons, terminology section, or revision notes. Necessary explanations or footnotes within the translation are allowed. | IF | 1 | 1 | 1 |
| R2 | Preserve the time frame of taking office or entering the White House/Oval Office; do not omit it or change it to campaigning or leaving office. | CC | 1 | 1 | 1 |
| R3 | Identify China as the trade-war counterpart; do not omit it or replace it with an unspecified or broader group of countries. | CC | 1 | 1 | 1 |
| R4 | Convey that the previous administration gave up or ceded initiative at the critical point, rather than actively confronting the situation or merely doing nothing. | CC | 1 | 1 | 1 |
| R5 | Preserve the full argument: this administration attacks rather than defends, seeks a decisive victory for American workers, and prevents the other side from winning by delay. Retain the beneficiary and agency. | CC | 1 | 1 | 1 |
| R6 | Render "fourth-and-goal" with the urgency of a fourth-down or must-score-a-touchdown football situation. A generic decisive moment without that image does not suffice. | EF | 1 | 1 | **0** |
| R7 | Render "punt" with the image of punting or surrendering possession; generic avoidance or compromise without that action is insufficient. | EF | 1 | 1 | 1 |
| R8 | Render "run out the clock" with a concrete sense of exhausting playing time, rather than a generic delaying tactic without time-consumption imagery. | EF | 1 | 1 | 1 |
| **Additional quality (PLUS)** | | | | | |
| R9 | Render "prevent defense" with a defensive sports image or explanation, not merely a conservative strategy unrelated to defense. | EF | 0.5 | 1 | 1 |
| R10 | Render "march down the field" with progression along the football field, rather than generic forward movement without that scene. | EF | 0.5 | 1 | 1 |
| R11 | Render "score a touchdown" with the touchdown action, not only winning a game. | EF | 0.5 | 1 | 1 |
| R12 | Use a consistent explanation strategy for the football metaphors: bracketed/footnote explanations or contextual integration. Do not disrupt their unity by mixing these approaches, such as a footnote at one point and a paraphrase without a sports reference at another. | CO | 0.5 | **0** | 1 |
| R13 | Reproduce the mobilizing tone of a political speech with forceful parallelism or contrast, rather than flat news-style reporting or a list of facts. | EF | 0.5 | 1 | 1 |
| R14 | Explicitly connect the contrast between the predecessor's retreat and the current administration's counterattack through connective words or punctuation, rather than loosely juxtaposing isolated sentences. | EF | 0.5 | 1 | 1 |
| R15 | Restructure long English sentences into natural Chinese clause order; avoid word-for-word English syntax or awkwardly stacked modifiers. | EF | 0.5 | 1 | 1 |

**Score calculation.** The total weight is $8 \times 1 + 7 \times 0.5 = 11.5$. A passes eight MUST and six PLUS items:

$$100S(A) = 100\frac{8 \times 1 + 6 \times 0.5}{11.5} = 95.65.$$

B passes seven MUST and seven PLUS items, yielding $(7 + 3.5)/11.5 = 91.30\%$ before gating. Its failed MUST item R6 makes $100S(B) = 0$. Both responses pass 14 of 15 items; the failed item's role determines the difference.

### A.7.2 AUDIENCE ADAPTATION: INTERNAL MESSAGE TO EXTERNAL EMAIL

This example changes the recipient and communicative purpose while preserving transaction conditions. The task and complete Chinese source are presented in English; selected rubric items retain their identifiers, roles, and weights.

Item `en_026` contains 19 rubric items (10 MUST, nine PLUS). *Task:* A colleague has sent the following WeChat message. Rewrite it as a formal English email to the counterparty's project lead for an upcoming conference call. Use a professional tone, precise terminology, and standard English business-email conventions. Output only the final email.

> Lao Wang, have you seen Shenyun Data's Q3 results? Revenue grew 15% year on year to RMB 230 million, but operating margin dropped from 13% to 11%, mainly because they spent heavily on the Singapore data centre. We need to revise the term sheet: secure one board seat and cap the pre-money valuation at RMB 150 million, or we will not invest. Hurry up and check the terms with AllBright; the bosses have a pre-IPO financing call with the other side at 3 p.m. on Friday. Also include the previous round's full-ratchet anti-dilution clause. We nearly got burned in the Xinglian Technology deal, so do not take this protection lightly. Their ESOP pool is 12%; as usual, we require pre-money dilution. Get their written confirmation.

*Selected rubric items:*

| Item | Role (weight) | Requirement |
|---|---|---|
| R2 | MUST (1) | Address the counterparty's project lead in a business email, with a salutation, communicative purpose, and closing. |
| R5 | MUST (1) | Preserve the pre-money valuation cap of RMB 150 million and the condition that investment will not proceed if it is unmet. |
| R9 | MUST (1) | Preserve the 12% ESOP pool and request written confirmation that dilution occurs before investment. |
| R13 | PLUS (0.5) | Use polite but clear external-facing language in place of the source's informal commands. |
| R18 | PLUS (0.5) | Turn internal instructions into actionable requests or discussion points for the recipient. |

Here, external-facing requests replace internal commands. The rubric checks both the financial conditions and the email's communicative function.

### A.7.3 CREATIVE REWRITING: SPECIFICATIONS TO MARKETING COPY

This example turns technical information into audience-oriented marketing copy. The task and complete Chinese source are rendered in English.

Item `en_042` contains 18 rubric items (nine MUST, nine PLUS). *Task:* Rewrite the Chinese product description as English home-cinema marketing copy for an Amazon North America product page. Address home-cinema enthusiasts with concise, professional copy emphasizing technology and cinema-quality imagery. Retain the 0.47-inch DMD chip and native 1080p resolution, 1,500 ANSI lumens, 10,000:1 contrast ratio, HDR10+ and HLG, MEMC, Amlogic T982, 4GB RAM, 64GB storage, and FengOS. Write 120–150 words and output only the final copy.

> This home-cinema projector uses a 0.47-inch DMD chip with native 1080p resolution. It supports HDR10+ and HLG high-dynamic-range decoding, provides 1,500 ANSI lumens and a 10,000:1 contrast ratio, and has a 1.2:1 throw ratio. It includes an Amlogic T982 quad-core processor, 4GB RAM, and 64GB storage, running the proprietary FengOS system. Automatic keystone correction, global MEMC motion compensation, and AI image enhancement support a 40–200-inch display. Connections include two HDMI 2.1 ports, one USB 3.0 port, and optical audio output, with plug-and-play operation.

*Selected rubric items:*

| Item | Role (weight) | Requirement |
|---|---|---|
| R2 | MUST (1) | Produce Amazon product-page marketing copy for North American home-cinema enthusiasts. |
| R3 | MUST (1) | Preserve the 0.47-inch DMD chip and native 1080p resolution; compatible resolution is not native resolution. |
| R9 | MUST (1) | Introduce no unsupported product features or promises, such as native 4K, a laser light source, or Netflix certification. |
| R13 | PLUS (0.5) | Organize the dense specifications into English information groups, such as picture quality, motion processing, and the smart platform. |
| R14 | PLUS (0.5) | Connect technical features to home-cinema use, such as movie nights or smooth sports and action scenes. |

The rubric combines factual boundaries with communicative adaptation: product claims remain supported by the source, while presentation can connect specifications to the viewing experience.

# B HUMAN ANNOTATION

Human review supports two distinct questions: whether generated evaluation material is valid, and whether automatic judgments agree with independent human labels. Thirty-nine multilingual experts contributed to the two annotation rounds below.

## B.1 ANNOTATORS AND PROCEDURE

**Recruitment and qualifications.** The experts in both rounds came from a translation company and worked on an annotation platform. Written requirements in both rounds asked for a background in translation studies, foreign languages and literatures, or the relevant language, and for competence in the assigned language pair. Round 1 further required a degree from a Double First-Class, Project 985, or Project 211 university; certified proficiency at TEM-8 good level or above, IELTS reading and writing scores of at least 7.5, TOEFL reading and writing scores of at least 28, or equivalent evidence; and experience with LLM-based translation, with MTI or CATTI credentials preferred. Round 2 accepted MTI, CATTI, or an equivalent qualification together with professional translation experience as evidence of competence, and prohibited machine translation and generative AI during annotation.

Experts were assigned to language pairs according to their working languages; eight of the 39 worked on more than one pair (Table 15). Round 1 involved 35 experts, each with at least five years of professional experience. Round 2 involved 23 experts, of whom 19 had also taken part in Round 1.

Table 15: Experts per language pair and round. Some experts worked on more than one pair, so the last column counts distinct experts.

| Round | Zh–En | Zh–Es | Zh–Fr | Zh–Ja | Zh–Ko | Distinct |
|---|---|---|---|---|---|---|
| Round 1: production audit | 12 | 8 | 10 | 5 | 10 | 35 |
| Round 2: blinded reference labels | 6 | 5 | 8 | 5 | 4 | 23 |
| Either round | | | | | | 39 |

**Round 1 procedure.** The guidelines defined the seven rubric dimensions and two audits for each rubric item (Table 16). The rubric audit also checked the item's dimension and MUST/PLUS role, and annotators could rewrite a flawed item. The Judge audit followed for items that were not rewritten; a rewritten item skipped it, because the visible verdict referred to the old text. After the last item of a question, annotators checked whether a core requirement was missing, and they could mark cases with unclear boundaries for arbitration.

**Round 2 procedure.** In Round 2, annotators acted as the Judge. Each unit contained the question, one anonymized response, and its rubric items with their roles and dimensions; files omitted model and Judge identities, Judge verdicts and rationales, earlier reference labels, reference answers, and source question identifiers. For each item, an annotator selected PASS, FAIL, or cannot judge and flagged any linguistic, domain, or boundary ambiguity; every FAIL and cannot-judge label carries a written reason that points to the error or missing content. The guidelines treated examples in an item as illustrations rather than required wording, and required every part of a multi-part item to be met for

Table 16: Checks in the Round 1 audit. A failed check is recorded with its tag.

| Check | Passing condition | Failure tag |
|---|---|---|
| *Rubric audit* | | |
| Judgeability | Given a response, the item can be judged PASS or FAIL consistently | Not judgeable |
| Grounding | The item derives from the question, source, task contract, or an explicit domain norm | Ungrounded |
| Task consistency | The item fits the language direction, deliverable, audience, and style goal | Task mismatch |
| Role | A MUST item states a minimum requirement; a PLUS item measures added quality | Wrong role |
| Example tolerance | An example rendering is not treated as the only acceptable answer | Example too narrow |
| Synonym tolerance | Valid synonymous, equivalent, or cross-sentence renderings are allowed | Too strict |
| Non-redundancy | The item does not duplicate or largely overlap another item of the question | Redundant |
| *Judge audit* | | |
| Verdict | The PASS/FAIL verdict matches the annotator's judgment | Wrong verdict |
| Evidence | The rationale cites specific content of the response | Insufficient evidence |
| Valid expressions | Valid synonymous, equivalent, or cross-sentence renderings are recognized | Synonym penalized |
| Examples | An example rendering in the item is not treated as the only correct answer | Locked to example |
| Formatting | Minor formatting, punctuation, or explanatory text is handled as the question requires | Format penalized |
| Core errors | Omitted, mistranslated, reversed, or added core information is failed | Core error missed |
| Boundary | An apparent Judge error caused by an unclear item boundary is identified as such | Unclear boundary |

a PASS. Twenty of the 100 questions, four per language pair, were assigned to two annotators, giving 720 response-level assignments for 600 responses. The two annotators labeled all six responses of each such question independently, and a reviewer then examined both labels and recorded the final one.

### B.2 PRODUCTION-QUALITY AUDIT

The first round assessed 500 earlier-version questions, with one selected model response per question and 100 questions per language pair. Its 8,683 rubric-item records cover wording, dimension assignments, aggregation roles, and automatic judgments. Responses came from DeepSeek V4 Pro for Zh–En, Qwen3.6-Max-Preview-Thinking for Zh–Es, Doubao Seed 2.0 Pro (high) for Zh–Fr and Zh–Ja, and GPT-5.5 (xhigh) for Zh–Ko.

Reviewers saw each question, response, rubric item, and the judge model's verdict and rationale. They first checked whether the requirement was valid and supported by the task, then checked its dimension and MUST/PLUS role. Verdict-and-rationale acceptance was assessed for valid requirements. Table 17 reports these checks separately: 99.32% of item texts were accepted, alongside 77.31% of dimension assignments and 96.19% of roles. This distinction identifies which part of a generated rubric requires correction. For the 453 retained questions, the release applies all 1,833 suggested dimension corrections and all 304 suggested role corrections; the remaining suggestions concern questions later replaced.

Table 17: Production-audit acceptance, with percentages in parentheses. Text, dimension, and role use all rubric items as denominators; Judge acceptance$^{\dagger}$ uses valid items (8,624 overall).

| Pair | Samples | Rubric items | Item text | Dimension | Role | Judge$^{\dagger}$ |
|---|---|---|---|---|---|---|
| Zh–Ja | 100 | 1,810 | 1,808 (99.89) | 1,323 (73.09) | 1,704 (94.14) | 1,807 (99.94) |
| Zh–Fr | 100 | 1,800 | 1,779 (98.83) | 1,236 (68.67) | 1,712 (95.11) | 1,770 (99.49) |
| Zh–En | 100 | 1,807 | 1,805 (99.89) | 1,533 (84.84) | 1,782 (98.62) | 1,798 (99.61) |
| Zh–Es | 100 | 1,593 | 1,575 (98.87) | 1,264 (79.35) | 1,557 (97.74) | 1,567 (99.49) |
| Zh–Ko | 100 | 1,673 | 1,657 (99.04) | 1,357 (81.11) | 1,597 (95.46) | 1,634 (98.61) |
| Total | 500 | 8,683 | 8,624 (99.32) | 6,713 (77.31) | 8,352 (96.19) | 8,576 (99.44) |

### B.3 BLINDED REFERENCE JUDGMENTS

The second round provided reference labels for the Judge study on 100 questions from the initial benchmark, 20 per language pair; 97 remain in the release, and the other three were replaced during renewal. All 100 had been audited in Round 1, but none of the 600 Round 2 responses repeats a response audited there. Experts judged six anonymized responses per question with model identities

and automatic verdicts hidden. The 1,740 rubric items yield 10,440 rubric-item–response positions. Final records contain 10,419 binary labels (5,448 MUST and 4,971 PLUS), with 21 unresolved positions excluded from item-agreement calculations.

Twenty questions, four per language pair, received independent double annotation across their 120 responses, followed by expert review that set the final labels. The remaining 80 questions received single annotations. Judge comparisons use the final reference labels.

# C BENCHMARK EVALUATION DETAILS

## C.1 MODELS, INPUTS, AND SCORING COVERAGE

The benchmark comparison includes ten general-purpose models and three translation-specialized models: Qwen-MT-Plus, HY-MT2-7B, and HY-MT2-1.8B. Each receives the complete task instruction and source passage, without rubric items or reference answers. Translation-specialized models are evaluated on the same translation and rewriting instructions as the general-purpose models. Qwen-MT-Plus uses its custom-prompt interface with an 8,192-token output ceiling and seed 20260921; inputs are preserved without truncation. GPT-5.5 Medium judges the fixed rubric items using the fairness-calibrated prompt, with a 16,384-token output ceiling. The prompt's calibration rules treat examples in a rubric item as boundaries rather than required wording and ask the Judge to read the surrounding context. A MUST item fails only when a required fact, constraint, relation, or purpose is missing, wrong, reversed, or materially weakened and the Judge can cite the failing expression; formatting and extra explanations count only when the task requires them, and the same standard applies across models.

For a matched comparison across all 13 models, Table 3 uses the 494 questions with nonempty responses and complete judgments for every model. These comprise 231 translation and 263 rewriting questions, distributed as 96 Zh–En, 99 Zh–Es, 100 Zh–Fr, 100 Zh–Ja, and 99 Zh–Ko. The six excluded questions contain empty responses, content-filter outcomes, or incomplete judgments. HY-MT2-7B has complete results on all 500 questions; HY-MT2-1.8B and Qwen-MT-Plus each have 498 valid results.

Table 18 reports the matched comparison alongside a fixed-500 score, assigning zero to unavailable positions. The production study uses this fixed-500 convention for the original ten-model panel. Model means give equal weight to questions and use Strictly Hybrid scoring (Equation 1); all displayed scores are on a 0–100 scale.

## C.2 EFFECT OF THE ESSENTIAL-REQUIREMENT GATE

The validity gate gives priority to satisfying essential requirements. To show how it affects aggregation, Table 19 compares Strictly Hybrid with the ungated weighted score, $100 \sum_i w_i z_i / \sum_i w_i$. Both calculations use the same 500 questions, ten models, rubric-item verdicts, weights, and treatment of unavailable positions.

Ungated scores range from 88.97 to 93.89, whereas Strictly Hybrid scores range from 40.29 to 63.71. The rankings share the leading model and have Spearman correlation $\rho = 0.891$. The difference follows from their measurement targets: the ungated mean credits requirements independently, while the gate makes essential failures decisive for task success. The item-level judgments remain available to identify the requirements responsible for each zero score.

The gate does not mechanically penalize longer rubrics. Across the 500 released questions, the share of zero scores varies little with the number of MUST items (44.4% for 5–7 items, 50.4% for 8–11, and 52.7% for 12–16), and the per-question zero rate is uncorrelated with that number (Spearman $\rho = 0.03$). Every model answers the same rubric, so the number of MUST items is a property of the question rather than of the model.

## C.3 BLINDED VALIDATION OF THE BENCHMARK JUDGE

We applied the benchmark Judge, GPT-5.5 Medium with the fairness-calibrated prompt of Appendix C.1, to the Round 2 responses of 66 singly annotated reference questions that remain in the

Table 18: Translation and rewriting scores on the shared valid subset, with a fixed-500 sensitivity check (0–100).

| Model | Overall (494) | Translation (231) | Rewriting (263) | Fixed-500 overall |
|---|---|---|---|---|
| *General-purpose models* | | | | |
| GPT-5.6-Sol(max) | 63.67 | 67.98 | 59.89 | 63.71 |
| Qwen-3.8-Max-Thinking | 52.94 | 57.33 | 49.08 | 52.70 |
| GLM-5.3(max) | 52.63 | 54.71 | 50.80 | 52.60 |
| Kimi-K3(max) | 49.24 | 49.26 | 49.23 | 49.05 |
| Claude-Opus-5(max) | 47.68 | 52.83 | 43.15 | 47.49 |
| Grok-4.6(xhigh) | 47.40 | 43.90 | 50.47 | 46.83 |
| Doubao-Seed-2.1-Pro(high) | 47.11 | 50.47 | 44.15 | 46.93 |
| HY-3.0(high) | 45.91 | 47.50 | 44.51 | 45.76 |
| Muse-Spark-1.2(xhigh) | 43.84 | 46.05 | 41.89 | 43.70 |
| Gemini-3.7-Flash(high) | 40.57 | 50.53 | 31.83 | 40.29 |
| *Translation-specialized models* | | | | |
| Qwen-MT-Plus | 31.80 | 37.84 | 26.50 | 31.42 |
| HY-MT2-7B | 25.14 | 31.44 | 19.61 | 25.42 |
| HY-MT2-1.8B | 13.84 | 14.08 | 13.63 | 13.88 |

Table 19: Sensitivity to aggregation on the same released benchmark. Model rows show rank / score (0–100). Only the validity gate changes.

| Model | Strictly Hybrid | Ungated weighted |
|---|---|---|
| GPT-5.6-Sol(max) | 1 / 63.71 | 1 / 93.89 |
| Qwen-3.8-Max-Thinking | 2 / 52.70 | 3 / 91.84 |
| GLM-5.3(max) | 3 / 52.60 | 2 / 91.89 |
| Kimi-K3(max) | 4 / 49.05 | 5 / 91.52 |
| Claude-Opus-5(max) | 5 / 47.49 | 4 / 91.56 |
| Doubao-Seed-2.1-Pro(high) | 6 / 46.93 | 6 / 91.17 |
| Grok-4.6(xhigh) | 7 / 46.83 | 10 / 88.97 |
| HY-3.0(high) | 8 / 45.76 | 7 / 89.29 |
| Muse-Spark-1.2(xhigh) | 9 / 43.70 | 9 / 89.19 |
| Gemini-3.7-Flash(high) | 10 / 40.29 | 8 / 89.20 |
| Panel mean | 48.91 | 90.85 |
| Highest–lowest model gap | 23.42 | 4.92 |
| Between-model population SD | 6.07 | 1.54 |

release with unchanged rubrics: 396 responses from six answer models and 6,948 rubric items. The experts labeled these responses with model identities and automatic verdicts hidden, and none of the responses appears in the leaderboard runs. As a reference level, the 20 double-annotated questions give the agreement between the two experts' independent labels before adjudication (120 responses; 2,045 rubric items with two valid labels). The two comparisons cover different questions, so Table 20 compares levels of agreement rather than paired differences.

Table 20: Agreement (%) with blinded expert labels. The MUST gate column counts responses for which both sides reach the same decision on whether every MUST item passes.

| Comparison | All items | MUST items | PLUS items | MUST gate |
|---|---|---|---|---|
| GPT-5.5 Medium vs. experts (66 questions) | 90.4 | 93.4 | 87.1 | 71.2 |
| Expert vs. expert (20 questions) | 88.3 | 89.6 | 87.0 | 61.7 |

The Judge agrees with the expert labels at least as often as the two experts agree with each other in every column. Agreement falls from items to the MUST gate in both comparisons, because one disputed MUST item decides the gate; the amplification therefore reflects conjunctive essential requirements rather than the automatic Judge.

# D FEEDBACK-GUIDED BENCHMARK RENEWAL

## D.1 FEEDBACK AND GENERATION STRATEGIES

The production loop transforms evaluation records into generation targets. Failures are grouped by language, task, competence, and linguistic phenomenon, while coverage summaries identify targets absent from the current pool. The resulting strategy specifies a capability to probe, a source phenomenon that elicits it, and the response requirements that distinguish success from failure. Questions and rubric items are then constructed from the same testpoint, following Section 3.2.

Three strategies examine the contribution of this feedback. *Full feedback* combines observed failures, coverage gaps, score spread, and difficulty targets. *Difficulty only* receives only the aggregate difficulty signal. *Static balancing* follows predetermined coverage targets without historical model feedback. The feedback came from a separate development round rather than from the evaluation panel: 60 questions answered by GPT-5.5 (xhigh), Kimi K2.6 Thinking, and Doubao Seed 2.0 Pro (high), judged under Strictly Hybrid scoring and aggregated into ten language–task-family groups. No strategy receives model identities, responses, Judge rationales, reference answers, rubric text, or per-question feedback. The released replacements contain 16, 16, and 15 questions from these strategies, respectively. Selection drew 39 questions from a 60-question pilot and eight from nine supplementary candidates, filling the language-pair slots of the removed questions (Zh–En 18, Zh–Fr 14, Zh–Es 9, Zh–Ja 6) without using panel scores. The comparison in Table 4 concerns these selected groups under nonrandomized allocation.

## D.2 COMPARISON UNDER A COMMON JUDGE PROTOCOL

The removed and replacement sets each contain 47 questions, evaluated using the same ten-model panel and GPT-5.5 Medium fairness-calibrated rubric-level protocol. The original 470 responses to the removed questions were rescored, yielding valid judgments for every response. Replacement scores use their completed release judgments. Each version retains its own questions and rubrics, and model scores average equally over questions.

The removed and replacement sets score 80.95 and 43.32, respectively. Across the complete 500-question versions, the mean falls from 52.19 to 48.91.

For the full-500 comparison, common questions use aligned response records and their version-specific rubrics; 17 positions missing from the original response archive use the corresponding release records in both versions. Five retained questions received rubric-wording revisions during release calibration (Appendix A.4). The 47-question replacement contributes a 3.54-point reduction to the full-benchmark mean; the common-question revisions increase it by 0.26 points, giving the observed net decrease of 3.28 points. Historical response errors and empty outputs retain the fixed-denominator treatment in Appendix C.1. The full-feedback, difficulty-only, and static-balancing groups in Table 4 partition the 47 replacements.

## D.3 DIFFICULTY AND MODEL DISCRIMINATION

All 47 replacements and 453 retained questions are evaluated by the same final GPT-5.5 Medium Judge and ten-model panel. For question $i$, let $s_{im} \in [0, 1]$ be model $m$'s Strictly Hybrid score. Difficulty and score dispersion are

$$H(i) = 1 - \frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} s_{im}, \qquad D(i) = \mathrm{Var}_{m \in \mathcal{M}}(s_{im}). \tag{2}$$

The reported mean score averages over questions and models. Mean item SD averages $100\sqrt{D(i)}$ over questions, measuring how strongly models differ on an average item. An all-zero item has $H(i) = 1$ and $D(i) = 0$: it is difficult for the panel but supplies no within-item score separation. Reporting all three measures distinguishes challenge from discrimination.

The full-feedback example in Appendix A.3 illustrates the resulting construction path. Failures involving content correctness and consistency, together with an uncovered structural-transfer target, lead to an administrative rewriting task that tests agency and permission boundaries. Its two

targeted requirements become separate MUST items, making the subsequent model behavior directly interpretable against the generation objective.

Table 21 checks whether the strategy ranking still holds when we score models that did not drive the replacement loop. The middle row drops GPT-5.6-Sol, Kimi-K3, and Doubao-Seed-2.1-Pro, because earlier models from those families supplied the development feedback. The bottom row uses only the three translation-specialized models, which never entered generation or selection.

Table 21: Replacement strategies on the full panel and on models outside the feedback loop (0–100; mean score / mean item SD; SDs are omitted for the three translation-specialized models, whose mostly low scores make the spread unstable). Judge: GPT-5.5 Medium.

| Evaluation models | Static | Difficulty only | Full feedback | All 47 | Retained 453 |
|---|---|---|---|---|---|
| Ten-model panel | 51.25 / 26.54 | 36.15 / 22.59 | 43.06 / 35.16 | 43.32 | 49.49 |
| Seven models from other families | 50.05 / 22.28 | 33.25 / 20.71 | 43.56 / 34.97 | 42.12 | 47.56 |
| Three translation-specialized models | 25.11 | 20.49 | 13.60 | 19.62 | 23.98 |

# E JUDGE RULE LEARNING

## E.1 DESIGN, SPLITS, AND COMPARISON ARMS

The Judge study uses 100 human-labeled questions, 20 per language pair, with six anonymized responses each. The answer models are GPT-5.5 (xhigh), HY-3.0 (high), DeepSeek V4 Pro, Claude Opus 4.8, Grok 4.5 (high), and Gemini 3.1 Pro. Qwen3.6-27B supplies both judgments and rule proposals with frozen weights. Four-fold cross-fitting evaluates 80 development questions, using 60 induction and 20 evaluation questions per fold. W0 evaluates another 20 questions using rules induced from all 80 development questions. W0 is excluded from current induction but had appeared in earlier development folds. The combined results describe these 60- and 80-question induction settings.

Five arms share the initial judgments: direct judging (S0), learned cards, resampling, scoped placebo, and blank placebo. Each rubric-item–response pair receives three initial draws. For each draw, the gated analysis reconstructs verdicts by replacing scope-matched initial FAIL decisions with their archived rechecks, then majority-votes the resulting three verdicts. S0 votes over the initial draws. Cards and placebos reuse the corresponding initial seed; resampling changes it. Scoped placebo retains the scope label while replacing the trigger and both boundaries with approximately character-length-matched neutral text; blank placebo also replaces the label. Routing uses the originating evidence cluster's scope, independently of rendered text, giving all rechecking arms the same eligible initial failures. Arms ran in separate inference batches.

## E.2 RULE INDUCTION AND ROUTING

Rule state $S_t = (R_t, L_t)$ contains active cards $R_t$ and their evidence ledger $L_t$:

$$S_{t+1} = \mathcal{U}(S_t, \mathcal{F}_t), \qquad J_t(x, r) = J_{\theta,P}(x, r; R_t), \tag{3}$$

where $x$ is a question–response input, $r$ a rubric item, and $\theta, P$ the frozen backbone and base instructions. Human disagreements supply feedback $\mathcal{F}_t$ after rubric-item validity checks. Each fold starts from empty $S_0$ and adds rules once. Structural, evidence-reference, and leakage checks precede evaluation.

Cards specify scope, trigger, decision boundary, and counter-boundary. The following abbreviated English translation of an actual tone card illustrates these fields:

> **Scope:** content correctness—tone.
> **Trigger:** the rubric explicitly requires sarcasm or irony and prohibits literal translation.
> **Decision boundary:** wording or sentence structure makes suspicion, mockery, or irony recognizable to the target reader, avoiding neutral or merely literal rendering.
> **Counter-boundary:** use the original rubric when the source is neutral or positive, or the rubric has no special tone requirement.

For initial verdict $b_{ik}$, cluster-scope predicate $g_i$, and arm-$a$ recheck $r_{ik}^{(a)}$, execution is

$$\tilde{z}_{ik}^{(a)} = \begin{cases} r_{ik}^{(a)}, & g_i = 1 \text{ and } b_{ik} = 0, \\ b_{ik}, & \text{otherwise}, \end{cases} \qquad \hat{z}_i^{(a)} = \text{majority}_{k=0,1,2}\, \tilde{z}_{ik}^{(a)}. \tag{4}$$

The mechanical scope predicate matches task type, rubric dimension, and MUST/PLUS role, and language when specified by the source cluster. The rechecking Judge interprets the card's natural-language trigger within this scope. Scope and the initial verdict determine routing; human labels enter rule induction and evaluation. Judge decoding uses temperature 1, top-$p$ 0.95, top-$k$ 20, a 32,768-token context, and an 8,192-token output ceiling. Proposal uses temperature 0, top-$p$ 1, a 131,072-token context, and a 16,384-token output ceiling.

Table 22: One-step rule induction by evaluation block.

| Evaluation block | Induction questions | Evidence clusters | Cards | Draft input tokens |
|---|---|---|---|---|
| W1 | 60 | 5 | 6 | 40,725 |
| W2 | 60 | 8 | 5 | 53,329 |
| W3 | 60 | 7 | 6 | 49,578 |
| W4 | 60 | 7 | 6 | 51,534 |
| W0 (evaluation) | 80 | 8 | 6 | 71,031 |

### E.3 METRICS AND ELIGIBLE SAMPLES

The 1,740 rubric items produce 10,440 rubric-item–response positions. Twenty-one human labels are unresolved: six PLUS in W0, nine PLUS in W1, and six MUST in W4. One W3 seed0 MUST verdict is invalid; voting over valid values produces a verdict for every position. Accuracy uses known reference labels. Repairs and harms count wrong-to-right and right-to-wrong changes from S0; question-level gains sum these changes across six responses.

Table 23: Eligible samples for three-draw voting. Known judgments count rubric-item–response pairs.

| Sample | Known judgments | Known MUST | Responses | MUST-complete |
|---|---|---|---|---|
| Development 80 | 8,349 | 4,386 | 480 | 474 |
| W0 (20) | 2,070 | 1,062 | 120 | 120 |
| Combined 100 | 10,419 | 5,448 | 600 | 594 |

Response failure is the positive label when any MUST item fails. Among the 594 MUST-complete responses, human labels identify 243 failures and 351 passes. False failures are false positives and false passes are false negatives; $F1 = 2TP/(2TP + FP + FN)$. Rankings compare six model means over 96 questions with complete rubric labels (576 responses). Judge-study Strictly Hybrid scores use Eq. 1 with weights 1 for MUST and 0.5 for PLUS; fly-corpus uses its exported rubric weights.

### E.4 THREE-DRAW RESULTS AND PAIRED COMPARISONS

Cards reduce MUST errors from 307 to 291 and, under three-draw voting, raise ranking Spearman correlation from 0.371 to 0.600; the single-draw comparison in Appendix E.5 does not reproduce this ranking gain. Resampling and the blank placebo reach 0.314 and 0.086; on MUST items alone, direct judging, cards, resampling, and the blank placebo reach 0.382, 0.456, 0.250, and 0.029. Within language pairs, cards fall below direct judging in four of five pairs, so the pooled ranking gain is descriptive. Table 25 separates development and W0 outcomes; Table 26 compares paired correctness directly.

Exact two-sided McNemar tests use discordant correctness pairs. Question-level tests compare the observed absolute sum of per-question differences $d_q$ against $|\sum_q \epsilon_q d_q|$, with independent equiprobable signs, using exact discrete convolution. A question-clustered percentile bootstrap with 20,000 resamples gives net gains of $+16$ $[+3, +29]$ for cards versus S0 and $+11$ $[+0, +23]$ versus blank placebo (95% intervals). A post hoc Holm check across four arm-versus-S0 rubric-item tests does not retain significance for cards. These exploratory analyses condition on the realized folds, rules,

Table 24: Correction outcomes across five Judge arms on 100 questions (600 responses), relative to direct judging.

| Measure | Direct | Cards | Resampling | Scoped placebo | Blank placebo |
|---|---|---|---|---|---|
| Questions improved ↑ | – | **20** | 15 | 17 | 13 |
| Questions worsened ↓ | – | **6** | 8 | 10 | 11 |
| Questions with zero net change | – | 74 | 77 | 73 | 76 |
| MUST errors ↓ | 307 | **291** | 301 | 299 | 302 |
| MUST error reduction (%) ↑ | – | **5.21** | 1.95 | 2.61 | 1.63 |
| Errors repaired ↑ | – | **29** | 23 | 24 | 24 |
| New errors introduced ↓ | – | **13** | 17 | 16 | 19 |
| Net additional correct decisions ↑ | – | **+16** | +6 | +8 | +5 |

Table 25: Split results after three-draw voting. Repairs, harms, and net gains are relative to S0; failure F1 is a percentage.

| Sample | Arm | Repairs / harms | Net | Failure F1 |
|---|---|---|---|---|
| Development 80 | S0 | – | – | 67.03 |
| | Cards | 23 / 13 | +10 | 63.56 |
| | Resampling | 17 / 15 | +2 | 65.14 |
| | Scoped placebo | 19 / 14 | +5 | 63.98 |
| | Blank placebo | 19 / 18 | +1 | 64.93 |
| W0 (20) | S0 | – | – | 72.38 |
| | Cards | 6 / 0 | +6 | 75.25 |

Table 26: Paired contrasts. Wins/losses count rubric-item judgments correct under only one arm. Rubric-item $p$ uses exact McNemar; question $p$ uses sign flips. Values are unadjusted.

| Sample | Contrast | Wins / losses | Rubric-item $p$ | Question $p$ |
|---|---|---|---|---|
| Development 80 | Cards–S0 | 23 / 13 | .1325 | .1342 |
| | Cards–scoped placebo | 14 / 9 | .4049 | .4284 |
| | Cards–resampling | 14 / 6 | .1153 | .1163 |
| | Cards–blank placebo | 18 / 9 | .1221 | .0933 |
| W0 (20) | Cards–S0 | 6 / 0 | .0313 | .1250 |
| | Cards–blank placebo | 5 / 3 | .7266 | .7812 |
| Combined 100 | Cards–S0 | 29 / 13 | .0195 | .0242 |
| | Cards–scoped placebo | 19 / 11 | .2005 | .2456 |
| | Cards–resampling | 19 / 9 | .0872 | .0872 |
| | Cards–blank placebo | 23 / 12 | .0895 | .0896 |
| | Scoped–blank placebo | 13 / 10 | .6776 | .6831 |

and draws; shared induction data and selecting gating after inspecting always-on results lie outside their uncertainty estimates. Blanking scope labels also makes cards within a fold indistinguishable, coupling label information with distinguishable card identities.

**Response-level outcomes.** False failures/false passes are 68/82 for S0, 54/96 for cards, 58/93 for resampling, 57/96 for scoped placebo, and 55/94 for blank placebo. Cards exchange 14 false failures for 14 false passes, leaving 150 response errors. Development failure F1 changes from 67.03% to 63.56%, with split results in Table 25. A repaired rejection can open the MUST gate while another genuine failure remains undetected; five of the 14 newly false-passing development responses contain no newly incorrect rubric-item judgment.

### E.5 FIXED SINGLE-DRAW COMPARISON

Table 27: Agreement with human judgments under three-draw voting and fixed single-draw evaluation. Cards recheck scope-matched initial failures. Accuracies are percentages; error counts concern essential rubric items.

| *Three-draw majority vote* | | | |
|---|---|---|---|
| Judge | Item acc. ↑ | MUST acc. ↑ | Rank $\rho$ ↑ |
| Qwen3.6-27B | 91.50 | 94.36 | 0.371 |
| Qwen3.6-27B + cards | **91.65** | **94.66** | **0.600** |
| *Single-draw comparison* | | | |
| Judge | Item acc. ↑ | MUST acc. ↑ | MUST errors ↓ |
| Qwen3.6-27B | 90.99 | 93.94 | 330 |
| Qwen3.6-27B + cards | **91.19** | **94.33** | **309** |

The single-draw evaluation uses Qwen's first draw (seed0) alone. Applying Eq. 4 to that draw gives one final verdict per item. Its invalid initial verdict remains invalid under the card policy, leaving 10,418 valid pairs, including 5,447 MUST pairs. Cards reduce MUST errors from 330 to 309, a 6.36% reduction (Table 27). Rankings use 95 complete questions after excluding the additional invalid-verdict question; on this draw, Spearman correlation is 0.429 without cards and 0.200 with cards.

### E.6 RULE EVIDENCE AND ROUTED CARDS

The tone card above cites two records from one Spanish-to-Chinese dialogue question. Marta's *Qué conveniente* must retain sarcasm. A literal-looking phrase occurs within an explicitly mocking passage; the human label is PASS and the initial Judge label FAIL. Together with an agreed failure, this evidence motivates a rule about recognizable pragmatic force.

A card reaches a rubric item when the item's task type, dimension, and role match the card's originating evidence cluster, and its language matches when the cluster specifies one. Each recheck prompt contains every card routed to its item. Tables 28 and 29 detail the two corrections, and Table 30 translates the cards routed to the two corrections in Section 5.4; because several cards reach each item, neither correction is attributed to a single card.

Table 28: Correction A: equivalent terminology (English→Chinese, evaluation block W0). English glosses are ours.

| Field | Record |
|---|---|
| Task | Translate an English report on a Ministry of Education guideline into Chinese for a WeChat article aimed at parents and educators. |
| Source excerpt | "The document emphasizes the need to integrate five areas of education—moral, intellectual, physical, aesthetic, and labor education—and to uphold a 'health first' philosophy." |
| Rubric item (MUST) | Render the phrase with an established Chinese education term such as 五育并举 or 德智体美劳五育融合; a generic "integrate five educational areas" that loses the five-education meaning fails. |
| Response (HY-3.0) | 要把德育、智育、体育、美育和劳动教育这五方面的教育融合起来 (*integrate these five kinds of education: moral, intellectual, physical, aesthetic, and labor education*) |
| Verdicts | Direct judging: FAIL in two of three draws → FAIL. Rule-guided rechecking: PASS in all three draws → PASS. Human: PASS. Response score: $0 \rightarrow 100$ (human score 100). |
| Routed cards | card-000003 (tone) and card-000004 (terminology), both induced from a translation / content-correctness / MUST evidence cluster. |
| Terminology card | *Trigger:* the rubric requires a compound policy term or proper noun to cover every object or dimension it contains. *Boundary:* accept if the translation explicitly lists or explains all core objects the rubric names, omitting none. *Counter-boundary:* not applicable if the rubric asks only for the term itself or allows a common abbreviation. |

Table 29: Correction B: contextual reference (Chinese→Japanese, evaluation block W1). English glosses are ours.

| Field | Record |
|---|---|
| Task | Translate a WeChat message to a close Japanese friend into casual Japanese, at most 80% of the source length, keeping all proper nouns; output only the translation. |
| Source excerpt | 嘿，小红！……上次说想辞职，现在考虑得咋样了？ (*Hey, Xiaohong! … You said last time you wanted to quit—how are you thinking about it now?*) |
| Rubric item (MUST) | Convey that Xiaohong said earlier she wanted to resign and ask how she now feels, e.g., 「前に辞めたいって言ってたけど、今どう考えてる？」; fails if the speaker or Xiaoqiang becomes the one resigning, or if a stiff literal rendering obscures who is addressed. |
| Response | ねえ小红！……辞めたい話、今どう？ (*Hey, Xiaohong! … The quitting thing—how's it going now?*) |
| Verdicts | Direct judging: FAIL in all three draws → FAIL. Rule-guided rechecking: FAIL, PASS, PASS → PASS. Human: PASS. |
| Direct rationale | "Too brief: it omits the time reference 'last time' and the explicit agent, so the relation 'Xiaohong had said she wanted to quit' is unclear and the reference becomes ambiguous out of context" (third draw, translated). |
| Routed cards | card-000002 (organization names), card-000003 (a person's identity and itinerary), and card-000004 (caller relations in relayed calls); none concerns participant reference. |
| Closest boundary | card-000003: accept if the translation conveys the person's core affiliation and itinerary purpose, even if length limits lead it to omit a title or not to match words such as "originally" or "attend", provided the role relation is not distorted. |

Table 30: Cards routed to the two corrected judgments (English translation). Every trigger additionally states that the card does not apply otherwise and that the rubric text then governs.

| Card | Trigger | Decision boundary | Counter-boundary |
|---|---|---|---|
| *Correction A (fold for W0): routed to the five-education item* | | | |
| card-000003 (tone) | The rubric requires the translation to convey the source's sarcasm, mockery, or irony explicitly and prohibits literal translation. | Accept if word choice or sentence structure lets target readers recognize the speaker's suspicion, mockery, or irony, without a neutral or literal rendering. | Not applicable if the source tone is neutral or positive, or the rubric sets no tone requirement. |
| card-000004 (terminology) | The rubric requires the rendering of a compound policy term or proper noun to cover all specific objects or dimensions it contains. | Accept if the translation explicitly lists or explains all core objects the rubric names, without omitting any specified dimension. | Not applicable if the rubric requires only the term itself or allows a common target-language abbreviation. |
| *Correction B (fold for W1): routed to the Xiaohong item* | | | |
| card-000002 (organization names) | The rubric requires an institution's name in full form with its country qualifier and forbids rendering it as a personal title. | Accept if the translation conveys the institution's official status without turning it into a personal title, even if its wording differs literally from the rubric's examples. | Not applicable if the name becomes a personal title, an ordinary position, or a term unrelated to the institution; the item then fails. |
| card-000003 (identity and itinerary) | The rubric requires the translation to state a person's identity and original itinerary and forbids changing their role relation. | Accept if the translation conveys the person's core affiliation and itinerary purpose, even if length limits lead it to omit a title or not to match words such as "originally" or "attend", provided the role relation is not distorted. | Not applicable if the person becomes the host, an employee of another company, or an unrelated role, or if the destination and nature of the trip are omitted; the item then fails. |
| card-000004 (caller relations) | The rubric requires the call relation to be presented correctly, identifying caller, callee, and the core content, and forbids misattributing the caller. | Accept if the translation conveys the calling party, the called party, and the core data and terms to confirm, even if prepositions or particles shift reference slightly, provided the relation is not reversed and no content is lost. | Not applicable if caller and callee are reversed or the core data and terms listed by the rubric are omitted; the item then fails. |