ATOMDIALBENCH: FROM AGGREGATE SCORES TO ATOMIC TASK PROFILES FOR DIALOGUE UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

As organizations increasingly deploy large language models (LLMs) for customer-service review and meeting analysis, model selection demands operation-level evidence rather than merely an aggregate rank. Existing resources offer valuable evaluations but differ in the operations they cover, the outputs they require, and how those outputs are scored. We introduce **AtomDialBench**, a Chinese–English diagnostic benchmark with **16 atomic tasks**, two composite evaluation units, and **3,459 instances**. Each task specifies an operation, output requirements, and a scoring procedure; the construction pipeline combines dialogue selection, schema-guided synthesis, and factual checks. Here, *atomic* means separately specified and scored rather than an independent latent ability. Across **21 LLMs from seven families**, the top five aggregate scores fall within **0.013**, yet task-level rankings differ: the aggregate leader ranks first or joint first on seven of 16 tasks but last on keyword extraction, where it trails the task leader by **0.111**. A 900-response human study yields high inter-annotator agreement (ICC(2,1) $= 0.918$) and a system–human Spearman correlation of $\rho = 0.697$. In a controlled comparison with similarly sized training sets, task-aligned fine-tuning yields a higher aggregate score but uneven operation-level changes. Together, AtomDialBench supports aggregate screening, operation-specific comparison, and response-level diagnosis while complementing existing evaluations.

1 INTRODUCTION

Customer-service review and meeting analysis use conversation records to inform operational decisions (Zhang et al., 2024c; Lin et al., 2021; Zhong et al., 2021). Misidentified intentions, overlooked commitments, or misattributed speakers can distort follow-up actions and resource allocation. A reviewer may need to identify a customer's intention or verify an agent's commitment; a meeting analyst may need a summary of decisions and their owners. In this paper, *dialogue analysis* means analyzing a supplied dialogue: an LLM need not participate in or continue the original interaction. This scope covers summaries, labels, and extracted values. Reliable analysis requires attributing evidence to speakers, integrating turns, and respecting the requested scope.

Established resources span three complementary areas: summaries of everyday, customer-service, and meeting conversations through SAMSum, CSDS, and QMSum (Gliwa et al., 2019; Lin et al., 2021; Zhong et al., 2021), task-oriented dialogue annotations through MultiWOZ (Budzianowski et al., 2018), and broader multi-task coverage through DialoGLUE and InstructDial (Mehri et al., 2020; Gupta et al., 2022). Their distinct task scopes provide complementary evidence rather than interchangeable measurements of the same behavior (§2).

Direct operation-level model comparisons across these resources remain difficult because each uses a different task scope, output format, and scoring contract. Aggregate scores also do not show whether a model missed relevant evidence, returned an incorrect value, or violated a requested restriction. We therefore introduce AtomDialBench to enable operation-level comparisons with inspectable responses, complementing existing evaluations.

AtomDialBench comprises 16 atomic tasks (ATs), two composite evaluation units, and 3,459 Chinese–English instances evaluated across 21 LLMs from seven families (Figure 1). Each AT

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.

pairs an operation with an output contract and scoring procedure, linking instructions to checkable behavior. Five reporting groups organize ATs without implying independent abilities. Multi-Task Dialogue Analysis (MTDA) combines requirements for one supplied dialogue in a single prompt; Instruction-Constrained Dialogue Analysis (ICDA) adds instruction constraints. The aggregate AtomDialBench score (ADB) equally averages the 18 unit scores.

The top five ADB scores fall within 0.013, yet task-level rankings differ substantially. The aggregate leader, GPT-5.6-Sol, ranks first or joint first on seven of 16 ATs but outside the top three on seven. Its largest gap occurs on keyword extraction, where it ranks last and trails the task leader by 0.111. Across the 16 ATs, its mean gap to the strongest observed model is 0.029, driven mainly by keyword extraction, summarization, and attribute inference. These are descriptive within-benchmark comparisons rather than held-out model-selection gains; Section 4.3 reports their uncertainty and selection limitations. The profiles also contain ranking reversals despite near-tied aggregate scores: GPT-5.6-Sol exceeds Claude-Opus-5 on candidate selection but trails it on exemption judgment. The relevant comparison therefore depends on the application's requested operation and output requirements.

For model comparison, aggregate scores provide an initial shortlist; operation-level profiles then support closer comparison of the required tasks and inspection of missed evidence, incorrect values, or constraint violations. We also use a controlled fine-tuning comparison to examine whether the profiles capture operation-specific changes. The task-aligned model has a higher aggregate score than a model trained on a similarly sized general-purpose dataset, but the changes are not uniform across operations. This comparison illustrates how AtomDialBench separates broad gains from operation-specific trade-offs.

Our contributions are:

1. **An atomic-and-composite task framework** that defines 16 operation-level ATs with explicit contracts and two composite evaluation units for dialogue analysis.
2. **A bilingual diagnostic resource** of 3,459 instances with inspectable outputs, task-specific scoring, and a 900-response human study of scoring agreement.
3. **Task-profile evidence across 21 LLMs** showing that near-tied aggregate scores can hide operation-level rank reversals, together with a controlled fine-tuning study that separates aggregate gains from operation-specific trade-offs.

2 Related Work

Dialogue-analysis resources. MultiWOZ supports task-oriented dialogue research, while SAMSum, DialogRE, MELD, and CIDER target summarization, relations, emotion, and consistency (Budzianowski et al., 2018; Gliwa et al., 2019; Yu et al., 2020; Poria et al., 2019; Zhang et al., 2024b). DialoGLUE and InstructDial broaden task coverage (Mehri et al., 2020; Gupta et al., 2022). AtomDialBench complements these application-specific resources with shared operation-level contracts for supplied-dialogue analysis. DialogBench (Ou et al., 2024) and TD-EVAL (Acikgoz et al., 2025) evaluate interactive dialogue capabilities; AtomDialBench instead targets post-hoc analysis of supplied transcripts (Table 1; extended comparison in Appendix A).

Diagnostic and compositional evaluation. FLASK assesses response-level skills; MT-Bench-101 evaluates multi-turn interaction; CheckList and SKILL-MIX provide behavioral and compositional tests (Ye et al., 2023; Bai et al., 2024; Ribeiro et al., 2020; Yu et al., 2024). Our distinction is the combination of given-dialogue operations, inspectable contracts, and composite prompts, rather than broader task coverage alone.

Score interpretation. Following measurement arguments that tie scores to evidence for their intended use (Cronbach & Meehl, 1955; Schlangen, 2021; Bean et al., 2025), we examine scoring agreement, task relationships, and fine-tuning responses (§4.5), not independent traits or cognitive mechanisms.

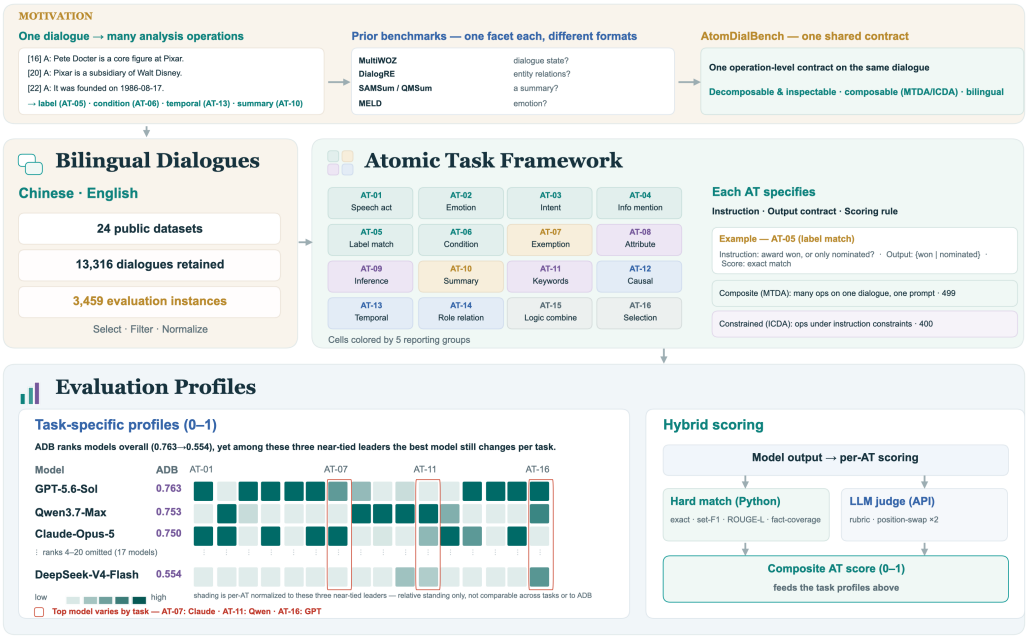

Figure 1: **AtomDialBench overview: from bilingual dialogues to operation-level task profiles.** From 13,316 retained dialogues, we construct 3,459 instances covering 16 ATs, MTDA, and ICDA. Explicit instructions and output contracts connect task-specific scoring to inspectable profiles (§3.4). ADB averages 18 unit scores and ranges from 0.763 to 0.554 across 21 LLMs. The heatmap highlights near-tied GPT, Qwen, and Claude models (ADB ranks 1, 4, and 5), with an omitted band and the aggregate-weakest model showing the full spread. Within the highlighted trio, the strongest model differs by task: Claude on AT-07, Qwen on AT-11, GPT on AT-16. Shading is normalized per AT to this trio: it shows within-task relative standing, not cross-task differences or absolute ADB scores.

Table 1: Representative evaluation designs. **Contract** means explicit task instructions, expected outputs, and scoring procedures across units; native task specifications without an instruction-level interface receive partial coverage. **Composite** covers multiple requested skills or constraints in one prompt; multi-turn accumulation receives partial coverage. Inspectable outputs include labels, structures, and free text. ✓ = stated coverage, ∂ = partial or differently scoped coverage, — = outside the stated scope. AtomDialBench's bilingual control fixes 80 Chinese and 80 English cases per AT. Dimensions reflect design choices relevant to post-hoc dialogue analysis, not a claim of universal importance.

Resource	Given-dialogue analysis	Contract across units	Inspectable outputs	Composite prompt	Bilingual control coverage
MultiWOZ	✓	∂	✓	—	—
DialoGLUE	∂	∂	✓	—	—
InstructDial	∂	✓	✓	✓	—
CheckList	—	∂	✓	—	—
SKILL-MIX	—	✓	✓	✓	—
MT-Bench-101	—	✓	✓	∂	—
AtomDialBench	✓	✓	✓	✓	✓

Table 2: AT-to-reporting-group mapping. 16 atomic tasks across 5 core reporting groups plus 2 composite evaluation units.

Report Group	ATs	Operations	Output Format
InfoExtract	AT-01–06	Detection/matching/conditions	Binary + evidence
SummaryGen	AT-07, 10	Exemption judgment/summary	Structured/text
ValueInfer	AT-08, 09, 11	Attribute/keyword extraction	Value/list
RelationExt	AT-13, 14	Temporal/entity relations	Categorical
ConstraintCheck	AT-12, 15, 16	Causal/logical/selection	Binary/ranked
MTDA	Composite	Shared-dialogue analysis	Per-AT output
ICDA	Composite	Format/content instruction	Constrained

3 METHOD

We present AtomDialBench in four stages: defining operational atomic tasks (§3.1), selecting and filtering dialogue materials (§3.2), constructing evaluation instances (§3.3), and scoring model responses (§3.4).

3.1 FROM AGGREGATE SCORES TO ATOMIC TASK PROFILES

Operational task definitions. Each AT specifies an instruction, parameters, output schema, and scoring contract. Recurring application requirements define the templates; public datasets supply dialogue material rather than determine the taxonomy. AT-01–06 cover detection, matching, and condition satisfaction; AT-07 exemption judgment; AT-08/09 explicit-attribute extraction and implicit-attribute inference; AT-10 summarization; AT-11 keyword extraction; AT-12–14 causal, temporal, and entity/role relations; AT-15 logical combination; and AT-16 best/top-K candidate selection under a supplied scoring rule. Appendix B gives the contracts and scoring limitations.

Task organization. Table 2 groups the separately executable and scored ATs for reporting, not as a factor model: SummaryGen includes exemption judgment as well as summarization. MTDA combines requirements for one supplied dialogue; ICDA adds instruction constraints. Neither denotes multi-turn interaction or dependency-graph execution. Their distinct case sets permit side-by-side reporting, not a causal estimate of composition.

What a contract makes inspectable. On one AT-06 condition-satisfaction instance, Qwen3.5-4B locates the evidence and marks both required leaf conditions true, yet returns `hit: false`. The structured output localizes an observable inconsistency to the final AND aggregation, beyond merely recording an error; one response does not establish a general reasoning deficit.

3.2 DIALOGUE SELECTION AND QUALITY CONTROL

We collect 339,129 real dialogues from 24 publicly accessible datasets across 9 domains and two languages. The *Dialogue Complexity Index* (DCI) combines length, speaker diversity, turn-length regularity, and information density as a reproducible analytical-complexity proxy, not a general quality measure. We rank within each source and separately by language, using a minimum-turn filter, source caps, and format cleaning rather than a fixed DCI threshold.

Selection retains 13,316 dialogues from 15 datasets while preserving all 9 domains, with recomputed mean DCI of 4.16 (ZH) and 4.90 (EN). These describe the retained set, not a causal effect or population representativeness; the archived pre-filter DCI baseline has unverified input coverage. Source caps limit but do not equalize source shares. Within Chinese, customer-service dialogues (CSDS) contribute 47% of retained material; this concentration reflects DCI-qualified availability, not domain-balanced design. Appendix B.1 gives the formula, thresholds, source shares, baseline provenance, and paired counts (Table 4).

3.3 Data Construction Pipeline

Construction turns selected material into explicit analysis requests. An LLM synthesizes instances from 16 AT-specific JSON schemas; targeted negatives for the nine judgment-type ATs reduce majority-class reliance. Multi-task composition yields 499 MTDA cases, and constraint extension yields 400 ICDA cases. Both draw on a design catalog of five combination types and six constraint families (F1–F6), which does not imply complete coverage in the frozen sets. Deterministic schema checks and LLM factual checks flag instances for repair, separate from the later human scoring study (Appendix B.1).

Each AT contributes 160 instances (80 Chinese, 80 English). For each AT and language, screening targets 40 cases on which both reference models fail and 40 on which both succeed (Qwen3.6-Flash and DeepSeek-V4-Flash-0731). These are construction quotas rather than verified final proportions or difficulty estimates (Appendix B.1). The total is $16 \times 160 + 499 + 400 = 3{,}459$ instances across 18 units; source dialogues may be reused.

3.4 Evaluation Pipeline

Cached inference, deterministic scoring, LLM judging, and aggregation are separate stages. Metrics follow output type: exact match, ROUGE-L, or F_1, with pure set-F_1 for AT-11 and AT-16. For each judged item, the main judge, Qwen3.6-Flash, scores both response orders with reference roles visible and averages the valid scores. This procedure reduces position bias but is not blind assessment. The main judge is excluded from the 21-model comparison; the separate multi-judge ablation uses a different set and includes self-scoring.

Hybrid paths equally weight positive hard and judge scores, using a calibrated judge fallback capped at one when the hard score is zero. AT-07 exemption judgment uses this text-based hybrid path, not ROUGE-L alone; ICDA uses a separate professional judge (Table 3). We aggregate cached final scores.

We first average instance scores within each evaluation unit, then weight the 18 unit means equally to obtain the AtomDialBench (ADB) score. Reporting-group profiles and external-suite averages are shown separately. Appendix G gives statistical details.[1]

4 Experiments

4.1 Setup

We evaluate **21 models from seven families** (ten from the Qwen family, reflecting API availability rather than a balanced developer sample) on the 18 AtomDialBench units. Seventeen models also cover 12 external dialogue-analysis benchmarks (4,471 cases) and four instruction-following (IF) benchmarks (3,353 cases); four have ADB-only coverage. External suites remain separate from ADB. Task-specific scoring follows §3.4; model and external-suite details appear in Appendix G.

4.2 Aggregate Scores and Task Profiles

Aggregate leaders are near-tied despite a wide overall spread. Figure 2 places GPT-5.6-Sol first at 0.763, with the top five within 0.013, while the full range extends to 0.554. Among 20 adjacent pairs, four have uncorrected 95% bootstrap intervals that exclude zero; only the #13–#14 pair remains after Bonferroni correction. The ranking therefore provides a coarse reference, not statistical tiers or evidence of a uniquely superior model (all scores in Table 7).

Task-level diagnostics. Task profiles are uneven, not scaled copies of aggregate rank. Figure 3 shows this across five reporting groups and two composite units. At AT resolution, speech-act detection and condition satisfaction have wide cross-model spreads (0.394 and 0.388), whereas keyword-extraction scores fall within 0.404–0.514 (mean 0.465). This narrow range indicates limited discrimination rather than a ceiling and may also reflect task-design limitations worth revisiting.

[1] r denotes Pearson correlation, ρ Spearman rank correlation, τ Kendall rank correlation, and Δ a mean score difference (treated minus control in training; comparator minus aggregate leader in selection analyses).

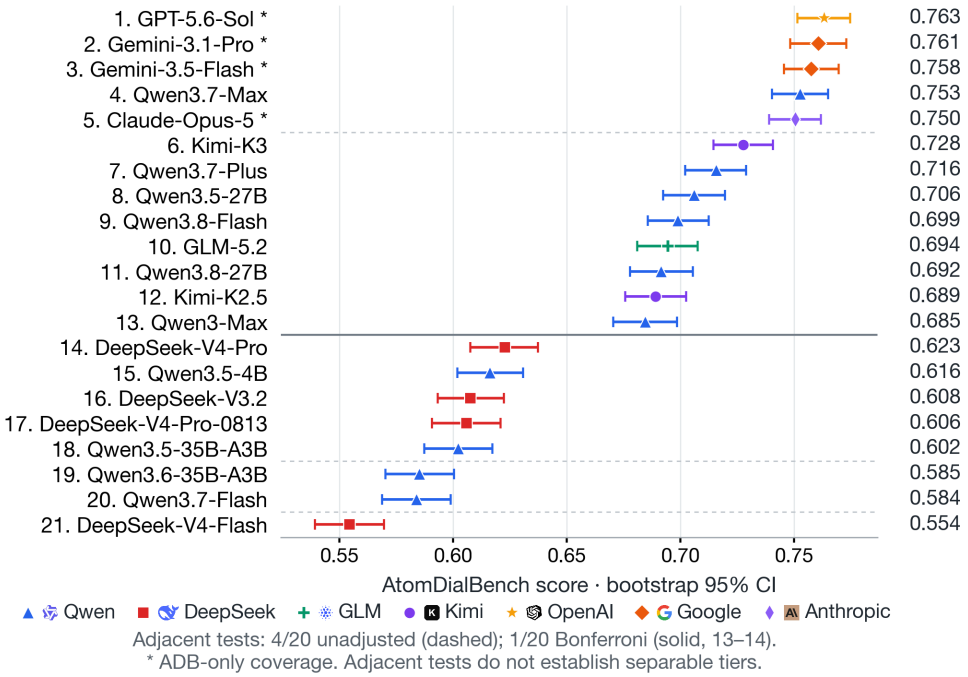

Figure 2: **AtomDialBench ranking with bootstrap 95% CIs** (21 models, 10,000 paired resamples; case indices shared across models within each benchmark). Only the #13–#14 boundary (solid) survives Bonferroni correction; the three dashed boundaries are uncorrected and shown for reference only. These four boundaries do not define statistical tiers.

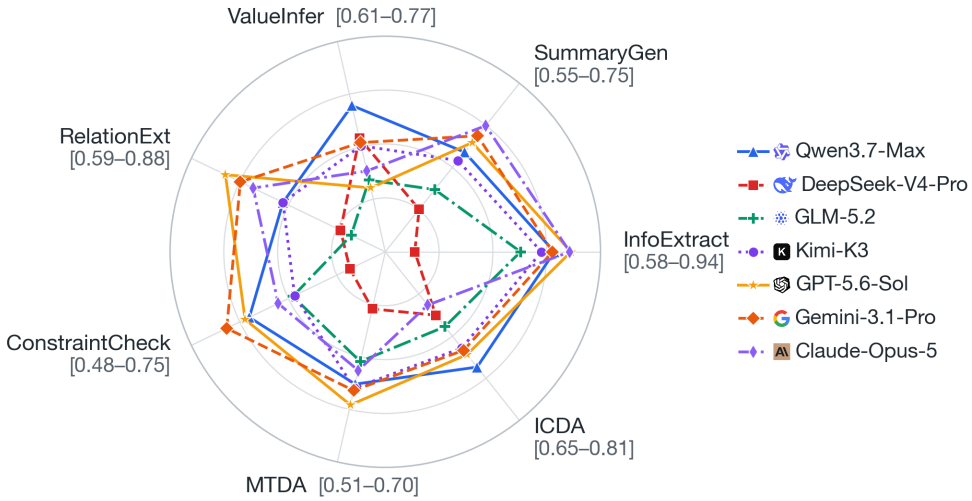

Figure 3: **Task profiles across five reporting groups and two composite units.** Seven representative models (one per family) are identified by markers and line styles. Each axis spans the displayed models' minimum–maximum scores with 0.05 padding, clipped to $[0, 1]$ (Appendix E). Compare models on the same axis only, not radii across axes or polygon areas; the groups are not established independent abilities.

Candidate selection has mean 0.499 and spread 0.237 (Table 8; Figure 10). These contrasts motivate within-operation comparison. Radar axes use separate ranges, so only same-axis model comparisons are meaningful—not cross-axis radii, polygon areas, or independent-ability interpretations.

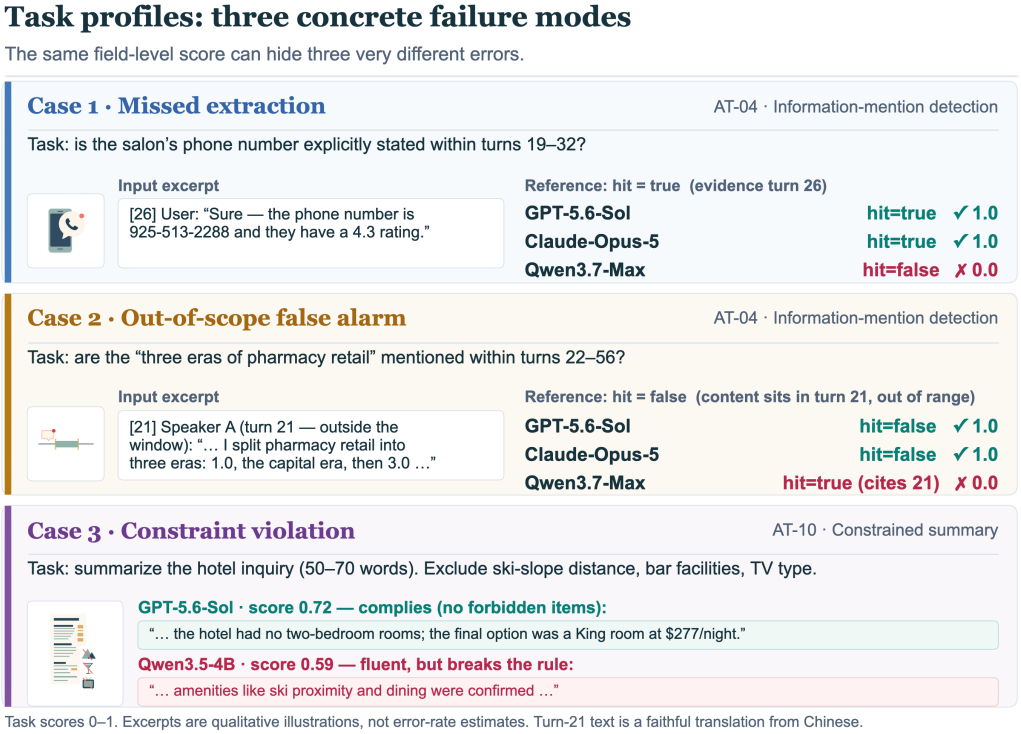

Figure 4: **Three inspectable failure modes revealed by task profiles.** In AT-04, Qwen3.7-Max (i) misses an explicitly stated phone number and (ii) reports an out-of-scope mention by citing turn 21 when the requested window begins at turn 22. In AT-10, Qwen3.5-4B produces a fluent summary that includes a forbidden item, whereas GPT-5.6-Sol excludes it. Scores are on $[0, 1]$; these selected cases are qualitative illustrations, not error-rate estimates or claims of overall model superiority.

4.3 Task-Aware Selection Beyond Aggregate Scores

For each AT, we compare GPT-5.6-Sol against the model with the highest observed mean on that AT, breaking exact ties lexicographically by model identifier. The aggregate leader ranks first or joint first on 7/16 ATs and outside the top three on 7; six models attain first or joint first on at least one AT. Its mean gap is 0.029, concentrated on keyword extraction (AT-11: 0.111, 95% CI $[0.069, 0.152]$, rank 21/21), summarization (AT-10: 0.077), and attribute inference (AT-09: 0.066). Seven of 16 ATs show zero observed gap. Together, these three account for most of the mean. All three intervals exclude zero before correction (Table 6). The intervals are uncorrected and use fixed observed comparators, without adjustment for same-data selection or shared-source-dialogue dependence; they are not held-out selection gains (Appendix E).

Near aggregate parity can therefore coexist with substantial task-specific differences: GPT-5.6-Sol leads Claude-Opus-5 by only 0.013 ADB but by 0.190 on candidate selection; Claude instead leads by 0.051 on exemption judgment. Which profile is preferable depends on the requested operation, not a universal ordering. Figure 4 presents three visible failures—missed information, an out-of-scope false alarm, and a constraint-violating summary—without attributing internal causes. The AT-16 contract check below further tests expression sensitivity (§5).

4.4 Profile Responses to Task-Aligned Fine-Tuning

The Qwen3.5-4B study examines changes in operation-level profiles after task-aligned fine-tuning. Its 21 runs include a general-purpose UltraData-only control. The all-task mixture (Exp1) and control each use approximately 114K examples; single-task and joint-task runs use smaller, unequal

budgets, varying exposure as well as composition (Appendix G.5). No exact train–evaluation duplicates were found by MD5, which does not exclude semantic overlap or pretraining exposure.

All-task training versus a size-matched control (Exp1). Relative to control, all-task training gains $+0.115$ over 16 ATs and $+0.107$ over 18 units, with a non-significant AT-04 decrease (Appendix G.5.1). External dialogue scores improve on 10/12 benchmarks (mean $+0.174$; eight improvements individually significant at uncorrected $\alpha = 0.05$), alongside two significant regressions and no detected IFBench gain (Appendix G.5.4).

Single-task and joint-task training (Exp2–3). Single-task tuning produces selective rather than uniform transfer: in the 16×16 AT block of the transfer matrix, 124/240 off-diagonal changes fall below -0.02 (Figure 13). AT-16-only training gains $+0.138$ on AT-16 but only $+0.007$ on AT-07. The joint AT-07+10+16 run shows positive changes on all three targeted operations: $+0.254$ on AT-07, $+0.028$ on AT-10, and $+0.094$ on AT-16 (Appendix G.5.3). These profiles expose configuration trade-offs, but unequal budgets prevent isolating task-composition effects. The exploratory 27B results lack a matched control and do not establish cross-scale replication (Appendix G.5.6).

4.5 Measurement Evidence

Scoring agreement. Human and multi-judge checks provide complementary scoring evidence. Three independent human annotators with computer-science backgrounds rated 900 stratified responses from two models across 18 units. The human–human ICC(2,1) is 0.918, and the system–human Spearman correlation is 0.697. This supports scoring agreement within the sampled scope, not a human ability baseline or validation across all 21 models (Appendix E). Across eight judges evaluating four ATs, 11 models, and 50 cases per AT, the mean pairwise ICC(2,1) is 0.696, although rankings vary by judge. In a separate 11-model scoring-fusion ablation, six of eight reported configurations retain high rank agreement ($\tau \geq 0.891$); hard-only scoring is the clear exception ($\tau = 0.236$) (Tables 10, 9; Figure 12). The main judge (Qwen3.6-Flash) shares a model family with ten evaluated models; same-family scoring preference has not been quantified and remains a validity limitation.

Task and cross-suite relationships. Task and external-suite relationships are heterogeneous. InfoExtract tasks correlate strongly across 21 models ($\rho \approx 0.84$–0.95), while AT-16 is less aligned (Figure 6). Of 288 cross-suite Spearman tests on 17 models, only AT-14–MWOZ-Int and AT-14–QMSum survive both Bonferroni and Benjamini–Hochberg correction; non-significance does not establish independence. External dialogue and IF rankings have low observed association ($\tau = 0.127$); ICDA has $\tau = 0.418$ with ADB and $\tau = 0.018$ with IF. Because ADB includes ICDA, the former is part–whole association, not independent validation (Appendix D). Together with the training responses, these checks show profile variation that a single aggregate does not express, while stopping short of a validated latent-factor model. Cross-suite correlations exclude the four ADB-only models (ranks 1–3 and 5); including them could shift association estimates.

5 Discussion

Configuration comparisons require task profiles. Within Qwen3.5, moving from 4B dense to 27B dense raises ADB by $+0.090$, while 35B-A3B MoE scores 0.104 below 27B dense. AT-11 and AT-16 nevertheless improve in the latter comparison, illustrating why an aggregate reversal does not imply uniform decline (Appendix H, Figure 18). ADB averages all 18 unit means, not the five reporting-group means. These are descriptive comparisons within one family and one MoE configuration: architecture, routing, and training may differ alongside total and activated parameters. They identify neither a pure scaling effect nor a better capacity predictor; no confidence intervals are supplied for these contrasts.

Score distributions guide inspection, not causal diagnosis. Score distributions vary by task design: InfoExtract is bimodal (zero or high, reflecting binary scoring), AT-11 has the largest zero-score share (42.9%), and AT-10 is dominated by partial scores. Fixed Top-5/Bottom-5 comparisons (by frozen ADB rank) show that zero-rate gaps and high-score share differences are group-specific, not uniform tier effects (Table 28; Appendix I).

AT-16 provides a bounded response-level follow-up. On 1,687 eligible records, candidate-identity normalization raises pooled mean F_1 from 0.6153 to 0.6906, indicating measurable expression sensitivity on this subset. The comparison uses the same records before and after normalization. Eligibility requires reliable input-only candidate catalogues and unique mappings; ambiguous or incomplete pairs are excluded. Of the 1,673 excluded records, 84% (1,407) lack reliable input catalogues, limiting coverage independently of model quality. This analysis does not repair semantic failures or replace the frozen scores and rankings (Appendix I.5).

Language differences are operation-specific. Nineteen of 21 models have higher English means over the 16 ATs; Qwen3-Max and DeepSeek-V4-Pro-0813 are the exceptions. Across models, the largest absolute AT mean gaps (EN$-$ZH) occur on keyword extraction ($+0.104$), entity relation ($+0.098$), and attribute extraction ($+0.060$), while summarization reverses direction (-0.035). Thus an overall English advantage is not shared by every operation. The small signed mean for AT-01–06 ($+0.002$) can conceal opposing differences rather than demonstrate robustness. Each AT has 80 cases per language, but these are different-content samples, not paired translations: neither language causality nor optimization priorities follow from these contrasts.

Appendix J retains per-model values and the unmatched 17:4 developer-region comparison, whose sample composition, configurations, training, and mostly unknown parameter counts preclude geographic or size attribution. Cross-model SDs in the language figures are not confidence intervals. AT-09 component means are also descriptive: conditional fallback prevents decomposing its final language gap from marginal hard-match and judge means.

Practical use. Specify the application's operation and output requirements, compare the corresponding profiles, then inspect outputs and test on application-specific data. Operation-level profiles support targeted model selection for specific dialogue tasks, systematic failure diagnosis across atomic operations, and identification of training gaps that aggregate scores would mask. Score bands, language splits, and training responses provide complementary checks.

Scope and limitations. These checks are not deployment guarantees. Coverage is limited to selected Chinese/English given-dialogue analysis, not ongoing interaction; synthesis and screening can introduce bias. The instance synthesizer (Qwen3.7-Max, rank 4) is also evaluated, and synthesis-related advantage has not been bounded. Human scoring covers two models, so generalization to other families requires further study. Appendix F details sampling, scoring, and transfer limits.

6 Conclusion

We introduce AtomDialBench, a Chinese–English diagnostic benchmark that specifies 16 atomic tasks and two composite units through explicit contracts and provides 3,459 bilingual instances with inspectable outputs. Across 21 LLMs, near-tied aggregate scores hide operation-level rank reversals, motivating a diagnostic sequence: aggregate scores for coarse screening, operation-level profiles for finer comparison, and response inspection for interpreting failures and training responses. The human scoring-agreement study and fine-tuning analyses support interpretation of the scores and their responsiveness to training changes.

Beyond aggregate ranking, the operation-level profiles can help practitioners select models for specific dialogue-analysis needs, diagnose failure patterns on individual tasks, and identify targeted training gaps. The current suite covers selected Chinese and English given-dialogue analysis; broader language and domain coverage, larger human-evaluation scope, and inclusion of additional model families are natural next steps. These extensions should be guided by application requirements rather than scale alone.

AI Use Statement

LLMs assisted dialogue-grounded instance and reference synthesis, quality checks, repair, and scoring: Qwen3.7-Max synthesized AT instances and Qwen3.6-Flash was the main AT judge (§3.4). AI also assisted manuscript writing, language revision, and visuals in the method-overview figure.

Schema/factual checks, human scoring agreement, and the limited AT-07 cache check do not establish complete experimental replay (Reproducibility Details). Human authors retain responsibility for all methods, results, text, and figures, including AI-assisted content. Code, permitted data artifacts, and frozen evaluation records will be released upon acceptance; restricted source dialogues will be documented with access pointers.

REFERENCES

Emre Can Acikgoz, Carl Guo, Suvodip Dey, Akul Datta, Takyoung Kim, Gokhan Tur, and Dilek Hakkani-Tur. TD-EVAL: Revisiting task-oriented dialogue evaluation by combining turn-level precision with dialogue-level comparisons. In *Proceedings of the 26th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pp. 113–132, Avignon, France, August 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.sigdial-1.7/.

Nazanin Alipourfard, Peter G. Fennell, and Kristina Lerman. Can you trust the trend?: Discovering simpson's paradoxes in social data. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pp. 19–27, 2018.

Ge Bai, Jie Liu, Xingyuan Bu, Yancheng He, Jiaheng Liu, Zhanhui Zhou, Zhuoran Lin, Wenbo Su, Tiezheng Ge, Bo Zheng, and Wanli Ouyang. MT-bench-101: A fine-grained benchmark for evaluating large language models in multi-turn dialogues. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7421–7454. Association for Computational Linguistics, 2024.

Andrew M. Bean, Ryan Othniel Kearns, Angelika Romanou, Franziska Sofia Hafner, et al. Measuring what matters: Construct validity in large language model benchmarks. In *Advances in Neural Information Processing Systems 38 (Datasets and Benchmarks Track)*, pp. 19868–19949, 2025.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. MultiWOZ – a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 5016–5026. Association for Computational Linguistics, 2018.

Iñigo Casanueva, Tadas Temčinas, Daniela Gerber, Milica Gasic, and Nikola Mrkšić. Efficient intent detection with dual sentence encoders. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pp. 38–45. Association for Computational Linguistics, 2020.

Yulong Chen, Yang Liu, Liang Chen, and Yue Zhang. DialogSum: A real-life scenario dialogue summarization dataset. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 5062–5074. Association for Computational Linguistics, 2021.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. Chatbot arena: An open platform for evaluating LLMs by human preference. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

Lee J. Cronbach and Paul E. Meehl. Construct validity in psychological tests. *Psychological Bulletin*, 52(4):281–302, 1955.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B. Hashimoto. Length-controlled AlpacaEval: A simple way to debias automatic evaluators. In *Proceedings of the First Conference on Language Modeling*, 2024.

Jack FitzGerald, Christopher Hench, Charith Peris, Scott Mackie, Kay Rottmann, Ana Sanchez, Aaron Nash, Liam Urbach, Vishesh Kakarala, Richa Singh, Swetha Ranganath, Laurie Crist, Misha Britan, Wouter Leeuwis, Gokhan Tur, and Prem Natarajan. MASSIVE: A 1M-example multilingual natural language understanding dataset with 51 typologically-diverse languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4277–4302. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-long.235.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pp. 70–79. Association for Computational Linguistics, 2019.

Prakhar Gupta, Cathy Jiao, Yi-Ting Yeh, Shikib Mehri, Maxine Eskenazi, and Jeffrey Bigham. InstructDial: Improving zero and few-shot generalization in dialogue through instruction tuning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 505–525. Association for Computational Linguistics, December 2022.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, et al. C-eval: A multi-level multi-discipline Chinese evaluation suite for foundation models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 62991–63010, 2023.

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 benchbuilder pipeline. *arXiv preprint arXiv:2406.11939*, 2024.

Xiang Lisa Li, Farzaan Kaiyom, Evan Zheran Liu, Yifan Mai, Percy Liang, and Tatsunori Hashimoto. AutoBencher: Towards declarative benchmark construction. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=ymt4crbbXh.

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. AlpacaEval: An automatic evaluator of instruction-following models. https://github.com/tatsu-lab/alpaca_eval, 2023.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 986–995. Asian Federation of Natural Language Processing, 2017.

Percy Liang, Rishi Bommasani, Tony Lee, et al. Holistic evaluation of language models. *Transactions on Machine Learning Research*, 2023. URL https://openreview.net/forum?id=iO4LZibEqW.

Zheng Wei Lim, Ekaterina Vylomova, Trevor Cohn, and Charles Kemp. Simpson's paradox and the accuracy-fluency tradeoff in translation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 92–103. Association for Computational Linguistics, 2024.

Bill Yuchen Lin, Yuntian Deng, Khyathi Chandu, Faeze Brahman, Abhilasha Ravichander, Valentina Pyatkin, Nouha Dziri, Ronan Le Bras, and Yejin Choi. WildBench: Benchmarking LLMs with challenging tasks from real users in the wild. In *Proceedings of the Thirteenth International Conference on Learning Representations*, 2025.

Chin-Yuen Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81. Association for Computational Linguistics, 2004.

Haitao Lin, Liqun Ma, Junnan Zhu, Lu Xiang, Yu Zhou, Jiajun Zhang, and Chengqing Zong. CSDS: A fine-grained Chinese dataset for customer service dialogue summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 4436–4451. Association for Computational Linguistics, 2021.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2511–2522. Association for Computational Linguistics, 2023.

Yuxuan Liu, Tianchi Yang, Shaohan Huang, Zihan Zhang, Haizhen Huang, Furu Wei, Weiwei Deng, Feng Sun, and Qi Zhang. Calibrating LLM-based evaluator. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 2638–2656. ELRA and ICCL, 2024.

Shikib Mehri and Maxine Eskenazi. Unsupervised evaluation of interactive dialog with DialoGPT. In *Proceedings of the 21st Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pp. 225–235. Association for Computational Linguistics, 2020a.

Shikib Mehri and Maxine Eskenazi. USR: An unsupervised and reference free evaluation metric for dialog. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 681–707. Association for Computational Linguistics, 2020b.

Shikib Mehri, Mihail Eric, and Dilek Hakkani-Tur. Dialoglue: A natural language understanding benchmark for task-oriented dialogue. *arXiv preprint arXiv:2009.13570*, 2020.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2023.

Jiao Ou, Junda Lu, Che Liu, Yihong Tang, Fuzheng Zhang, Di Zhang, and Kun Gai. DialogBench: Evaluating LLMs as human-like dialogue systems. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 6137–6170, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.341. URL https://aclanthology.org/2024.naacl-long.341/.

ChaeHun Park, Minseok Choi, Dohyun Lee, and Jaegul Choo. PairEval: Open-domain dialogue evaluation metric with pairwise comparisons. In *Proceedings of the First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=y6aGT625Lk.

Philip M. Podsakoff, Scott B. MacKenzie, Jeong-Yeon Lee, and Nathan P. Podsakoff. Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology*, 88(5):879–903, 2003.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 527–536. Association for Computational Linguistics, 2019.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4902–4912. Association for Computational Linguistics, 2020.

David Schlangen. Targeting the benchmark: On methodology in current natural language processing research. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pp. 670–674. Association for Computational Linguistics, 2021.

Elizabeth Shriberg, Raj Dhillon, Sonali Bhagat, Jeremy Ang, and Hannah Carvey. The ICSI meeting recorder dialog act (MRDA) corpus. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue*, pp. 97–100. Association for Computational Linguistics, 2004.

Edward H. Simpson. The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 13(2):238–241, 1951.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*, 2023. URL https://openreview.net/forum?id=uyTL5Bvosj.

Joe Stacey, Pasquale Minervini, Haim Gaifman, and Marek Rei. Atomic inference for NLI with generated facts as atoms. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 10188–10204. Association for Computational Linguistics, 2024.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26 (3):339–374, 2000. URL https://aclanthology.org/J00-3003/.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems 32*, 2019.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9440–9450. Association for Computational Linguistics, 2024a.

Yidong Wang, Zhuohao Yu, Zhengran Zeng, Linyi Yang, Cunxiang Wang, Hao Chen, Chaoya Jiang, Rui Xie, Jindong Wang, Xing Xie, Wei Ye, Shikun Zhang, and Yue Zhang. Pandalm: An automatic evaluation benchmark for LLM instruction tuning optimization. In *The Twelfth International Conference on Learning Representations*, 2024b.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13484–13508. Association for Computational Linguistics, 2023.

Colin White, Samuel Dooley, Manley Roberts, Arka Pal, Ben Feuer, Siddhartha Jain, Ravid Shwartz-Ziv, Neel Jain, Khalid Saifullah, Siddartha Naidu, Chinmay Hegde, Yann LeCun, Tom Goldstein, Willie Neiswanger, and Micah Goldblum. LiveBench: A challenging, contamination-limited LLM benchmark. In *Proceedings of the Thirteenth International Conference on Learning Representations*, 2025.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. Wizardlm: Empowering large pre-trained language models to follow complex instructions. In *The Twelfth International Conference on Learning Representations*, 2024.

Seonghyeon Ye, Doyoung Kim, Sungdong Kim, Hyeonbin Hwang, Seungone Kim, Yongrae Jo, James Thorne, Juho Kim, and Minjoon Seo. FLASK: Fine-grained language model evaluation based on alignment skill sets. *arXiv preprint arXiv:2307.10928*, 2023.

Dian Yu, Kai Sun, Claire Cardie, and Dong Yu. Dialogue-based relation extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4927–4940. Association for Computational Linguistics, 2020.

Dingli Yu, Simran Kaur, Arushi Gupta, Jonah Brown-Cohen, Anirudh Goyal, and Sanjeev Arora. SKILL-MIX: A flexible and expandable family of evaluations for AI models. In *The Twelfth International Conference on Learning Representations*, 2024.

Fei Yuan, Longtu Zhang, Bojun Huang, and Yaobo Liang. Simpson's bias in NLP training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 14276–14283, 2021a.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. BARTscore: Evaluating generated text as text generation. In *Advances in Neural Information Processing Systems*, volume 34, 2021b.

Hanlei Zhang, Hua Xu, Xin Wang, Qianrui Zhou, Shaojie Zhao, and Jiayan Teng. MIntRec: A new dataset for multimodal intent recognition. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 1688–1697, 2022. doi: 10.1145/3503161.3547906.

Jianguo Zhang, Kun Qian, Zhiwei Liu, Shelby Heinecke, Rui Meng, Ye Liu, Zhou Yu, Huan Wang, Silvio Savarese, and Caiming Xiong. DialogStudio: Towards richest and most diverse unified dataset collection for conversational AI. In *Findings of the Association for Computational Linguistics: EACL 2024*, pp. 2299–2315. Association for Computational Linguistics, 2024a.

Mian Zhang, Lifeng Jin, Linfeng Song, Haitao Mi, and Dong Yu. Inconsistent dialogue responses and how to recover from them. In *Findings of the Association for Computational Linguistics: EACL 2024*, pp. 220–230, St. Julian's, Malta, March 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-eacl.16. URL https://aclanthology.org/2024.findings-eacl.16/.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2204–2213. Association for Computational Linguistics, 2018.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. BERTscore: Evaluating text generation with BERT. In *International Conference on Learning Representations*, 2020.

Xinghua Zhang, Haiyang Yu, Yongbin Li, Minzheng Wang, Longze Chen, and Fei Huang. The imperative of conversation analysis in the era of LLMs: A survey of tasks, techniques, and trends. *arXiv preprint arXiv:2409.14195*, 2024c.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, pp. 46595–46623, 2023.

Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan Awadallah, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, and Dragomir Radev. QMSum: A new benchmark for query-based multi-domain meeting summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 5905–5921. Association for Computational Linguistics, 2021.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. Towards a unified multi-dimensional evaluator for text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 2023–2038. Association for Computational Linguistics, 2022.

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

Kaijie Zhu, Jiaao Chen, Jindong Wang, Neil Zhenqiang Gong, Diyi Yang, and Xing Xie. Dyval: Graph-informed dynamic evaluation of large language models for reasoning tasks. In *The Twelfth International Conference on Learning Representations*, 2024.

Lianghui Zhu, Xinggang Wang, and Xinlong Wang. Judgelm: Fine-tuned large language models are scalable judges. In *The Thirteenth International Conference on Learning Representations*, 2025.

A Extended Related Work

This appendix provides the full survey deferred from the compressed §2.

A.1 Dialogue Evaluation Benchmarks

Task-specific resources make distinct aspects of dialogue behavior observable. MultiWOZ supports state tracking and slot-oriented analysis (Budzianowski et al., 2018); SAMSum evaluates summaries (Gliwa et al., 2019); DialogRE targets relation extraction (Yu et al., 2020); MELD and DailyDialog provide emotion and dialogue-act annotations (Poria et al., 2019; Li et al., 2017). MIntRec supplies multimodal intent-recognition data (Zhang et al., 2022), while PersonaChat concerns persona-grounded interaction (Zhang et al., 2018). These resources should not all be treated as interchangeable analysis benchmarks.

DialoGLUE groups seven NLU datasets, InstructDial organizes 48 dialogue tasks, and DialogStudio curates over 80 datasets (Mehri et al., 2020; Gupta et al., 2022; Zhang et al., 2024a). More specialized sources include CSDS, QMSum, and DialogSum for summarization (Lin et al., 2021; Zhong et al., 2021; Chen et al., 2021); Switchboard and ICSI for dialogue acts (Stolcke et al., 2000; Shriberg et al., 2004); and BANKING77 and MASSIVE for intent analysis (Casanueva et al., 2020; FitzGerald et al., 2023).

USR, FED, and PairEval assess response quality (Mehri & Eskenazi, 2020b;a; Park et al., 2024). MT-Bench-101 offers fine-grained multi-turn evaluation over 13 tasks, 1,388 dialogues, and 4,208 turns (Bai et al., 2024). Its interactive setting differs from analyzing a supplied dialogue, but its task organization is directly relevant to diagnostic evaluation.

A.2 Multi-task Evaluation Frameworks

BIG-Bench, HELM, MMLU, SuperGLUE, and C-Eval provide broad task or metric coverage (Srivastava et al., 2023; Liang et al., 2023; Hendrycks et al., 2021; Wang et al., 2019; Huang et al., 2023). MT-Bench, WildBench, Arena-Hard, and AlpacaEval evaluate assistant responses and interactions (Zheng et al., 2023; Lin et al., 2025; Li et al., 2024; Dubois et al., 2024). LiveBench emphasizes current evaluation material (White et al., 2025), and IFEval tests verifiable instructions (Zhou et al., 2023). Their distinct goals motivate complementary comparisons, not a blanket claim that prior frameworks lack fine-grained tests.

A.3 LLM-as-Judge

MT-Bench and AlpacaEval study model-based judging, while Chatbot Arena provides human preference comparisons (Zheng et al., 2023; Li et al., 2023; Chiang et al., 2024). JudgeLM and PandaLM train dedicated evaluators (Zhu et al., 2025; Wang et al., 2024b). Position bias and calibration remain concerns (Wang et al., 2024a; Liu et al., 2024). Our reference-aware position swap is intended to mitigate order effects; the restricted eight-judge ablation documents both agreement and ranking variation, rather than proving bias removal.

A.4 Task Decomposition in NLP Evaluation

CheckList organizes behavioral tests by capability and test type (Ribeiro et al., 2020). Atomic Inference for NLI decomposes inference steps (Stacey et al., 2024); SKILL-MIX tests skill combinations (Yu et al., 2024); FActScore assesses atomic factual claims (Min et al., 2023). AtomDialBench adopts an operational task unit for dialogue analysis, with separate instructions, outputs, and scoring. This differs from asserting that cognition itself is atomic.

A.5 Evaluation Metrics and Scoring Design

ROUGE, exact match, and F_1 provide reproducible comparisons with references (Lin, 2004). BERTScore, BARTScore, UniEval, and G-Eval use learned representations or evaluators (Zhang et al., 2020; Yuan et al., 2021b; Zhong et al., 2022; Liu et al., 2023). FLASK explicitly assesses

12 skills (Ye et al., 2023). We combine task-specific scoring paths with agreement checks, while retaining limitations such as AT-07's text-based scoring of an exemption decision.

A.6 Automatic Benchmark Construction

Self-Instruct and WizardLM/Evol-Instruct demonstrate instruction synthesis and evolution (Wang et al., 2023; Xu et al., 2024); DyVal generates structured evaluation examples (Zhu et al., 2024); AutoBencher targets evaluation gaps (Li et al., 2025). Our construction uses dialogue-grounded templates, schema checks, and LLM review. The separate human study checks scoring agreement, not every synthesized instance.

A.7 Aggregation and Score Interpretation

Aggregation effects have been studied in social data, NLP training data, and machine translation evaluation (Simpson, 1951; Alipourfard et al., 2018; Yuan et al., 2021a; Lim et al., 2024). Our ranking reversals need not meet the formal conditions of Simpson's paradox. Construct validity instead asks what evidence supports a score interpretation (Cronbach & Meehl, 1955; Schlangen, 2021; Bean et al., 2025); it is a continuing evaluation responsibility, not a property we assume absent from all previous benchmarks.

B Atomic Tasks and Reporting Groups

Table 3: Hard match metric assignment per AT type (deferred from §3.4).

AT Type	Hard Match	Scoring Target
Hit/Relation (AT-01–06, 12–15)	Exact match	Binary/categorical fields
Exemption/Summary (AT-07, 10)	ROUGE-L	Hard component of hybrid path
Value (AT-08, 09)	Fact extraction	ROUGE-L fallback
Keywords (AT-11)	Set F_1	Extracted keywords
Candidate selection (AT-16)	$F_1 = 2PR/(P+R)$	Selected candidate set
Multi-task (MTDA)	Per-AT hard match	Combined with judge
Instruction (ICDA)	None (judge only)	Instruction rubric

The five group names organize reporting, not independent latent constructs. Table 3 lists deterministic components rather than all final scores. AT-07 exemption judgment shares a text-based hybrid path with AT-10: the cached final `score`, not `hard_score` alone, is aggregated. The surviving judge-only fallback can apply the AT-07 factor of 1.18 before capping at one. This path limits interpretation as a pure decision measure. AT-16 set F_1 assesses selected candidates, not every supplied score or rank; ICDA uses a separate professional-judge rubric.

Information Extraction (AT-01–AT-06). The operations are speech-act detection, emotional-state detection, intent matching, information-mention detection, label matching, and condition satisfaction, respectively. They return binary judgments with task-specific evidence fields. Several require semantic interpretation or combining conditions, rather than literal entity retrieval alone.

Summary Generation (AT-07, AT-10). AT-07 determines whether the specified exemption condition applies and returns exemption and leaf-level judgments. AT-10 produces a summary under the supplied scope, length, style, and content requirements. Sharing this reporting group does not make AT-07 a summarization task.

Value Inference (AT-08, AT-09, AT-11). AT-08 extracts an explicitly stated attribute value; AT-09 infers an implicit participant attribute, with default handling; AT-11 extracts keywords or phrases under type and quantity requirements. Only AT-09 necessarily targets an unstated attribute.

Relation Extraction (AT-13, AT-14). AT-13 judges temporal relations between events (before, after, overlap, or unknown). AT-14 selects an entity/role relation type from the supplied candidates, including an unknown fallback.

Table 4: Dialogue counts and coverage before and after filtering: fewer dialogues and source datasets, with all 9 domains retained.

Metric	Before	After
# Dialogues	339,129	13,316
# Source datasets	24	15
# Domains	9	9

Constraint Checking (AT-12, AT-15, AT-16). AT-12 judges causal relations between events; AT-15 combines previously obtained task results through logical operators; AT-16 selects one or top-K candidates under a supplied scoring rule, with bonuses, penalties, normalization, and tie-breaking as specified. Candidate selection is the observable operation; generic constraint satisfaction is not a synonym for this contract.

Composite Units. MTDA (499 cases after deduplication) combines multiple AT analysis requirements for one shared dialogue in a single prompt. ICDA (400 cases) adds instruction constraints to given-dialogue analysis. The combination and constraint catalogs below describe construction options, not dependency-graph execution or verified complete and balanced coverage of the frozen cases.

B.1 Data Construction Details

Dialogue Complexity Index (DCI). The multiplicative composite used for dialogue selection (§3.2):

$$\mathrm{DCI} = \log\big(1 + N \cdot \bar{L}\big) \;\cdot\; D \;\cdot\; R \;\cdot\; I,$$

where N is the turn count, $\bar{L}$ the mean turn length, $D = \log(1{+}S)/\log 3$ a speaker-diversity factor (S = number of distinct speakers), $R = 1/(1{+}\mathrm{CV})$ a rhythm-regularity factor (CV = coefficient of variation of turn lengths), and I the fraction of turns exceeding 5 characters (information-density ratio). Chinese and English dialogues are scored separately to respect cross-lingual length norms.

Dialogue selection. We apply a minimum of 25 turns, then retain each source's highest-DCI dialogues under per-dataset caps (ZH 3,000; EN 2,000), followed by format cleaning of speaker labels and tokenization artifacts. DCI ranks dialogues within each source; it is neither a fixed admission threshold nor a general quality measure. Table 4 reports paired changes in counts and coverage.

Archived DCI baseline. The archived pre-filter means, 3.49 (ZH) and 3.99 (EN), are reported with the earlier 22-source count labels: 299,129 dialogues (ZH 89,513; EN 209,616). The inventory later expanded to 24 sources and 339,129 dialogues (ZH 89,513; EN 249,616). Although the archive specifies the same DCI formula, the means' actual input coverage, including whether they cover the 24-source pool, cannot be verified. They therefore do not establish same-population quantitative gains from selection.

Selection statistics. Using the same DCI formula (§3.2), independent recomputation gives retained-set means of 4.16 (ZH) and 4.90 (EN). Retained dialogues average 35.4 (ZH) and 29.9 (EN) turns. Per-source caps reduce single-source domination but do not equalize shares: within-language, the largest source is CSDS at 47% (ZH) and SGD at 29% (EN). These shares reflect availability of long, high-DCI dialogues per source and do not guarantee population representativeness.

AT output schemas. Each AT has task-specific required fields. AT-01–03 require `hit`, `evidence_turns`, and `evidence_quote`; AT-04 instead adds `matched_text`, AT-05 requires only the hit and evidence-turn fields, and AT-06 adds `leaves`. AT-07 requires `exempted`, `matched_exemption_id`, `evidence_turns`, and `leaf_results`. AT-13 returns `relation`, both anchor turns, and evidence turns. AT-16 returns a `top_k` list with candidate, score, evidence-turn, and rank fields. Full definitions for all 16 ATs are included in the repository.

Pipeline stages. *Step 1 (AT synthesis)* prompts an LLM with the source dialogue, the AT definition, and the target output schema; the prompt includes one worked example per AT type. Temperature is set to 0.7; each dialogue generates instances for all applicable ATs. *Step 2 (negative balancing)* applies to the 9 binary-judgment ATs: for each positive instance, a targeted negative is synthesized by perturbing the condition while preserving dialogue context. *Step 3 (multi-task composition)* combines multiple AT analysis requirements into one prompt about a shared dialogue, yielding 499 MTDA cases after deduplication. The five design types are sequential, nested, conditional, contrastive, and mixed request layouts; these names do not imply execution topologies or a fixed AT count per case. *Step 4 (constraint extension)* yields 400 ICDA cases. Its design catalog offers six families (F1 format; F2 content; F3 structural; F4 editing; F5 multilingual formatting; F6 complex composites) and up to three difficulty levels; the catalog does not establish complete or balanced family-by-level coverage in the frozen 400 cases.

Quality-gate rules. The deterministic layer enforces: (i) JSON well-formedness and AT-schema field completeness, (ii) truncation-marker detection, (iii) semantic consistency checks (evidence turns within the specified window; `hit` consistent with `leaf_results` for compound ATs), (iv) trivial-relation filtering for AT-14, and (v) vendor-QC pass-through for Step 4 records. The LLM-based review layer then samples failed-gate instances for factual verification, with automatic repair attempted before final rejection.

Difficulty-screen targets and fallback. For each AT and language, the construction screen targets 40 cases on which both Qwen3.6-Flash and DeepSeek-V4-Flash-0731 fail and 40 on which both succeed. These are task-dependent heuristic outcomes, not final evaluation scores: checks include decision/relation-field equality, keyword overlap, character-set overlap for some text/value outputs, and first-candidate agreement for selection. Parse failures count as non-successes; records with API-error responses are skipped. A dual-failure shortfall is supplemented with single-failure cases; a dual-success shortfall uses unused cases from the full same-language pool, followed by a remaining-pool fallback if the AT total is still short. Thus, 40/40 denotes target quotas, not a verified final hard/easy split. The saved exports do not recover the final proportions in every pool.

Coverage. Source and language metadata support inspection of the constructed instances. Evaluation fixes 80 Chinese and 80 English instances per AT; this does not imply that every AT–source–language combination is populated or that the sample represents deployment frequencies.

C Extended Exploratory Correlation Analysis

The all-pairs scan covers $\binom{34}{2} = 561$ comparisons among 18 SynthAT units and 16 external benchmarks. This is distinct from the 288 cross-suite Spearman tests in §4.5. The associations below use $N = 17$ full-coverage models and uncorrected p-values.

Summarization across suites. AT-10 vs. CSDS: Spearman $\rho = -0.419$, $p = 0.094$ (not significant at $\alpha = 0.05$).

Consistency vs. slot extraction. CIDER vs. MultiWOZ slot extraction: Spearman $\rho = -0.553$, $p = 0.021$ (uncorrected).

Summarization vs. candidate selection. SAMSum vs. AT-16: Spearman $\rho = -0.439$, $p = 0.078$ (not significant).

Within instruction following. IFBench vs. GuideBench: Spearman $\rho = -0.394$, $p = 0.118$ (not significant).

Positive contrasts. Within InfoExtract: AT-01 $\leftrightarrow$ AT-06 ($\rho = +0.898$), AT-04 $\leftrightarrow$ AT-06 ($\rho = +0.908$), AT-01 $\leftrightarrow$ AT-05 ($\rho = +0.886$).

Interpretation. Together with AT-16–CSDS, the reported examples include two negative associations below the uncorrected 0.05 threshold and three weaker negative estimates. They remain

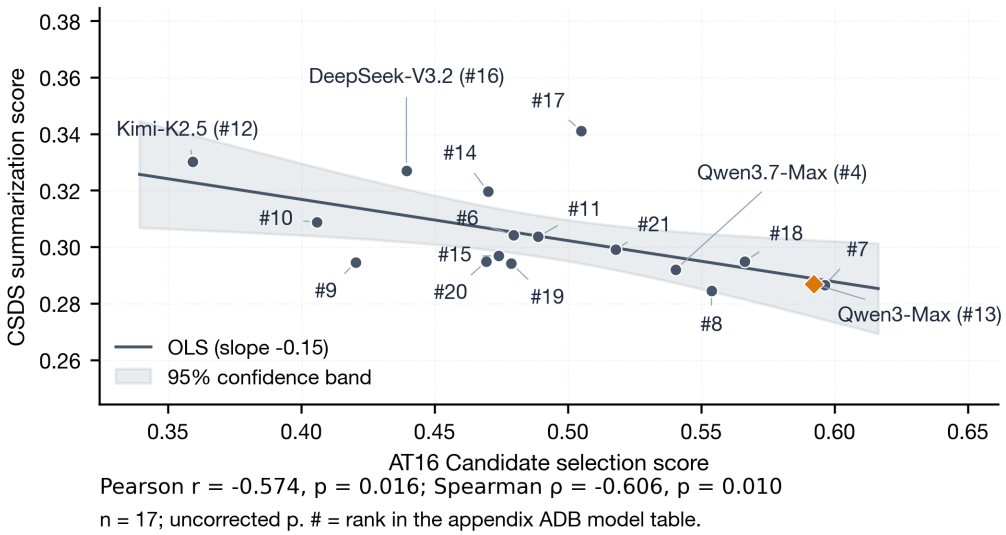

Figure 5: **Exploratory association between candidate selection (AT-16) and customer-service summarization (CSDS).** All 17 points are paired observations; numbers identify ADB ranks in Table 7, not scatter ordering. OLS is shown with a pointwise 95% confidence band for the fitted mean; Pearson $r = -0.574$, $p = 0.016$ (uncorrected). This association does not survive correction over the 288 cross-suite tests.

exploratory findings from a broad scan, not evidence for a shared precision–fluency mechanism. The reported pointwise bootstrap interval for AT-16–CSDS does not address this multiplicity; that pair does not survive either saved correction over the 288 cross-suite tests.

D Cross-Suite Relationships

Table 5: Observed cross-suite rank correlations (Kendall τ). Low association in this model sample does not establish independent abilities.

Comparison	τ	p	**Interpretation**
Ext-Dialogue ↔ IF	0.127	0.648	Low association
ADB → ICDA	0.418	—	Positive association
Ext-Dialogue → ICDA	0.236	—	Weaker association
IF → ICDA	0.018	—	Near-zero association

E Additional Figures and Tables from Main Text

Fixed-comparator selection protocol. Competition ranks retain exact ties in unrounded observed means. GPT-5.6-Sol is first outright on AT-03, AT-05, AT-13, and AT-14, and joint first on AT-01, AT-04, and AT-06. Six models attain first or joint first across the ATs. For bootstrap comparisons only, ties are broken lexicographically by model identifier, yielding five distinct fixed comparators; this is a different count from the number of top-scoring models. We use 10,000 draws with seed 42, sharing resampled case indices between the aggregate leader and comparator within each unit. When they are the same model, the paired gap is identically zero. The equal-weight mean across 16 AT gaps is a descriptive summary of task-specific metrics. These intervals condition on the observed model choices and case-level resampling: they do not account for selection optimism or dependence among instances sharing a source dialogue, and are not estimates of held-out selection-policy performance.

Table 6: **Fixed-comparator selection gaps for the aggregate leader (GPT-5.6-Sol) on every evaluation unit.** For each unit the comparator is the model with the highest observed mean (ties broken lexicographically); *Gap* is the paired-bootstrap difference (10,000 draws, seed 42, shared case indices) and CIs are 95% percentile intervals. *Rank* is the aggregate leader's competition rank among the 21 evaluated models on that unit. Intervals that exclude zero are individually reliable but uncorrected for multiple comparisons; they condition on fixed observed comparators and are not held-out selection-policy estimates.

Unit	Type	Fixed best model	Best	Leader	Gap	95% CI	Rank
AT-01	atomic	Claude-Opus-5	0.894	0.894	0.000	$[-0.044, 0.044]$	1
AT-02	atomic	Claude-Opus-5	0.875	0.838	0.037	$[-0.019, 0.094]$	4
AT-03	atomic	GPT-5.6-Sol	0.894	0.894	0.000	$[0.000, 0.000]$	1
AT-04	atomic	Claude-Opus-5	0.931	0.931	0.000	$[-0.031, 0.031]$	1
AT-05	atomic	GPT-5.6-Sol	0.863	0.863	0.000	$[0.000, 0.000]$	1
AT-06	atomic	Claude-Opus-5	0.919	0.919	0.000	$[-0.025, 0.025]$	1
AT-07	atomic	Claude-Opus-5	0.765	0.714	0.051	$[0.017, 0.085]$	3
AT-08	atomic	Qwen3.7-Plus	0.904	0.882	0.021	$[-0.006, 0.050]$	6
AT-09	atomic	Qwen3.7-Max	0.745	0.679	0.066	$[0.033, 0.100]$	12
AT-10	atomic	Qwen3.7-Max	0.724	0.647	0.077	$[0.065, 0.089]$	20
AT-11	atomic	Qwen3.7-Max	0.514	0.404	0.111	$[0.069, 0.152]$	21
AT-12	atomic	Claude-Opus-5	0.800	0.750	0.050	$[0.000, 0.100]$	6
AT-13	atomic	GPT-5.6-Sol	0.856	0.856	0.000	$[0.000, 0.000]$	1
AT-14	atomic	GPT-5.6-Sol	0.800	0.800	0.000	$[0.000, 0.000]$	1
AT-15	atomic	Gemini-3.1-Pro	0.719	0.694	0.025	$[-0.006, 0.062]$	3
AT-16	atomic	Qwen3.7-Plus	0.596	0.577	0.019	$[-0.035, 0.076]$	4
MTDA	comp.	GPT-5.6-Sol	0.649	0.649	0.000	$[0.000, 0.000]$	1
ICDA	comp.	Gemini-3.5-Flash	0.763	0.750	0.013	$[-0.003, 0.029]$	3

Radar scaling. In Figure 3, let s_{mj} be model m's raw score on reporting axis j, averaging the constituent ATs where applicable (Table 2). For the seven displayed models $\mathcal{M}_7$, each axis uses

$$l_j = \max(0, \min_{m \in \mathcal{M}_7} s_{mj} - 0.05), \qquad u_j = \min(1, \max_{m \in \mathcal{M}_7} s_{mj} + 0.05), \qquad r_{mj} = \frac{s_{mj} - l_j}{u_j - l_j}.$$

The axis labels show the raw bounds; radius supports comparison between models on the same axis only. Cross-axis radii and polygon areas are not comparable. The right-side legend lists seven models, with matching markers and line styles. Figure 11 instead uses a common raw-score range of $[0, 1]$.

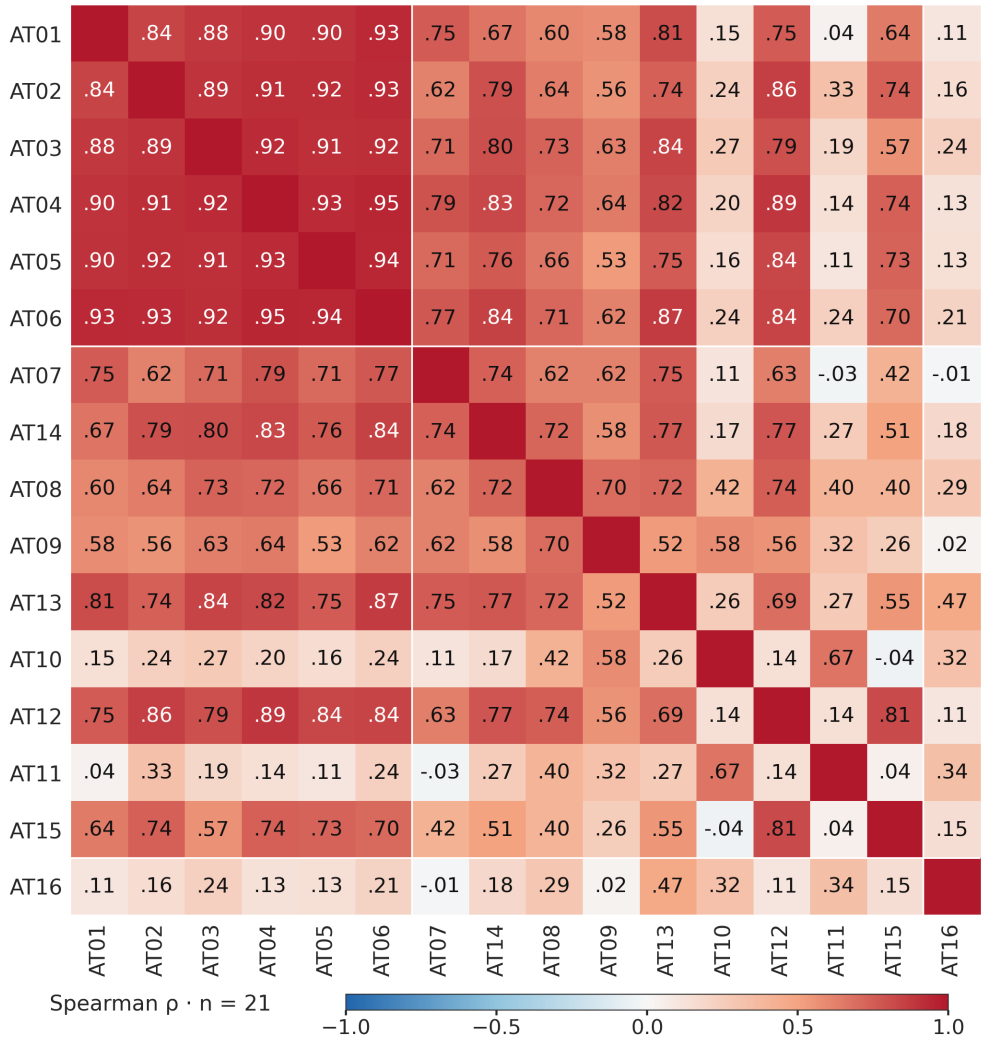

Figure 6: **AT correlation spectrum (N=21 models).** The 16 × 16 matrix shows Spearman ρ on a fixed $[-1, 1]$ color scale centered at zero; AT identifiers refer to Appendix B. InfoExtract tasks are strongly associated, AT-16 candidate selection is less aligned, and other tasks have intermediate associations. These regions are descriptive, not statistical tiers or independent constructs; the unit diagonal denotes self-correlation only.

Table 7: Full AtomDialBench (SynthAT-18) ranking of all 21 evaluated models. ADB averages the 18 SynthAT evaluation units (16 ATs + MTDA + ICDA); ⋆ marks the 4 ADB-only external models. Figure 2 reports uncertainty; adjacent-pair tests do not establish group-level equivalence.

Rank	Model	ADB	MTDA	ICDA	AT-01–16	Family
1	GPT-5.6-Sol⋆	0.7633	0.648	0.750	0.771	OpenAI
2	Gemini-3.1-Pro⋆	0.7607	0.636	0.746	0.769	Google
3	Gemini-3.5-Flash⋆	0.7576	0.626	0.763	0.765	Google
4	Qwen3.7-Max	0.7527	0.630	0.762	0.760	Qwen
5	Claude-Opus-5⋆	0.7505	0.618	0.704	0.762	Anthropic
6	Kimi-K3	0.7276	0.634	0.744	0.732	Kimi
7	Qwen3.7-Plus	0.7157	0.615	0.715	0.722	Qwen
8	Qwen3.5-27B	0.7060	0.607	0.727	0.711	Qwen
9	Qwen3.8-Flash	0.6989	0.617	0.710	0.703	Qwen
10	GLM-5.2	0.6944	0.611	0.724	0.698	GLM
11	Qwen3.8-27B	0.6916	0.623	0.733	0.693	Qwen
12	Kimi-K2.5	0.6890	0.613	0.713	0.692	Kimi
13	Qwen3-Max	0.6846	0.602	0.704	0.689	Qwen
14	DeepSeek-V4-Pro	0.6227	0.564	0.714	0.621	DeepSeek
15	Qwen3.5-4B	0.6163	0.528	0.663	0.619	Qwen
16	DeepSeek-V3.2	0.6076	0.545	0.708	0.605	DeepSeek
17	DeepSeek-V4-Pro-0813	0.6058	0.568	0.706	0.602	DeepSeek
18	Qwen3.5-35B-A3B	0.6024	0.544	0.702	0.600	Qwen
19	Qwen3.6-35B-A3B	0.5852	0.556	0.705	0.580	Qwen
20	Qwen3.7-Flash	0.5840	0.558	0.699	0.578	Qwen
21	DeepSeek-V4-Flash-0731	0.5545	0.522	0.679	0.549	DeepSeek

External benchmark panels. Figures 7, 8, and 9 cover all 12 external dialogue benchmarks (17 full-coverage models, 204 scores). Every panel uses the same model order, descending by mean over the 12 benchmarks, and the same 0–1 score axis. Panels are read left-to-right, then top-to-bottom.

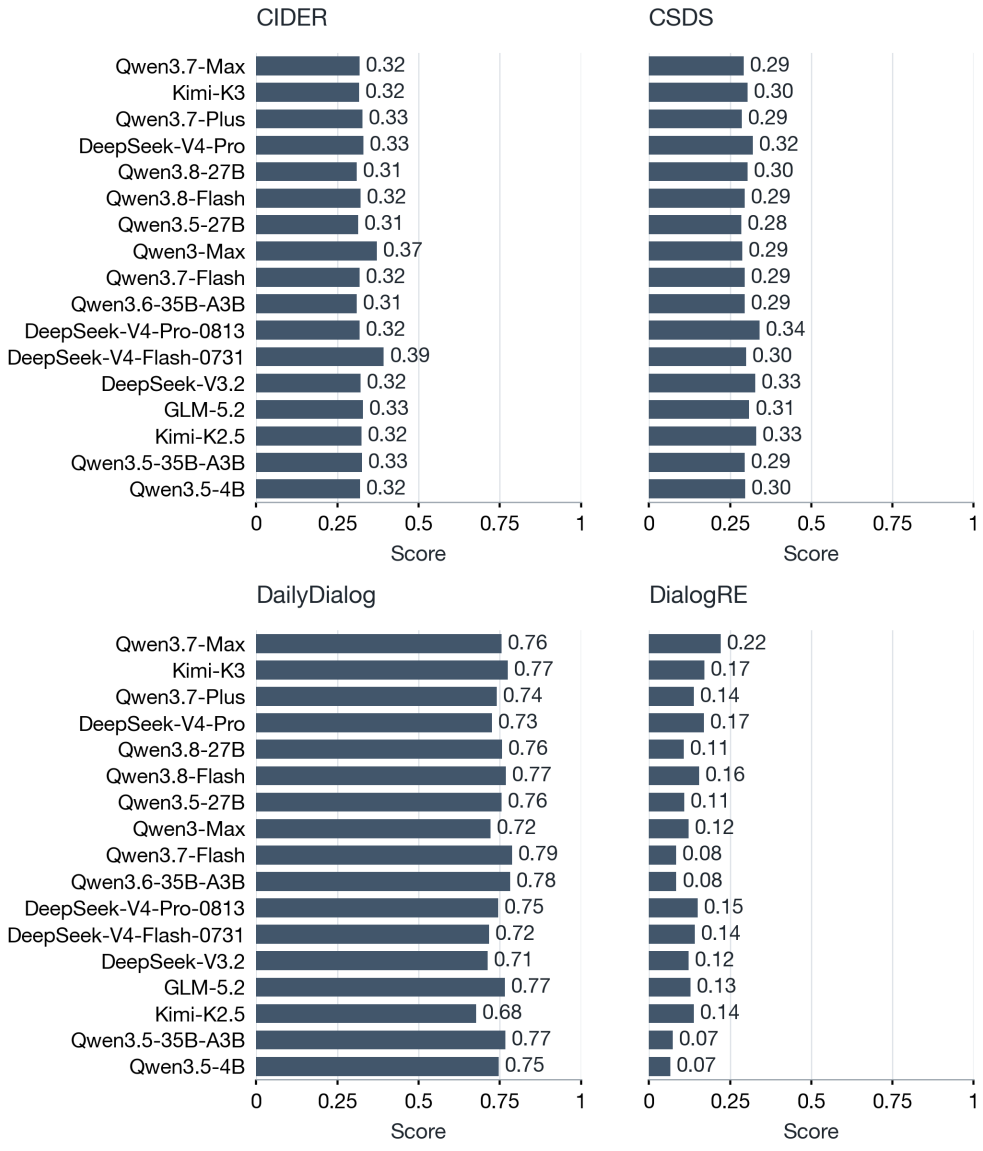

Figure 7: **External dialogue benchmarks: CIDER, CSDS, DailyDialog, and DialogRE.** First group of four benchmarks; all panels use the same 17-model order and 0–1 score axis as Figures 8 and 9.

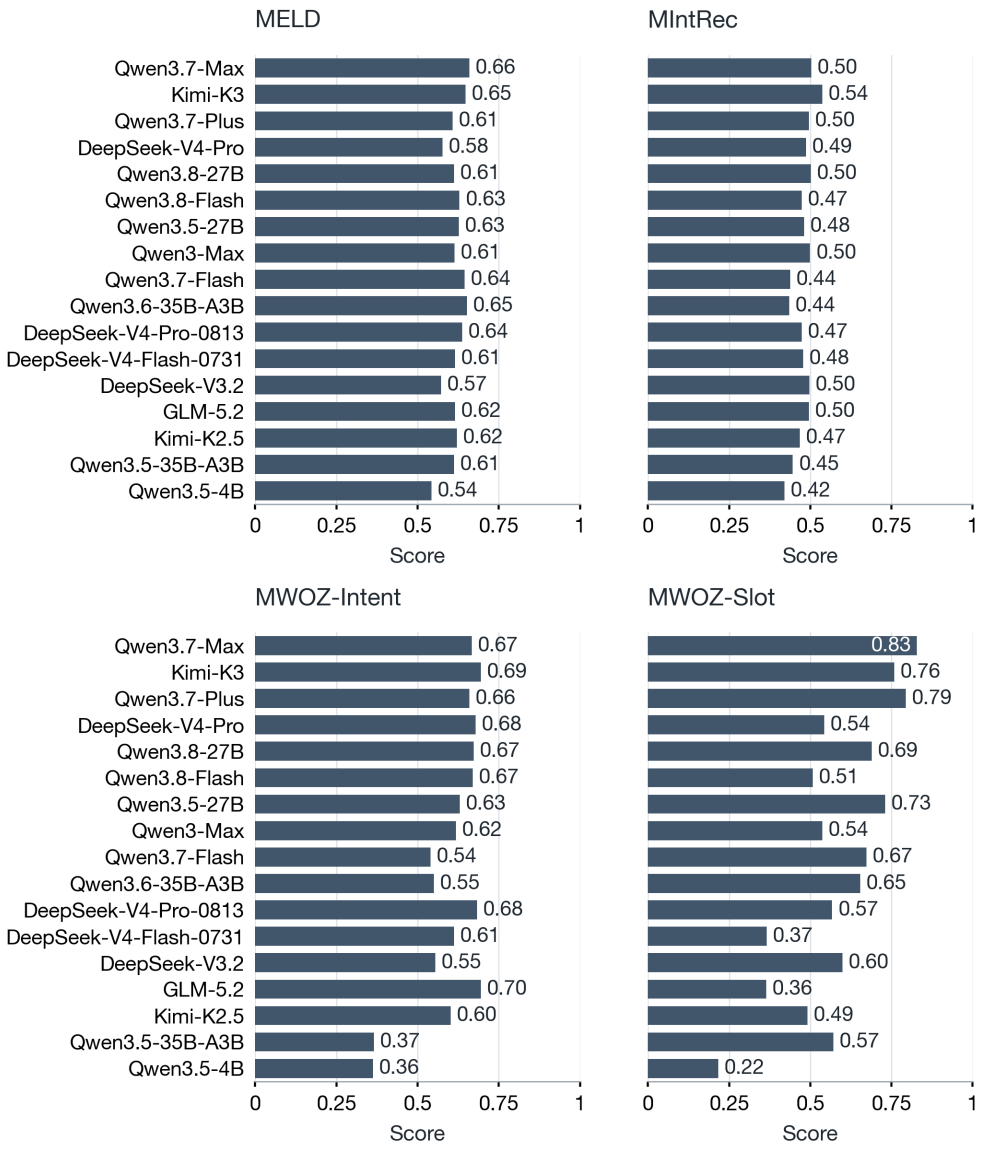

Figure 8: **External dialogue benchmarks: MELD, MIntRec, MWOZ-Intent, and MWOZ-Slot.** Second group of four benchmarks; model order and score axes match Figures 7 and 9.

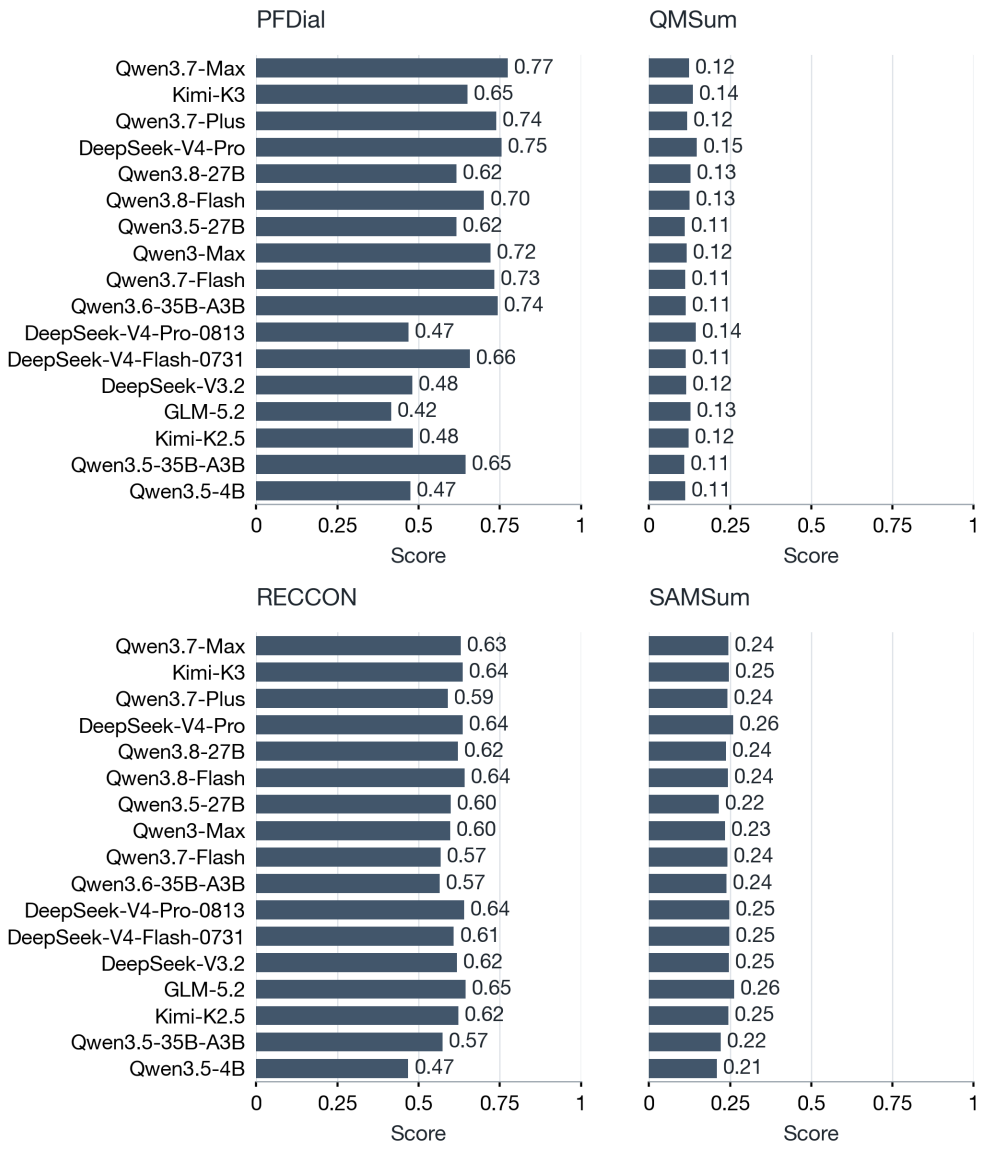

Figure 9: **External dialogue benchmarks: PFDial, QMSum, RECCON, and SAMSum.** Final group of four benchmarks; model order and score axes match Figures 7 and 8.

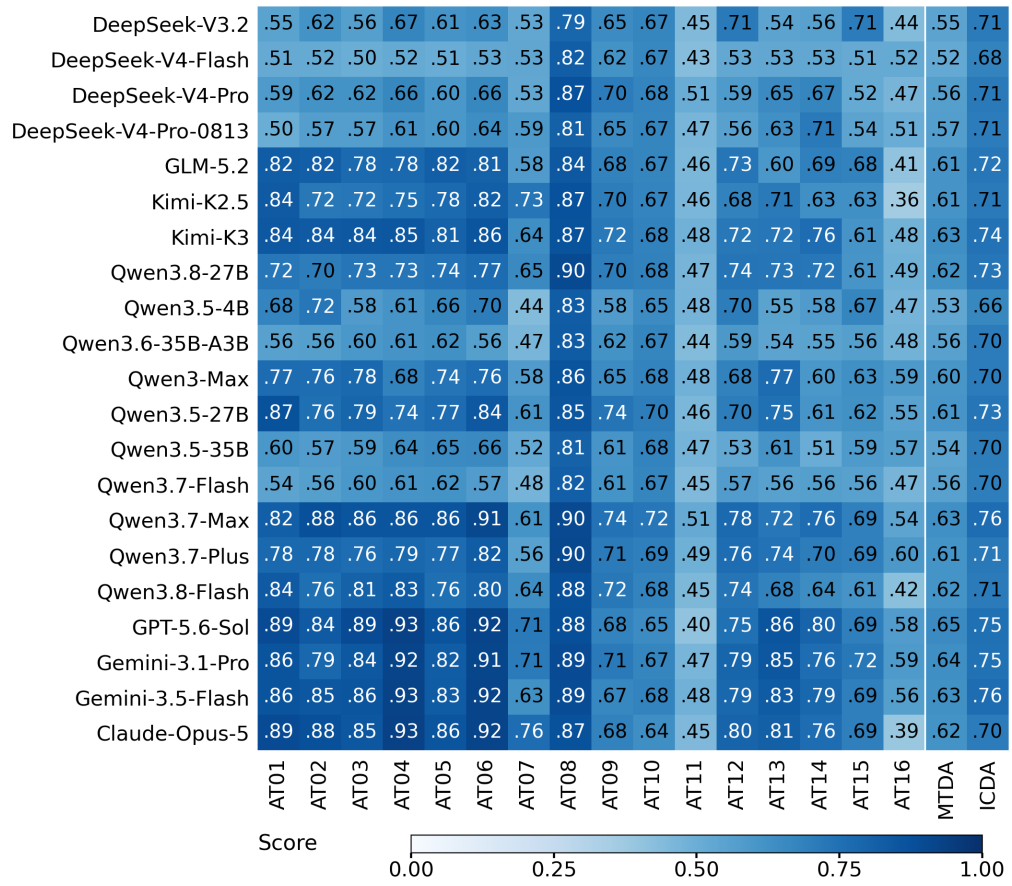

Figure 10: **Per-unit score matrix (21 models × 18 evaluation units).** Cells show raw scores for AT-01–16, MTDA, and ICDA on one sequential 0–1 color scale, without column normalization. Model and task order are shared across the matrix.

Table 8: AT-level discrimination: mean, best/worst model, and spread for representative ATs (21 evaluated models). Statistics use frozen per-task `avg_score` values with equal model weights, then half-up rounding to three decimals for display. Best lists one of the tied leaders where applicable.

AT	Mean	Best (Score)	Worst (Score)	Spread
AT-06	0.762	GPT-5.6-Sol (0.919)	DS-V4-Flash (0.531)	0.388
AT-01	0.731	GPT-5.6-Sol (0.894)	DS-V4-0813 (0.500)	0.394
AT-16	0.499	Qwen3.7-Plus (0.596)	Kimi-K2.5 (0.359)	0.237
AT-11	0.465	Qwen3.7-Max (0.514)	GPT-5.6-Sol (0.404)	0.111

Human scoring protocol and scope. The archived study contains 25 responses per model–unit cell from Qwen3.7-Max and Qwen3.6-Flash, giving 450 responses per model and 900 in total. System-correctness stratification yields 540 positive and 360 negative cases; this sampling is not an estimate of deployment prevalence. All three human annotators rated all 900 responses. Figure 12 groups responses by evaluation unit, not scoring method. The actual scoring-path counts are 600 deterministic, 226 hybrid, and 74 judge-only or calibrated-fallback cases.

F Full Limitations Analysis

1. **Statistical power and multiplicity.** On the 21-model AtomDialBench ranking, only 4/20 adjacent-rank pairs separate under the reported 10,000-resample paired bootstrap analysis, and

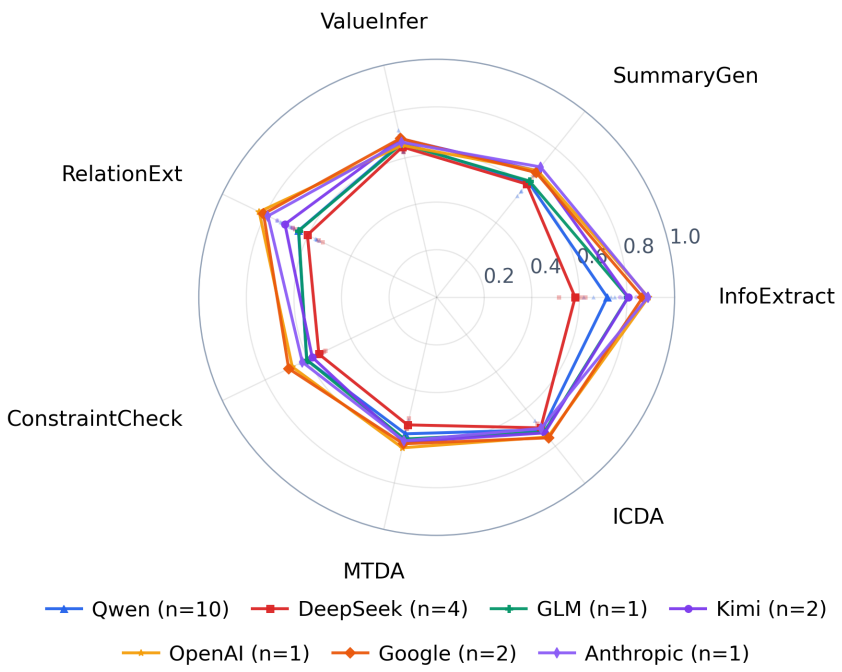

Figure 11: **Family-level reporting profiles (21 models, 7 families).** Lines show family means and faint markers show all 147 member-axis scores across five reporting groups and two composite units, on a common 0–1 scale; legend n counts family members. AT-11 and AT-16 single-task comparisons require per-AT results (Figure 10), not these group means.

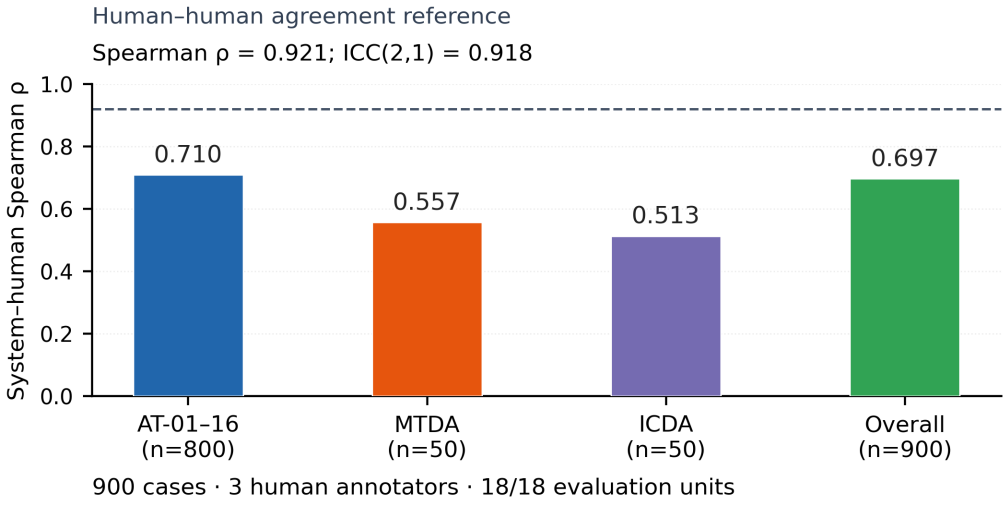

Figure 12: **System–human scoring agreement with three independent human annotators** (researchers with computer-science backgrounds). Spearman ρ: AT-01–16 0.710 ($n = 800$), MTDA 0.557 ($n = 50$), ICDA 0.513 ($n = 50$), and overall 0.697 ($n = 900$). These are evaluation-unit groups, not scoring-method categories. The dashed line is the human–human agreement reference (mean pairwise $\rho = 0.921$), not a performance ceiling; inter-annotator ICC(2,1) is 0.918.

Table 9: Independent 11-model fusion ablation: Kendall's τ against this ablation's Adaptive baseline, not the main 21-model ranking. Most configurations, including Adaptive, show high rank agreement ($\tau \geq 0.891$); hard-only scoring is the clear exception. Top-1/Top-3 stability is relative to Adaptive. Per-AT instance counts are not recoverable from the saved fusion summary; the separate four-AT, 50-case multi-judge scope does not apply.

Strategy	τ	**Top-1**	**Top-3?**
A. Adaptive	1.000	Qwen3.7-Max	—
B. Continuous	0.745	Qwen3.7-Max	Yes
C. Multiplicative	0.927	Qwen3.7-Max	Yes
D. 40/60	0.964	Qwen3.7-Max	Yes
E. Judge-only	0.891	Qwen3.7-Max	Yes
F. Hard-only	0.236	Qwen3.7-Plus	No
G. Noisy-OR	0.891	Qwen3.7-Max	Yes
H. Harmonic	0.927	Qwen3.7-Max	Yes

Table 10: Measurement-evidence summary. The judge ICC is a mean over 105 valid judge-pair–AT combinations from four ATs, 11 models, and the first 50 cases per AT.

Metric	**Value**	**Interpretation**
Bootstrap CI width (ADB)	0.027	Tight bounds
Adjacent pairs sig. (bootstrap)	4/20[†]	1/20 after Bonferroni
8-judge ICC(2,1), mean	0.696	Restricted ablation sample
Inter-annotator ICC(2,1)	0.918	Excellent
System–Human ρ (all 900 cases)	0.697	Strong

[†]Under 10,000-resample paired bootstrap tests (per-benchmark case indices shared across models), the uncorrected boundaries are #5–#6, #13–#14, #18–#19, and #20–#21. Only #13–#14 (Qwen3-Max vs. DeepSeek-V4-Pro) survives Bonferroni correction over the 20 adjacent pairs (§4.2); these tests do not establish statistical tiers.

only the #13–#14 pair (Qwen3-Max vs. DeepSeek-V4-Pro) survives Bonferroni correction. Cross-suite correlations use only $N{=}17$ models. AT-16–CSDS is exploratory: its Spearman association does not survive Bonferroni or Benjamini–Hochberg correction over 288 cross-suite tests. Pointwise bootstrap intervals do not resolve multiple-comparison concerns, and the 561 all-pairs scan is a separate family.

2. **Metric and endpoint differences.** AT-16 set F1 and CSDS ROUGE-L reflect different tasks and metric behavior. Exp3 measures AT-07 exemption judgment, AT-10 summarization, and AT-16 selection; it does not directly test the AT-16–CSDS relationship or exclude metric effects. AT-07's current text-scoring path also warrants a dedicated contract–metric audit.
3. **AT-09 scorer disagreement.** Hard-match and judge scores emphasize different aspects of participant-attribute inference. The AT-09 ablation has mean pairwise ICC 0.769; applying Spearman–Brown with eight judges predicts 0.964. This is neither measured reliability of the production scorer nor evidence of orthogonal constructs.
4. **Training budgets and scale.** Single-AT, joint, and all-AT runs differ in example counts and batch sizes despite sharing three epochs. Only Exp1 and its general-data control match the approximately 114K-example budget. No matched 27B control is available; 27B reference differences do not establish cross-scale replication.
5. **AT-11 limited discrimination.** Across 21 models, mean score is 0.465 with range approximately 0.404–0.514 and spread 0.111. The low, narrow range is not a ceiling and does not show model equivalence. Broader instance and metric checks remain useful.
6. **AT-16 scoring choice.** Set F_1 measures candidate overlap. It need not capture the correctness of supplied scores, tie-breaking, or the complete ranking.
7. **Judge dependence.** The main judge is excluded from the 21-model comparison, and the eight-judge ablation (AT-07–10, 11 models, first 50 cases per AT) includes self-scoring. After excluding a few non-comparable pairs, 105 valid pair–AT ICC(2,1) estimates average 0.696 (moderate).

Rankings shift somewhat across judges, so agreement does not by itself guarantee identical orderings.

8. **Coverage and synthesis.** Four models have ADB-only coverage and do not enter external correlations. Qwen3.7-Max both synthesizes instances and is evaluated; its rank does not rule out synthesis-related bias. Balanced Chinese/English sampling and source filtering do not guarantee population representativeness.

Scoring robustness. Hybrid fallback can reduce penalties when deterministic matching fails, but a zero hard score may also reflect an incorrect answer. Its calibration is a scoring choice, not a guarantee of zero false negatives. Language, format, and task-specific disagreements should be inspected alongside aggregate agreement.

Domain generalization. Business-domain OOD evaluation (finance/e-commerce/consulting) yields mean $\Delta = +0.045$ with 13/16 ATs improving. Regressions on AT-16, AT-11, and AT-14 warrant separate examination of format sensitivity, metric behavior, and task performance; they do not isolate a construct-level explanation.

Evaluation vs. training relationships. The AT$\times$AT matrix describes cross-model score association, whereas the transfer matrix describes changes under particular training interventions. Their differences can reflect task, metric, and training-method effects (Podsakoff et al., 2003); neither matrix alone establishes construct independence.

G Additional Experimental Details

This appendix provides the detailed benchmark tables deferred from §4 and the training-experiment data underlying §4.4. Unless noted otherwise, the 10-model columns are a representative subset of the 17 full-coverage models (Qwen-Max = Qwen3.7-Max, Qwen-Plus = Qwen3.7-Plus, DS-V4-Pro = DeepSeek-V4-Pro, DS-V4-Flash = DeepSeek-V4-Flash-0731, GLM = GLM-5.2, Kimi = Kimi-K2.5, DS-V3.2 = DeepSeek-V3.2, Qwen-27B = Qwen3.5-27B, Qwen-35B = Qwen3.5-35B-A3B, Qwen3-Max = Qwen3-Max). The Qwen3.6-Flash judge model does not appear in these tables because it is excluded from the main comparison.

G.1 Model List

Table 11 enumerates all 21 evaluated models with their developer, benchmark coverage, and evaluation window.

G.2 Per-AT Scores

Table 12 reports 18 evaluation units for a 10-model subset. Its Mean column, including MTDA 0.585 and ICDA 0.715, describes only these ten models, not the complete 21-model comparison.

The AT-07 marker (†) identifies exemption judgment scored through the cached text-based hybrid path. `aggregate_results.py` reads final `score` values from result records. Existing AT-07 caches contain `hybrid_judge_summary` and calibrated judge-only methods; they do not support describing AT-07 as pure ROUGE-L or the 1.18 fallback as universally retired. For example, one Qwen3.7-Max record has final score 0.90625, hard score 0.8125, and judge score 1.0. An AT-07-only check used `compute_stats` on cached final scores for the 21 paper models and, separately, `exp1_all_at_4B`. Case counts (160 each) and means (to four decimal places) matched the frozen `eval_results.json`, including Qwen3.7-Max's mean of 0.6144. This check did not cover other tasks, rerun inference or judging, or replay the full pipeline (Reproducibility Details).

AT-09 (‡) reports the production hybrid score, not an eight-judge aggregate. The separate AT-09 value 0.964 is predicted reliability under Spearman–Brown. AT-16 (§) uses set F_1 for candidate selection, which does not assess all ranking fields.

G.3 External Dialogue Benchmarks

Table 13: External dialogue benchmarks: 12 established dialogue-analysis benchmarks (10-model representative subset).

Bench	Qwen-Max	Qwen-Plus	DS-V4-Pro	DS-V4-Flash	GLM	Kimi	DS-V3.2	Qwen-27B	Qwen-35B	Qwen3-Max	**Mean**
CIDER	0.318	0.327	0.330	0.393	0.329	0.325	0.322	0.314	0.325	0.372	0.336
CSDS	0.292	0.287	0.320	0.299	0.309	0.330	0.327	0.285	0.295	0.287	0.303
DailyDialog	0.756	0.740	0.726	0.718	0.765	0.677	0.712	0.756	0.768	0.721	0.734
DialogRE	0.222	0.138	0.169	0.140	0.127	0.140	0.122	0.110	0.073	0.122	0.136
MELD	0.660	0.608	0.577	0.615	0.616	0.621	0.573	0.627	0.612	0.614	0.612
MIntRec	0.504	0.496	0.487	0.478	0.496	0.468	0.497	0.481	0.446	0.499	0.485
MWOZ-Int	0.667	0.660	0.678	0.612	0.695	0.602	0.554	0.630	0.365	0.618	0.608
MWOZ-Slot	0.828	0.794	0.542	0.366	0.364	0.491	0.599	0.730	0.571	0.537	0.582
PFDial	0.775	0.739	0.755	0.658	0.415	0.483	0.481	0.617	0.645	0.722	0.629
QMSum	0.123	0.118	0.148	0.113	0.128	0.122	0.115	0.111	0.109	0.116	0.120
RECCON	0.629	0.590	0.636	0.609	0.645	0.622	0.618	0.599	0.573	0.597	0.612
SAMSum	0.244	0.241	0.259	0.248	0.263	0.245	0.246	0.215	0.221	0.235	0.242

G.4 Instruction-Following Benchmarks (External Calibration)

Table 14: IF benchmarks for external calibration: 4 instruction-following benchmarks (10-model representative subset). CFBench uses 999 evaluation instances.

Bench	Qwen-Max	Qwen-Plus	DS-V4-Pro	DS-V4-Flash	GLM	Kimi	DS-V3.2	Qwen-27B	Qwen-35B	Qwen3-Max	**Mean**
IFBench	0.525	0.537	0.500	0.443	0.563	0.433	0.478	0.477	0.443	0.475	0.487
CFBench	0.541	0.506	0.466	0.470	0.494	0.466	0.471	0.491	0.462	0.508	0.487
InverseIF	0.284	0.259	0.234	0.221	0.238	0.212	0.193	0.218	0.196	0.232	0.229
GuideBench	0.828	0.669	0.681	0.839	0.671	0.813	0.744	0.665	0.809	0.694	0.741

G.5 Training Experiment Details

This subsection provides the detailed data underlying §4.4.

Hyperparameters. The 21 4B runs include the UltraData-only control, not 21 AT-mixture interventions. Exp1 all-AT training uses a 70% AT / 30% UltraData mixture; Exp1 and control each contain approximately 114K examples with global batch size 128. Single-AT runs contain approximately 7K examples and the AT-07+10+16 joint run approximately 21K, both with batch size 64.

Table 11: Evaluated models: developer, coverage, and evaluation window.

Model	**Developer**	**Coverage**	**Eval Period**
Qwen3.7-Max	Qwen	Full	2026-07
Qwen3.7-Plus	Qwen	Full	2026-07
Qwen3.5-27B	Qwen	Full	2026-07
Qwen3.8-Flash	Qwen	Full	2026-07
Qwen3.8-27B	Qwen	Full	2026-07
Qwen3-Max	Qwen	Full	2026-07
Qwen3.5-4B	Qwen	Full	2026-07
Qwen3.5-35B-A3B	Qwen	Full	2026-07
Qwen3.6-35B-A3B	Qwen	Full	2026-07
Qwen3.7-Flash	Qwen	Full	2026-07
DeepSeek-V4-Pro	DeepSeek	Full	2026-07
DeepSeek-V3.2	DeepSeek	Full	2026-07
DeepSeek-V4-Pro-0813	DeepSeek	Full	2026-08
DeepSeek-V4-Flash-0731	DeepSeek	Full	2026-07
GLM-5.2	GLM	Full	2026-07
Kimi-K3	Kimi	Full	2026-08
Kimi-K2.5	Kimi	Full	2026-07
GPT-5.6-Sol⋆	OpenAI	ADB only	2026-08
Gemini-3.1-Pro⋆	Google	ADB only	2026-08
Gemini-3.5-Flash⋆	Google	ADB only	2026-08
Claude-Opus-5⋆	Anthropic	ADB only	2026-08

Table 12: Per-AT scores (18 evaluation units × 10-model representative subset).

AT	Qwen-Max	Qwen-Plus	DS-V4-Pro	DS-V4-Flash	GLM	Kimi	DS-V3.2	Qwen-27B	Qwen-35B	Qwen3-Max	**Mean**
AT-01	0.819	0.781	0.588	0.512	0.825	0.838	0.550	0.869	0.600	0.769	0.715
AT-02	0.875	0.781	0.619	0.519	0.819	0.719	0.619	0.756	0.575	0.762	0.704
AT-03	0.856	0.762	0.619	0.500	0.775	0.725	0.562	0.794	0.594	0.781	0.697
AT-04	0.856	0.788	0.662	0.519	0.781	0.750	0.669	0.744	0.638	0.681	0.709
AT-05	0.856	0.769	0.600	0.506	0.819	0.775	0.612	0.769	0.650	0.738	0.709
AT-06	0.906	0.825	0.662	0.531	0.806	0.819	0.631	0.844	0.656	0.762	0.744
AT-07†	0.614	0.558	0.529	0.532	0.579	0.727	0.528	0.613	0.520	0.581	0.578
AT-08	0.899	0.903	0.871	0.824	0.845	0.871	0.790	0.854	0.806	0.862	0.852
AT-09‡	0.745	0.709	0.704	0.619	0.680	0.702	0.647	0.737	0.607	0.647	0.680
AT-10	0.724	0.695	0.675	0.666	0.671	0.673	0.670	0.700	0.678	0.677	0.683
AT-11	0.514	0.491	0.507	0.433	0.459	0.462	0.447	0.461	0.475	0.483	0.473
AT-12	0.775	0.762	0.588	0.525	0.731	0.681	0.712	0.700	0.525	0.681	0.668
AT-13	0.725	0.744	0.650	0.531	0.600	0.712	0.544	0.750	0.612	0.769	0.664
AT-14	0.762	0.700	0.669	0.531	0.688	0.631	0.556	0.606	0.506	0.600	0.625
AT-15	0.688	0.688	0.519	0.512	0.681	0.631	0.706	0.625	0.588	0.631	0.627
AT-16§	0.540	0.596	0.470	0.518	0.406	0.359	0.439	0.554	0.566	0.592	0.504
MTDA	0.630	0.615	0.564	0.522	0.611	0.613	0.545	0.607	0.544	0.602	0.585
ICDA	0.762	0.715	0.714	0.679	0.724	0.713	0.708	0.727	0.702	0.704	0.715

Table 15: Training hyperparameters for the 4B fine-tuning experiments (21 runs; 27B differences in the text).

Hyperparameter	**Value**
Base model	Qwen3.5-4B
Adapter	LoRA, rank $= 256$, $\alpha = 256$
Epochs	3
Learning rate	2×10^{-4}, min 2×10^{-5}
Warmup fraction	0.01
Global batch size	128 (114K mixes) / 64 (7K–53K mixes)
Max sequence length	16,384
Parallelism	Tensor parallel 2, sequence parallel, packing
Exact-duplicate check	MD5 fingerprints: zero exact overlap

The ICDA run uses approximately 53K examples. All use three epochs, so equal epoch counts do not imply matched example or token exposure (Table 15). UltraData is the general-purpose instruction-tuning mixture.

The exploratory 27B runs use LoRA rank$/\alpha = 128$ and smaller global batches (64 for 114K mixes, 32 for 7K–53K mixes), with the same nominal epochs, learning rate, sequence length, and parallelism. They lack a matched 27B control and are interpreted separately.

G.5.1 EXP1: PER-AT EFFECT SIZES

Table 16 compares `exp1_all_at_4B` with `control_ultra_only_4B`. Both use the approximately 114K-example budget. Gains are positive on 17/18 units, with mean $+0.115$ over 16 ATs and $+0.107$ over all 18 units. Eleven improvements pass paired two-sided t-tests with Bonferroni threshold $\alpha = 0.05/18 \approx 0.0028$. The zero MD5 overlap check excludes exact duplicates, not semantic overlap or prior pretraining exposure.

G.5.2 EXP2: TRANSFER MATRIX SUMMARY

The diagonal mean Δ is $+0.075$ (13/16 entries exceed $+0.02$). In the 16×16 atomic-task block of the matrix (i.e., the 240 off-diagonal AT-to-AT entries, excluding the two composite-unit columns), 80/240 show positive transfer above $+0.02$ (33.3%), 124/240 negative transfer below -0.02 (51.7%), and 36/240 near-zero change (± 0.02). These changes compare approximately 7K-example single-AT runs with the approximately 114K-example general-data control. They describe transfer under different training budgets, not isolated composition effects or independent constructs. Figure 13 shows the full 16×18 matrix.

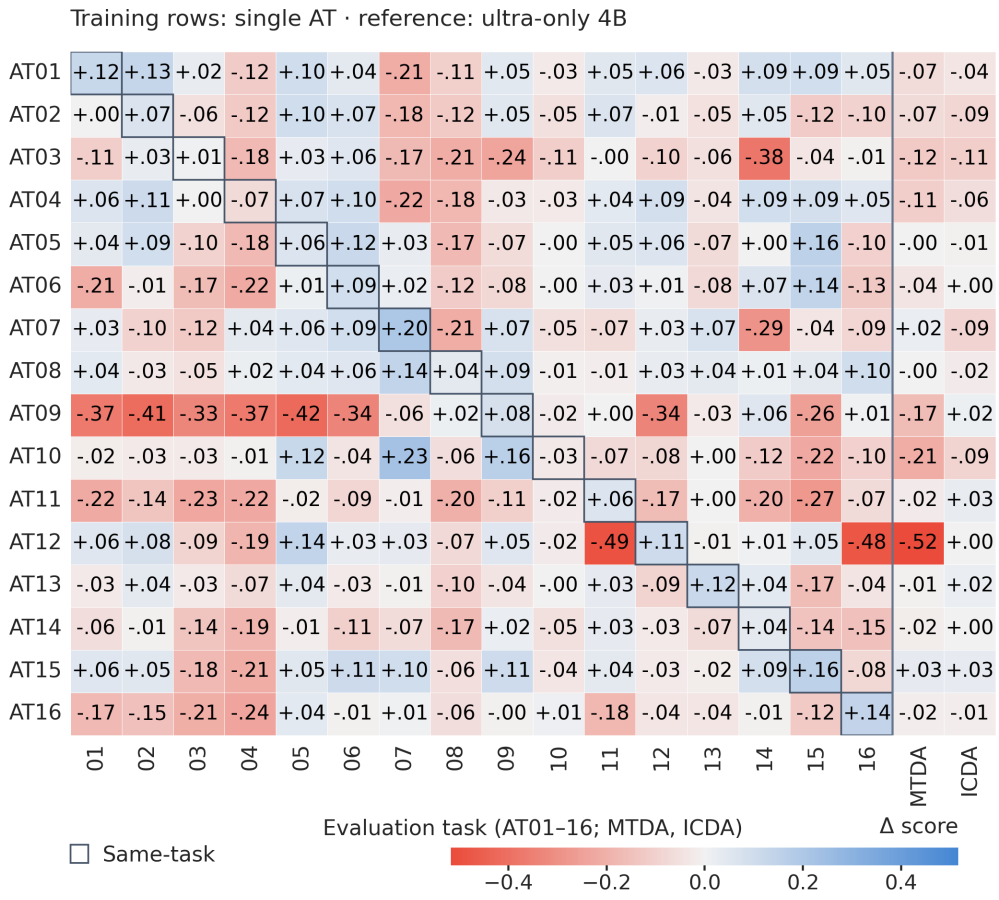

Figure 13: **Training transfer matrix (16 single-AT models × 18 evaluation units).** Rows identify training tasks, columns evaluation units; cells show trained-model minus ultra-only 4B control scores on a symmetric, zero-centered color scale. Outlined diagonal cells mark same-task evaluation, not necessarily improvement (13/16 exceed $+0.02$). Off-diagonal transfer rates use $|\Delta| > 0.02$ in the 16×16 atomic-task block, excluding MTDA/ICDA. These responses do not establish construct independence. Full matrix: `transfer_matrix.npy`.

Table 16: Exp1 per-AT effect sizes (all-AT training vs. control). Sig. column marks Bonferroni-significant improvements.

AT	Control	Exp1	Δ	Cohen's d	p-value	Sig.
AT-01	0.725	0.856	+0.131	0.30	2.0e-04	yes
AT-02	0.713	0.838	+0.125	0.30	2.0e-04	yes
AT-03	0.706	0.750	+0.044	0.09	2.6e-01	no
AT-04	0.806	0.794	−0.013	−0.03	7.2e-01	no
AT-05	0.706	0.775	+0.069	0.16	4.8e-02	no
AT-06	0.781	0.875	+0.094	0.26	1.6e-03	yes
AT-07	0.422	0.626	+0.205	0.69	6.0e-15	yes
AT-08	0.695	0.774	+0.080	0.26	1.4e-03	yes
AT-09	0.503	0.649	+0.145	0.55	1.2e-10	yes
AT-10	0.648	0.699	+0.051	0.44	1.1e-07	yes
AT-11	0.495	0.557	+0.062	0.22	6.3e-03	no
AT-12	0.775	0.856	+0.081	0.19	1.5e-02	no
AT-13	0.581	0.763	+0.181	0.40	1.7e-06	yes
AT-14	0.544	0.775	+0.231	0.44	1.3e-07	yes
AT-15	0.763	0.944	+0.181	0.45	5.7e-08	yes
AT-16	0.483	0.650	+0.167	0.39	1.8e-06	yes
MTDA	0.541	0.584	+0.044	—	—	—
ICDA	0.561	0.605	+0.044	—	—	—

Table 17: Exp3 score differences relative to `control_ultra_only_4B` for all four configurations. The 4B configurations have different exposure budgets; the 27B rows are exploratory cross-scale reference differences, not matched 27B gains.

Model	AT-07 Δ	AT-10 Δ	AT-16 Δ
AT-16 only (4B)	+0.007	+0.008	+0.138
AT-07+10+16 joint (4B)	+0.254	+0.028	+0.094
AT-16 only (27B)	+0.280	−0.015	+0.027
AT-07+10+16 joint (27B)	+0.293	+0.040	+0.118

G.5.3 Exp3: Responses to Training Configurations

At 4B, AT-16-only and joint AT-07+10+16 training produce different profiles on exemption judgment, summarization, and candidate selection. They use approximately 7K and 21K examples, respectively, versus the approximately 114K-example control; equal epochs do not isolate composition. These endpoints do not directly test AT-16–CSDS. Table 17 and Figure 14 retain exploratory 27B results relative to the same 4B control. No matched 27B control is available and the 27B resources are incomplete.

G.5.4 OOD Generalization (Five Categories)

Table 18 and Figure 15 summarize the transfer profile of `exp1_all_at_4B` against `control_ultra_only_4B`, not Exp2. The model improves on average across external dialogue and business-domain tasks, while IFBench shows no detected gain. Among the 12 external dialogue benchmarks, 10 improve ($\Delta > 0$); 8 of these improvements are individually significant (paired t-tests, α=0.05, uncorrected), as are the two regressions (PFDial -0.336; MultiWOZ slot extraction -0.168). This observation does not establish a general boundary for instruction-following transfer. Differences use `control_ultra_only_4B` (pure UltraData, no AT data) as the reference.

G.5.5 Business-Domain OOD (Per-AT)

Business-domain OOD evaluates the Exp1-trained model on 16 ATs constructed from business-customer dialogues in finance, e-commerce, and consulting, using the saved hard-match results. The domain labels describe this transfer setting; they do not establish absence of related content from the general-data mixture or pretraining. These hard-match results are not identical in scoring scope to every hybrid path in AtomDialBench. As shown in Table 19, 13/16 ATs improve under cross-

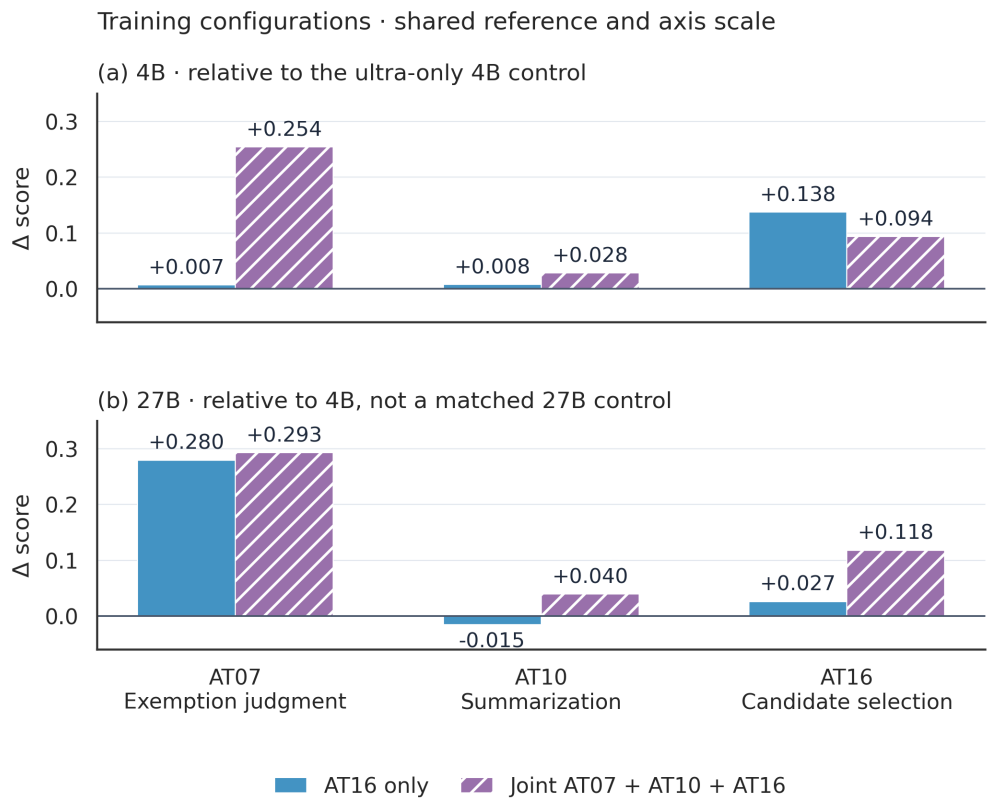

Figure 14: **Task responses to training configurations (Exp3).** Two panels share the same score-difference scale: (a) 4B configurations with different exposure budgets; (b) exploratory 27B comparisons. All 12 bars use the same ultra-only 4B control reference (Table 17). AT-07 is exemption judgment, AT-10 summarization, and AT-16 candidate selection. Without a matched 27B control, panel (b) does not estimate matched-scale gains or establish cross-scale replication.

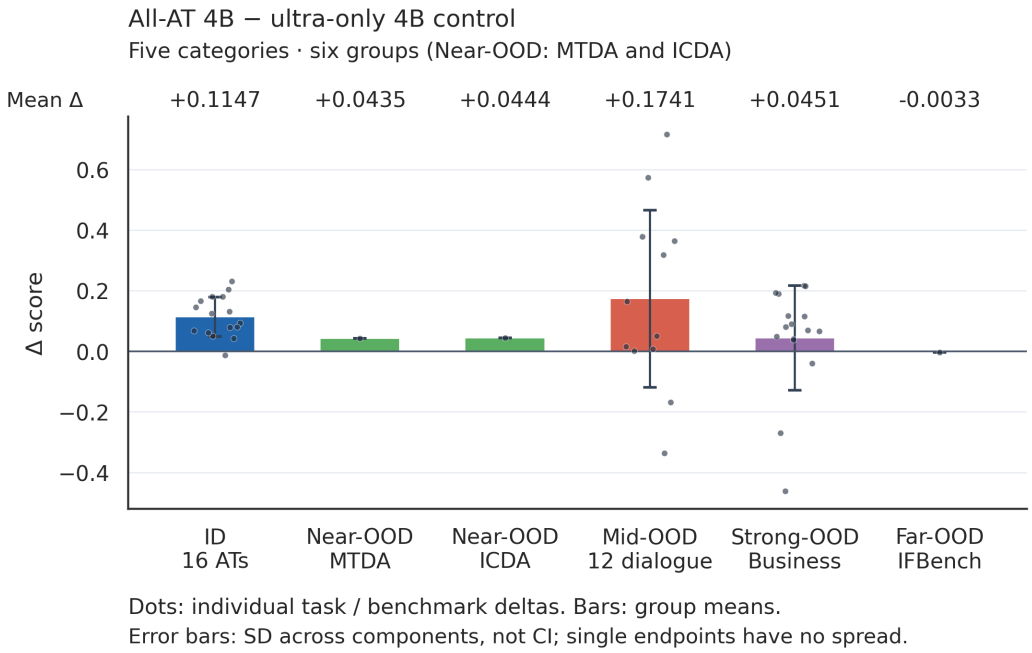

Figure 15: **OOD profile of all-AT 4B training.** The horizontal axis lists five discrete categories in six groups: ID (16 ATs), Near-OOD (MTDA and ICDA separately), Mid-OOD (12 dialogue benchmarks), Strong-OOD (16 business-domain ATs), and Far-OOD (IFBench); spacing does not measure distributional distance. The vertical axis is score difference from the ultra-only 4B control. Dots show all 47 component differences; bars show group means. Error bars are SD across tasks or benchmarks, not confidence intervals; single endpoints have no spread.

Table 18: Five-category OOD generalization profile across discrete evaluation settings. The Positive column counts $\Delta > 0$ with its row-specific denominator: 16 ATs (ID), 12 external benchmarks (Mid-OOD), 16 ATs (Strong-OOD), and 1 endpoint (Far-OOD). The ID Δ is the 16-AT mean; the separate 18-evaluation-unit mean, including MTDA and ICDA, is $+0.107$. Figure 15 additionally shows the Near-OOD ICDA endpoint.

OOD Category	Evaluation	Δ	Positive	N
ID	AtomDialBench SynthAT (16 ATs)	$+0.115$	15/16	2560
Near-OOD	MTDA (composite)	$+0.044$	—	499
Mid-OOD	Ext-Dialogue (12 benchmarks)	$+0.174$	10/12	4471
Strong-OOD	Business-domain OOD	$+0.045$	13/16	1327
Far-OOD	IFBench (instruction following)	-0.003	0/1	300

Table 19: Business-domain OOD per-AT results (AT-trained vs. control).

AT	AT-trained	Control	Δ	Direction
AT-01	0.950	0.900	$+0.050$	up
AT-02	0.983	0.867	$+0.117$	up
AT-03	1.000	0.930	$+0.070$	up
AT-04	0.867	0.800	$+0.067$	up
AT-05	0.973	0.892	$+0.081$	up
AT-06	0.850	0.733	$+0.117$	up
AT-07	0.343	0.301	$+0.042$	up
AT-08	0.761	0.570	$+0.191$	up
AT-09	0.595	0.401	$+0.195$	up
AT-10	0.544	0.505	$+0.038$	up
AT-11	0.468	0.737	-0.269	down
AT-12	0.917	0.700	$+0.217$	up
AT-13	0.750	0.659	$+0.091$	up
AT-14	0.782	0.821	-0.039	down
AT-15	0.933	0.717	$+0.217$	up
AT-16	0.127	0.588	-0.462	down
Mean	**0.740**	**0.695**	$+0.045$	**13/16**

domain transfer. Three regressions require closer inspection: AT-16 ($\Delta = -0.462$) coincides with a reported field-name mismatch (`option` in training vs. `item` in business-domain OOD data); AT-11 ($\Delta = -0.269$) and AT-14 ($\Delta = -0.039$) may involve domain-specific scoring behavior. These observations do not isolate format effects from changes in task performance.

G.5.6 Exploratory 27B Results and Negative Findings

The 27B entries in Table 17 are retained only as exploratory results. Resources are incomplete and there is no matched 27B control; all four rows use `control_ultra_only_4B` as the reference. Consequently, the 27B differences combine model-scale and training differences and cannot identify a capacity mechanism or matched-scale gain. They also do not establish directional replication: from AT-16-only to joint training, the ratio $\Delta\text{AT16}/\Delta\text{AT07}$ decreases from approximately 19.71 to 0.370 at 4B, but increases from 0.096 to 0.403 at 27B using the reported rounded differences.

For completeness, Exp1 shows no significant improvement on AT-03 intent matching ($p = 0.264$, $\Delta = +0.044$), while AT-04 information-mention detection has $p = 0.716$, $\Delta = -0.013$, and control score 0.806. External regressions include PFDial ($\Delta = -0.336$) and MultiWOZ slot extraction ($\Delta = -0.168$). Business-domain regressions include AT-16 ($\Delta = -0.462$) and AT-11 ($\Delta = -0.269$). Format and vocabulary differences are plausible contributors, but these results do not by themselves separate scoring effects from capability changes.

H SCALING ANALYSIS

This appendix provides the full scaling evidence deferred from the Discussion. The frozen 21-model list is evaluated on all 16 atomic tasks (AT-01–16) and two composite units (MTDA and ICDA). Every cell uses the final unit mean score (`avg_score`), with no accuracy substitution; ADB is the equal-weight mean of these 18 units, not the equal-weight mean of the five reporting groups. The four-decimal unit values are the inputs to the comparisons below. The six-decimal frozen profile aggregates serve as a cross-source check with absolute tolerance 5.1×10^{-5} and zero relative tolerance, accommodating the different source precision. Differences and increase/decrease/tie counts are computed before display rounding, with tolerance 10^{-9} for ties.

The focal comparisons use Qwen3.5-4B, Qwen3.5-27B, and Qwen3.5-35B-A3B. Qwen3.8-27B and Qwen3.6-35B-A3B are distinct evaluated models, not additional members of that focal trio. As throughout the paper, "dialogue understanding" denotes the ability probed by these dialogue analysis tasks. Both comparisons are descriptive: shared family branding does not establish identical training, and the dense-to-MoE comparison also changes architecture and activation.

H.1 QWEN3.5 FAMILY: RANKING AND AGGREGATE GAINS

Table 20 isolates the three focal models. The 27B dense model improves ADB by $+0.0897$ relative to 4B dense (rank #15 $\rightarrow$ #8), but not on every AT. The 35B-A3B MoE model has 35B total and 3B activated parameters; its ADB difference is -0.1036 relative to 27B dense (rank #8 $\rightarrow$ #18) and -0.0139 relative to 4B dense. The frozen profile ADB values are respectively 0.616297, 0.705968, and 0.602362; their pairwise differences are $+0.089671$, -0.103606, and -0.013935. Small last-digit differences from recomputing the 18-unit means reflect source precision, not a change of metric. Figure 16 places the trio within the full 21-model ranking.

Table 20: **Qwen3.5 configurations: ADB aggregate and rank.** ADB is recomputed from the 18 unit mean scores and shown to four decimals. Ranks cover all 21 models. The second comparison changes architecture as well as parameter counts; neither comparison identifies a causal effect of total or activated capacity.

Model	Size (disclosed)	ADB	Rank	Difference
Qwen3.5-4B	4B dense	0.6163	15	—
Qwen3.5-27B	27B dense	0.7060	8	$+0.0897$ (vs. 4B)
Qwen3.5-35B-A3B	35B MoE (3B active)	0.6024	18	-0.1036 (vs. 27B)

H.2 PAIRWISE CHANGES AND THE AGGREGATE REVERSAL

Table 21 reports all 18 unit scores and two absolute score differences: $\Delta_D = 27\text{B dense} - 4\text{B dense}$ and $\Delta_M = 35\text{B-A3B MoE} - 27\text{B dense}$. Among the 16 ATs, Δ_D has 13 increases, 2 decreases (AT-11 and AT-15), and 1 tie (AT-12); Δ_M has 2 increases and 14 decreases, with no ties. The only MoE-minus-27B increases are AT-11 ($+0.0141$) and AT-16 ($+0.0125$); AT-08 instead declines by -0.0474. The five largest drops are AT-01 speech-act detection (-0.2688), AT-03 intent matching (-0.1999), AT-06 condition satisfaction (-0.1876), AT-02 emotion detection (-0.1812), and AT-12 causal judgment (-0.1750). Four belong to InfoExtract. MTDA and ICDA decline by -0.0629 and -0.0245, respectively, and are excluded from the 16-AT counts.

Exactly 12 ATs rise from 4B to 27B and then fall at 35B-A3B: AT-01–10, AT-13, and AT-14. The remaining patterns are down/up (AT-11), flat/down (AT-12), down/down (AT-15), and up/up (AT-16). Thus 14 pairwise decreases do not mean 14 three-point rise-then-fall trajectories. The aggregate reversal remains, but the task-level evidence distinguishes the two comparisons explicitly.

Figure 18 visualizes both pairwise comparisons; Figure 17 supplies the corresponding absolute scores. Together they distinguish the aggregate reversal from individual AT trajectories.



Figure 16: **All 21 models ranked by ADB.** Each row has a score-positioned point and independent columns for the model name, ADB value, total parameters, activated parameters, and architecture. The three Qwen3.5 focal rows are highlighted. Parameter information is retained for the five disclosed-size models; a dash denotes unavailable metadata, not a numeric position on a parameter axis. The 27B focal model ranks above both 4B and 35B-A3B, but this ordering does not isolate a parameter-scaling mechanism.

H.3 Scaling Benefit by Reporting Group

Tables 22 and 23 aggregate only the actual members of each reporting group, weighting its ATs equally. Membership is not obtained by partitioning AT-01–16 into consecutive numeric ranges, and the two composite units remain outside these groups. The dense difference is positive for all five groups: largest for InfoExtract ($+0.1365$) and smallest for ConstraintCheck ($+0.0120$). SummaryGen gains $+0.1120$, ValueInfer $+0.0558$, and RelationExt $+0.1125$. All five MoE-minus-27B group differences are negative, but MoE-minus-4B is still positive for SummaryGen ($+0.0550$) and ValueInfer ($+0.0013$). These localized gains coexist with the lower MoE ADB; they do not establish uniform deterioration relative to 4B or a universal architecture effect.

H.4 Complete Model-by-unit Matrix

The matrix is split into two continued panels so that all 378 values remain readable at the fixed text width. Both panels use exactly the same ADB-ranked model order and 0–1 color range; grouping

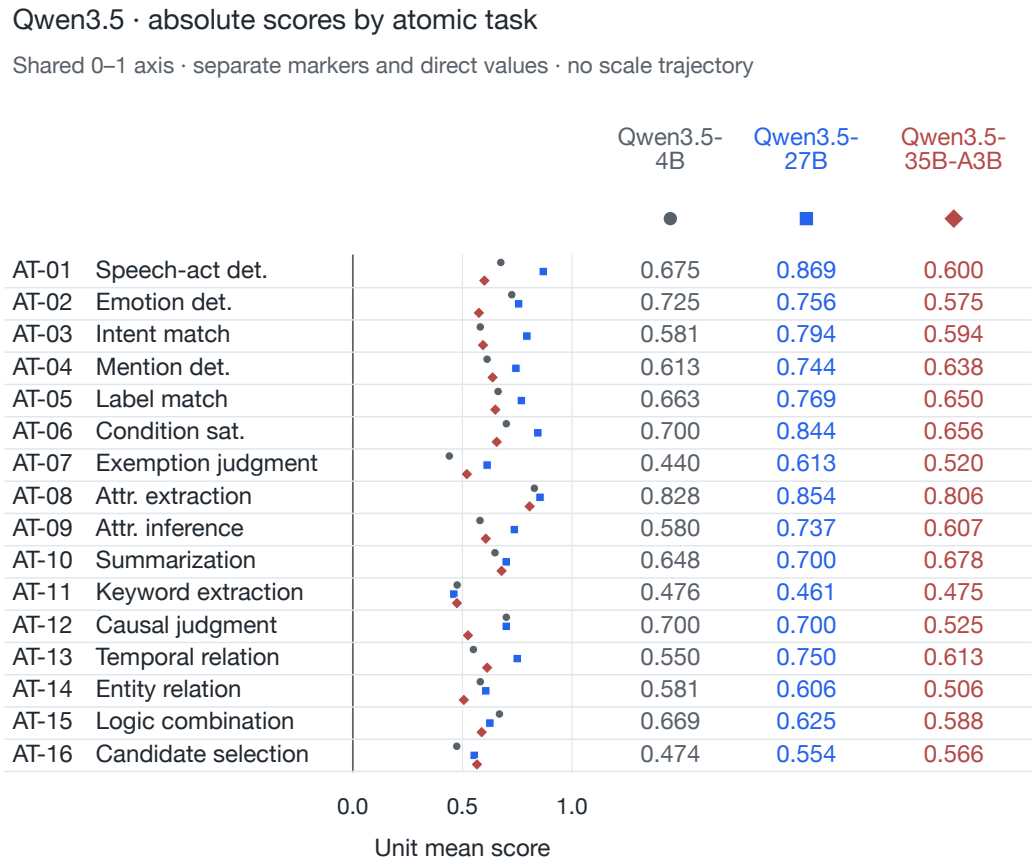

Figure 17: **Absolute per-AT scores for the Qwen3.5 trio.** Sixteen aligned rows share a 0–1 score axis; circle, square, and diamond markers identify 4B dense, 27B dense, and 35B-A3B MoE, respectively. Small vertical offsets keep coincident scores distinguishable. The three right-hand columns directly label all 48 scores to three decimals. Points are not connected into a causal scale trajectory; Table 21 gives both pairwise differences and four-decimal values.

follows task membership rather than an artificial consecutive-number partition. The second panel includes the two composite units, displayed as MTDA and ICDA rather than their storage keys.

Scope. The reversal is observed within one family and one MoE configuration. Architecture, expert routing, activated parameter count, and possible training differences are confounded; 3B active does not imply the effective capacity of a sub-4B dense model. These results do not establish activated parameters as a stronger predictor than total parameters, and no confidence intervals or significance claims are supplied for these contrasts. Unknown parameter metadata remains missing rather than inferred from product names. Application-level selection therefore requires direct evaluation on the relevant ATs, not extrapolation from a parameter count or an MoE label.

Table 21: **Qwen3.5 per-unit scores and both comparisons.** MoE denotes 35B-A3B (35B total, 3B active). Δ_D is 27B dense minus 4B dense; Δ_M is MoE minus 27B dense. Pattern lists increase (up), decrease (down), or tie (flat) for Δ_D/Δ_M, using unrounded differences and tolerance 10^{-9}. Scores and absolute differences are shown to four decimals. Composite units are separated and do not enter the 16-AT counts.

Unit	4B	27B	MoE	Δ_D	Δ_M	Pattern (D/M)
AT-01	0.6750	0.8688	0.6000	+0.1938	−0.2688	up/down
AT-02	0.7250	0.7562	0.5750	+0.0312	−0.1812	up/down
AT-03	0.5813	0.7937	0.5938	+0.2124	−0.1999	up/down
AT-04	0.6125	0.7438	0.6375	+0.1313	−0.1063	up/down
AT-05	0.6625	0.7688	0.6500	+0.1063	−0.1188	up/down
AT-06	0.7000	0.8438	0.6562	+0.1438	−0.1876	up/down
AT-07	0.4401	0.6126	0.5200	+0.1725	−0.0926	up/down
AT-08	0.8280	0.8538	0.8064	+0.0258	−0.0474	up/down
AT-09	0.5800	0.7367	0.6065	+0.1567	−0.1302	up/down
AT-10	0.6483	0.6997	0.6784	+0.0514	−0.0213	up/down
AT-11	0.4758	0.4606	0.4747	−0.0152	+0.0141	down/up
AT-12	0.7000	0.7000	0.5250	+0.0000	−0.1750	flat/down
AT-13	0.5500	0.7500	0.6125	+0.2000	−0.1375	up/down
AT-14	0.5813	0.6062	0.5062	+0.0249	−0.1000	up/down
AT-15	0.6687	0.6250	0.5875	−0.0437	−0.0375	down/down
AT-16	0.4740	0.5538	0.5663	+0.0798	+0.0125	up/up
MTDA	0.5278	0.6073	0.5444	+0.0795	−0.0629	up/down
ICDA	0.6631	0.7266	0.7021	+0.0635	−0.0245	up/down

Table 22: **Reporting-group membership and absolute means.** Member entries are AT suffixes, not assumed contiguous ranges. Each mean weights member ATs equally and uses final unit mean scores. MoE is Qwen3.5-35B-A3B. ADB is instead computed over all 18 units and cannot be recovered by equally averaging these five group means.

Group	Member ATs	4B	27B	MoE
InfoExtract	01–06	0.6594	0.7959	0.6188
SummaryGen	07, 10	0.5442	0.6562	0.5992
ValueInfer	08, 09, 11	0.6279	0.6837	0.6292
RelationExt	13, 14	0.5657	0.6781	0.5594
ConstraintCheck	12, 15, 16	0.6142	0.6263	0.5596

Table 23: **Three reporting-group comparisons.** Differences are taken before rounding the group means and shown to four decimals. All five groups gain in the dense comparison. InfoExtract, not SummaryGen, gains the most; SummaryGen and ValueInfer retain positive MoE-minus-4B differences.

Group	27B − 4B	MoE − 27B	MoE − 4B
InfoExtract	+0.1365	−0.1771	−0.0406
SummaryGen	+0.1120	−0.0570	+0.0550
ValueInfer	+0.0558	−0.0545	+0.0013
RelationExt	+0.1125	−0.1188	−0.0063
ConstraintCheck	+0.0120	−0.0667	−0.0546

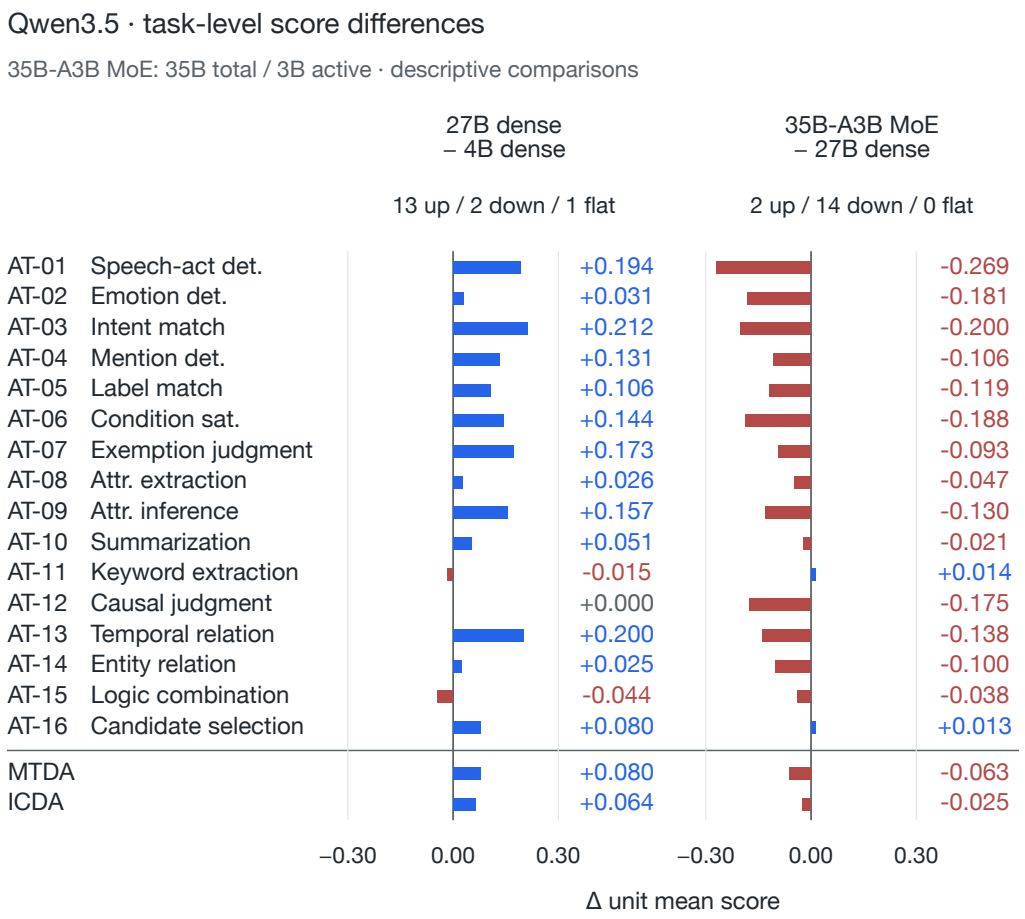

Figure 18: **Two aligned score comparisons for the Qwen3.5 family.** The left column is 27B dense minus 4B dense; the right is 35B-A3B MoE minus 27B dense (35B total, 3B active). Both use the same $[-0.30, +0.30]$ axis for absolute differences in unit mean score, with independent signed values: blue/right indicates an increase and red/left a decrease. Header counts cover only the 16 ATs: $13/2/1$ increases/decreases/ties on the left and $2/14/0$ on the right; MTDA and ICDA are separated below. Counts use unrounded differences with tolerance 10^{-9}; displayed differences are rounded to three decimals. The architecture-changing comparison is descriptive, not a controlled scaling trajectory.

Qwen3.5 · reporting-group comparisons

MoE = 35B-A3B (35B total / 3B active); group means weight member ATs equally

	Absolute mean score			Δ unit mean score		
	4B dense	27B dense	35B-A3B MoE	27B – 4B	MoE – 27B	MoE – 4B
InfoExtract 01–06	0.659	0.796	0.619	+0.136	-0.177	-0.041
SummaryGen 07, 10	0.544	0.656	0.599	+0.112	-0.057	+0.055
ValueInfer 08, 09, 11	0.628	0.684	0.629	+0.056	-0.055	+0.001
RelationExt 13–14	0.566	0.678	0.559	+0.112	-0.119	-0.006
ConstraintCheck 12, 15, 16	0.614	0.626	0.560	+0.012	-0.067	-0.055

All difference strips use –0.20 to +0.20; numbers retain the sign.

Figure 19: **Aligned reporting-group means and differences.** Five rows retain the same group order and show their actual AT members. Three columns give the absolute means; three independent columns give 27B minus 4B, MoE minus 27B, and MoE minus 4B differences. Signed values accompany lightweight diverging strips, all on $[-0.20, +0.20]$ scales, rather than truncated bars encoding absolute scores. Values are displayed to three decimals.

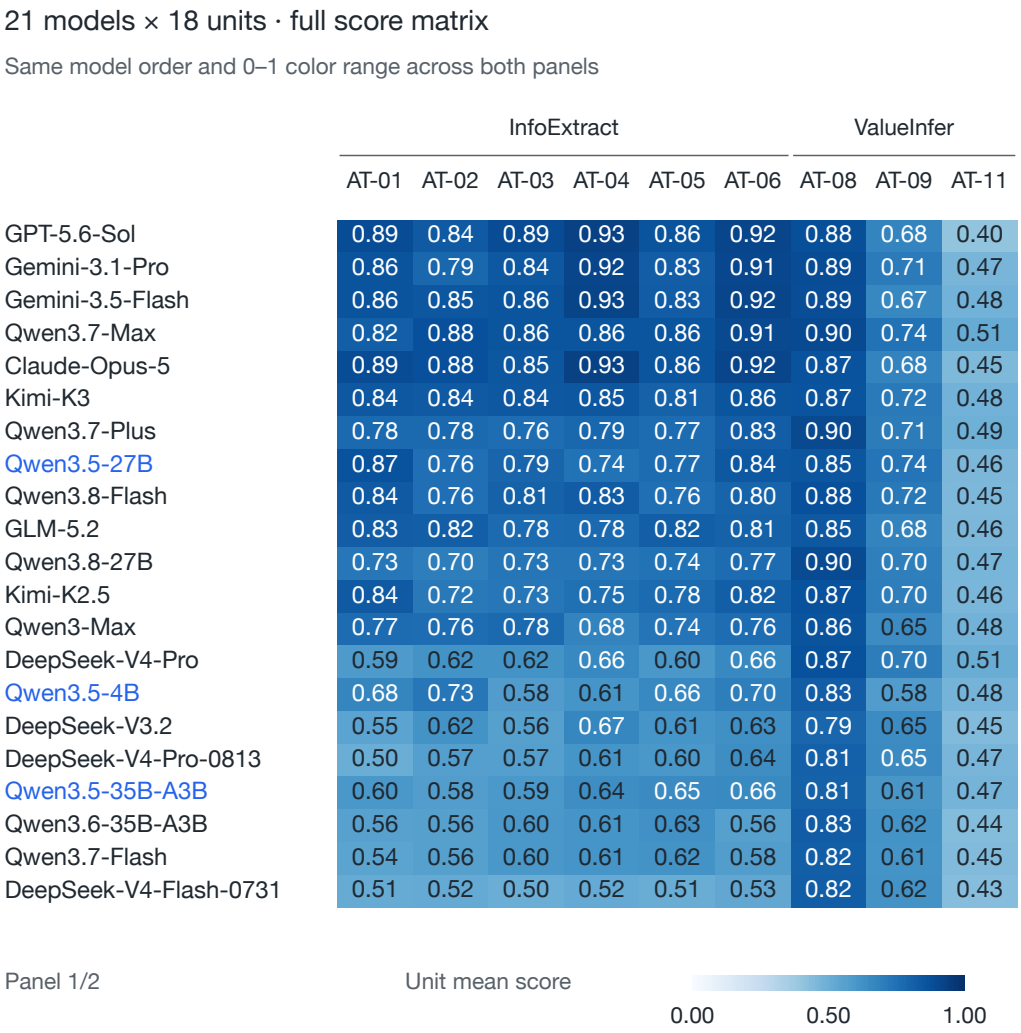

Figure 20: **Full 21-model score matrix, panel 1 of 2.** The nine columns cover InfoExtract (AT-01–06) and ValueInfer (AT-08, AT-09, AT-11). All cells show final unit mean scores to two decimals on a shared 0–1 color scale. Rows follow the 21-model ADB ranking, with the focal trio highlighted in the names. Panel 2 (Figure 21) preserves this order and scale and completes the 378-value matrix.

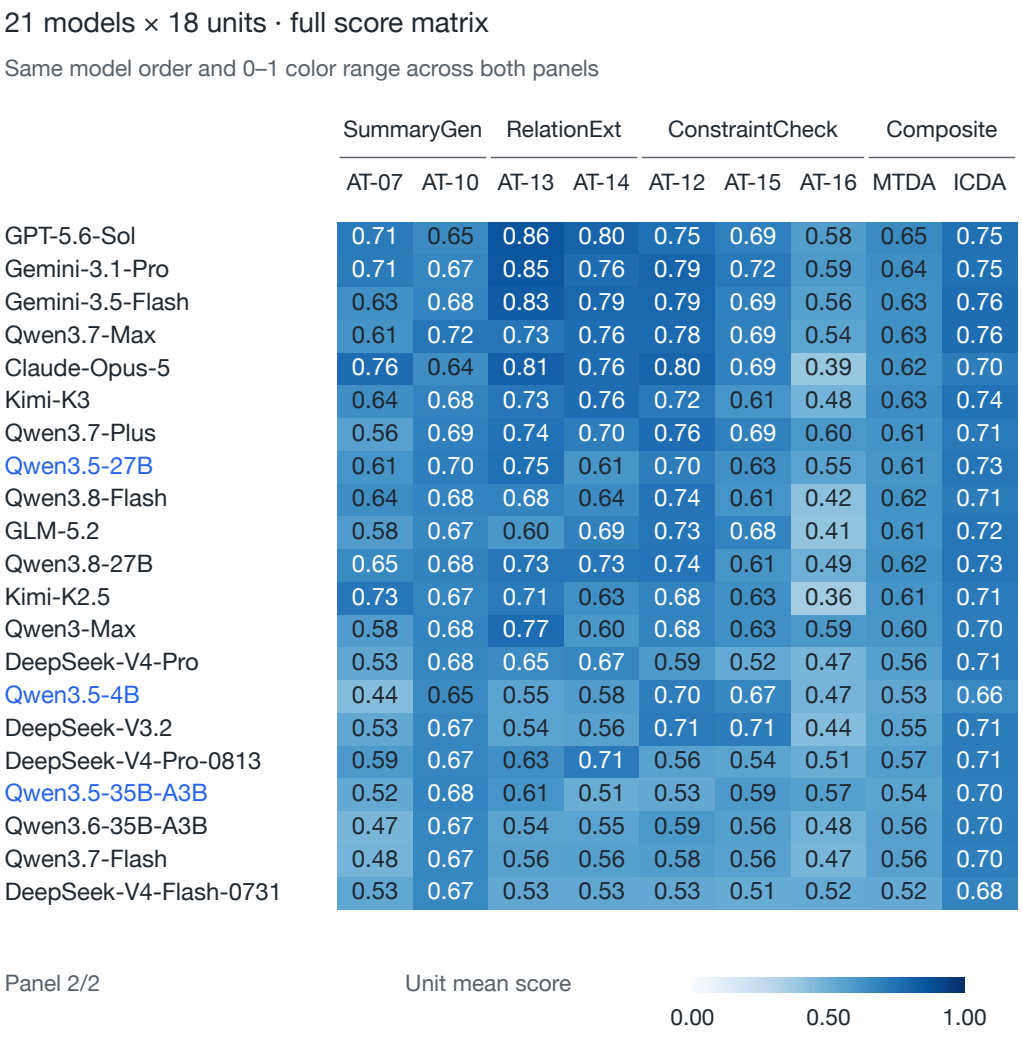

Figure 21: **Full score matrix, panel 2 of 2 (continued).** The remaining nine columns are SummaryGen (AT-07, AT-10), RelationExt (AT-13, AT-14), ConstraintCheck (AT-12, AT-15, AT-16), MTDA, and ICDA. Model order, score scale, and decimal formatting are identical to Figure 20; the two panels are one complete model-by-unit comparison, not separate subsets of selected models.

I Error Analysis: Observed Score Distributions

This appendix recomputes the error-analysis baseline from the frozen 21-model list used by the headline evaluation, excluding the judge endpoint. It uses 53,760 final case-level scores ($21 \times 16 \times 160$), not the earlier 22-model pool. For each AT, all models cover exactly the same 160 unique task IDs; IDs are also distinct across ATs. Missing or duplicate records and non-finite or out-of-range scores stop the analysis rather than becoming zeros. Each model–AT case mean is checked against its four-decimal `avg_score` with absolute tolerance 0.00005000001 and zero relative tolerance (half a four-decimal rounding unit plus floating-point allowance). The largest observed absolute difference is 0.00005. No accuracy fallback is used. ADB ranking is fixed from the frozen profile's 18-unit aggregate, cross-checked against the mean of the 18 aggregate scores with absolute tolerance 5.1×10^{-5}; it is not reranked from these 16 ATs.

I.1 Descriptive Bands, Not Cognitive Error Causes

Table 24 defines the four mutually exclusive score bands. They replace the earlier causal-sounding labels: score bins alone do not identify binary misses, logic collapse, or structural constraint violations. The same numeric interval can have different meanings under different scoring functions. These are descriptive summaries for deciding which outputs to inspect, not a common-scale taxonomy of cognitive failure.

Table 24: **Four observed score bands.** Definitions apply to final scores before rounding. Counts pool 53,760 model–case records; they do not count independent models. High-score does not mean perfectly correct.

Band	Interval	Count	Share
Zero-score	$s = 0$	12,509	23.3%
Low-score	$0 < s < 0.5$	3,269	6.1%
Partial-score	$0.5 \leq s < 0.8$	6,795	12.6%
High-score	$0.8 \leq s \leq 1$	31,187	58.0%

I.2 Score Distribution by AT and Reporting Group

Table 25 gives all 16 ATs. AT-01–06 and AT-12–15 have empty low- and partial-score bands. AT-11 has the largest zero-score share, 42.9%; AT-10 has none but 90.0% partial-score records. These statements concern the observed score distribution, not intrinsic task difficulty or output adequacy. AT-07 and AT-10 also differ substantially despite sharing a reporting group.

The five reporting groups use the actual members in Table 26, not consecutive ranges labeled Detection/Generation/Reasoning. Counts pool all models and cases within member ATs; because each model–AT has 160 cases, this is also the equal-model, equal-member-AT average of band proportions. It is not an equal-weight average over the five groups.

I.3 Hard Cases: Universal and Discriminative

A *universal-hard* case has zero scores from at least 80% of the frozen 21 models, i.e., at least 17 models. We retain the original analysis's executed $\geq 80\%$ rule; its earlier prose sometimes said $> 80\%$, which yields the same integer threshold for 21 models. The heatmap itself encodes per-model, per-AT zero-score rates, not a separate per-case difficulty threshold. A *discriminative* case has a mean score of at least 0.7 across the fixed ADB Top-5 and strictly less than 0.3 across the fixed Bottom-5. An unused 40–60% zero-rate threshold in the old script was never part of this criterion. The tier means each have five models; missing models are not dropped from these denominators. Tier membership is specified below and never reselected using the case being classified.

Across 2,560 distinct AT cases, 270 (10.5%) are universal-hard and 375 (14.6%) are discriminative; these labels overlap on 10 cases and should not be added as disjoint categories. The highest universal-hard proportions are AT-11 (36.9%), AT-16 (26.9%), and AT-15 (23.1%). AT-04 and AT-06 have the most discriminative cases, 47 each (29.4%). These counts describe this model sample under fixed thresholds, not a claim that the selected ATs optimally rank unseen models.

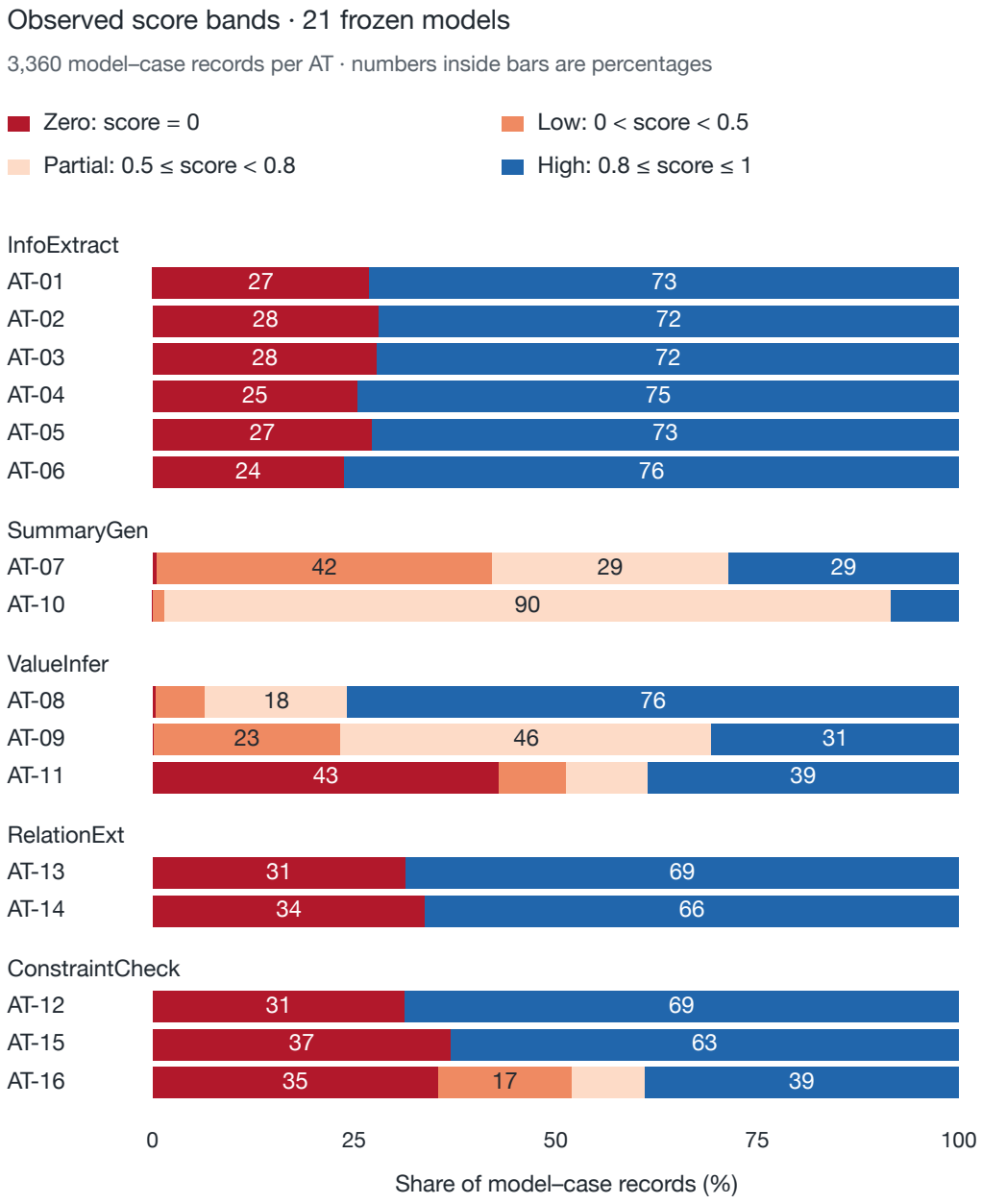

Figure 22: **Observed score-band distribution by AT.** Each horizontal bar pools 3,360 records. AT order follows the five reporting groups, including their non-contiguous membership. Numbers label segments of at least 12%, rounded to whole percentages; Table 25 gives every band. Colors encode score intervals, not inferred error causes.

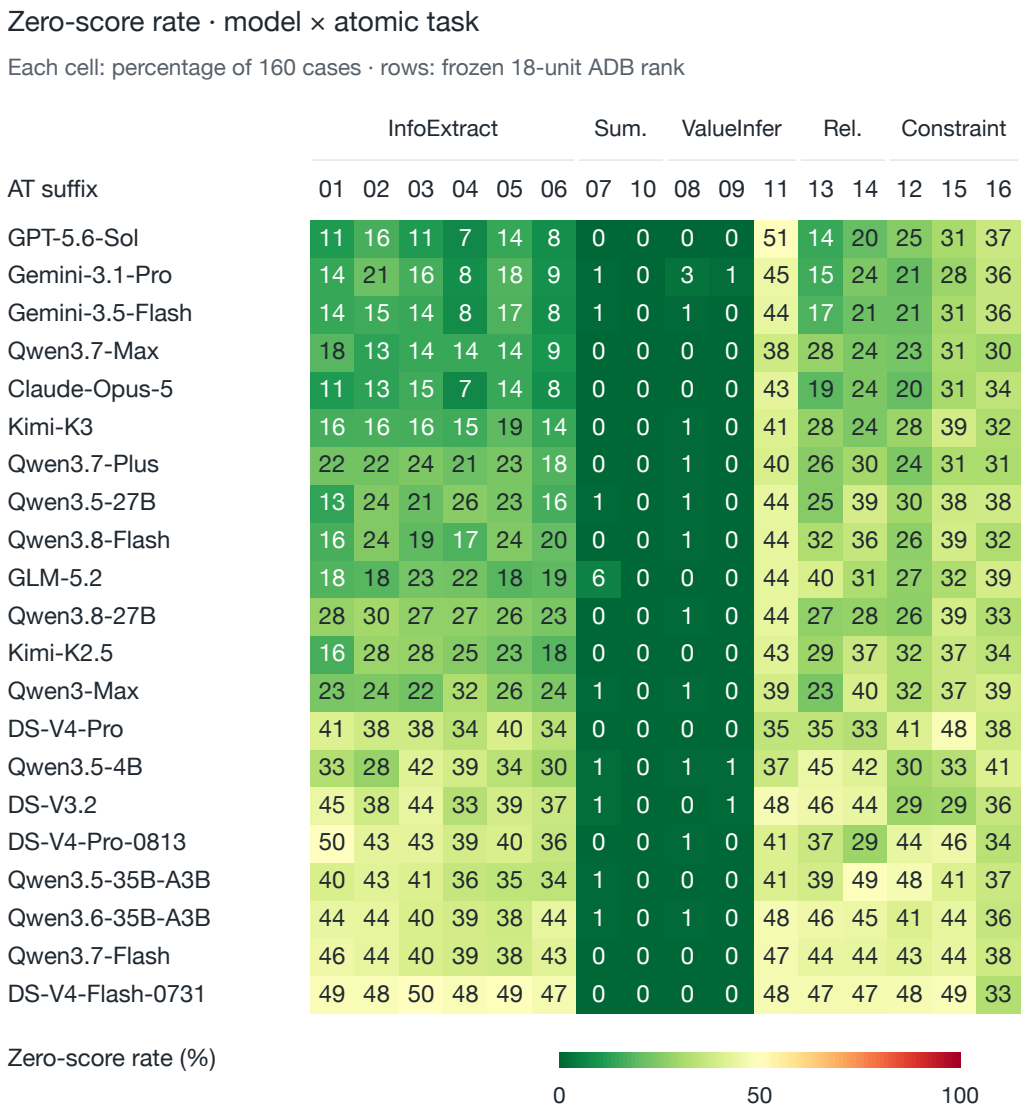

Figure 23: **Model × AT zero-score rates, not a model × case map.** Each of 336 cells is the percentage of 160 cases scoring zero, rounded to a whole percentage. Rows follow frozen 18-unit ADB rank; columns follow actual group membership (Sum. = SummaryGen, Rel. = RelationExt, Constraint = ConstraintCheck; DS = DeepSeek). Colors share a 0–100% range. Aggregate cells do not reveal which cases are shared failures; Table 27 instead uses the aligned case IDs.

Table 25: **Final score distribution by AT.** Each row has 3,360 model–case records. Entries are percentages with counts in parentheses; percentages are rounded independently to one decimal using half-up rounding.

AT	Zero-score	Low-score	Partial-score	High-score
AT-01	26.9% (905)	0.0% (0)	0.0% (0)	73.1% (2455)
AT-02	28.1% (943)	0.0% (0)	0.0% (0)	71.9% (2417)
AT-03	27.9% (937)	0.0% (0)	0.0% (0)	72.1% (2423)
AT-04	25.4% (855)	0.0% (0)	0.0% (0)	74.6% (2505)
AT-05	27.2% (915)	0.0% (0)	0.0% (0)	72.8% (2445)
AT-06	23.8% (798)	0.0% (0)	0.0% (0)	76.3% (2562)
AT-07	0.6% (19)	41.5% (1395)	29.4% (987)	28.5% (959)
AT-08	0.4% (14)	6.1% (206)	17.6% (593)	75.8% (2547)
AT-09	0.1% (4)	23.2% (780)	46.0% (1546)	30.7% (1030)
AT-10	0.0% (0)	1.5% (51)	90.0% (3025)	8.5% (284)
AT-11	42.9% (1443)	8.4% (282)	10.1% (338)	38.6% (1297)
AT-12	31.3% (1051)	0.0% (0)	0.0% (0)	68.7% (2309)
AT-13	31.4% (1055)	0.0% (0)	0.0% (0)	68.6% (2305)
AT-14	33.8% (1136)	0.0% (0)	0.0% (0)	66.2% (2224)
AT-15	37.0% (1243)	0.0% (0)	0.0% (0)	63.0% (2117)
AT-16	35.4% (1191)	16.5% (555)	9.1% (306)	38.9% (1308)

Table 26: **Reporting-group membership and score-band shares.** Member entries are AT suffixes. N is the model–case denominator; Zero, Low, Partial, and High are percentages in the four bands defined above.

Group	Members	N	Zero	Low	Partial	High
InfoExtract	01–06	20160	26.6	0.0	0.0	73.4
SummaryGen	07, 10	6720	0.3	21.5	59.7	18.5
ValueInfer	08, 09, 11	10080	14.5	12.6	24.6	48.4
RelationExt	13, 14	6720	32.6	0.0	0.0	67.4
ConstraintCheck	12, 15, 16	10080	34.6	5.5	3.0	56.9

Table 27: **Universal-hard and discriminative cases.** Every AT has 160 cases, each scored by all 21 models. Universal-hard means at least 17 zero scores; discriminative uses Top-5 mean ≥ 0.7 and Bottom-5 mean < 0.3. The total denominator is 2,560 distinct cases, not 53,760 records.

AT	Cases	Universal hard	Univ. ratio	Discrim.
AT-01	160	12	7.5%	46
AT-02	160	16	10.0%	31
AT-03	160	17	10.6%	33
AT-04	160	11	6.9%	47
AT-05	160	11	6.9%	38
AT-06	160	5	3.1%	47
AT-07	160	0	0.0%	1
AT-08	160	0	0.0%	1
AT-09	160	0	0.0%	0
AT-10	160	0	0.0%	0
AT-11	160	59	36.9%	0
AT-12	160	16	10.0%	35
AT-13	160	20	12.5%	40
AT-14	160	23	14.4%	30
AT-15	160	37	23.1%	24
AT-16	160	43	26.9%	2
Total	**2560**	**270**	**10.5%**	**375**

I.4 Fixed Model Samples and Tier Comparisons

The family-sample comparison retains the existing figure's membership without interpreting geography as a cause. Sample A has 17 models: Qwen (10, including the three pre-production aliases), DeepSeek (4), GLM (1), and Kimi (2). Sample B has four: GPT (1), Gemini (2), and Claude (1). These are unequal, unmatched samples, not representative estimates for regions or developers. Their overall zero-score shares are 25.1% (10,929/43,520) and 15.4% (1,580/10,240), respectively. No size, architecture, or training effect is identified, and pooled records are not treated as independent models.

Top-5 consists of GPT-5.6-Sol, Gemini-3.1-Pro, Gemini-3.5-Flash, Qwen3.7-Max, and Claude-Opus-5 (frozen ADB ranks 1–5). Bottom-5 consists of DeepSeek-V4-Pro-0813, Qwen3.5-35B-A3B, Qwen3.6-35B-A3B, Qwen3.7-Flash, and DeepSeek-V4-Flash-0731 (ranks 17–21). Ranking uses the frozen 18-unit profile, with lexicographic model-key tie-breaking; the former 16-AT reranking is not used.

Table 28 shows the largest zero-score difference on InfoExtract (29.7 pp), followed by RelationExt (22.1 pp) and ConstraintCheck (12.6 pp). SummaryGen has identical zero-score counts, yet high-score shares are 23.3% versus 13.5%. ValueInfer has nearly identical zero-score shares but high-score shares of 50.1% versus 43.5%. Consequently, similar zero-score rates do not establish similar output quality or failure causes.

Table 28: **Zero-score rates by fixed ADB tier and reporting group.** N is the record denominator for each tier. Rates are rounded to one decimal; Δ is Bottom-minus-Top in percentage points, computed before rounding. ValueInfer's unrounded difference is -0.0417 pp, shown as 0.0 at this precision.

Group	N	Top-5	Bottom-5	Δ (pp)
InfoExtract	4800	12.6%	42.3%	29.7
SummaryGen	1600	0.1%	0.1%	0.0
ValueInfer	2400	15.0%	15.0%	0.0
RelationExt	1600	20.5%	42.6%	22.1
ConstraintCheck	2400	29.0%	41.5%	12.6
Overall	12800	15.5%	31.8%	16.3

I.5 AT-16 Output-Contract Sensitivity

Table 29 locates an output-expression sensitivity in AT-16: 159 records improve under candidate-identity normalization, with 109 reaching F1 $= 1$. The official metric is candidate-set F1; this diagnostic leaves frozen scores, ADB values, and rankings unchanged.

Table 29: Identity normalization increases pooled mean F1 by 0.0753 on the same 1,687 eligible AT-16 model–task records, without altering frozen scores.

Quantity	Value
Coverage	
Reliable input-only catalogue tasks	93/160 (58.1250%)
Eligible paired model–task records	1,687/3,360 (50.2083%)
Same-subset sensitivity (1,687 records)	
Original mean F1	0.6153
Identity-normalized mean F1	0.6906
Paired change (normalized − original)	+0.0753
Outcomes	
Improved	159/1,687
Of improved: fully recovered (F1 = 1)	109/159
Unchanged	1,528/1,687
Decreased	0/1,687

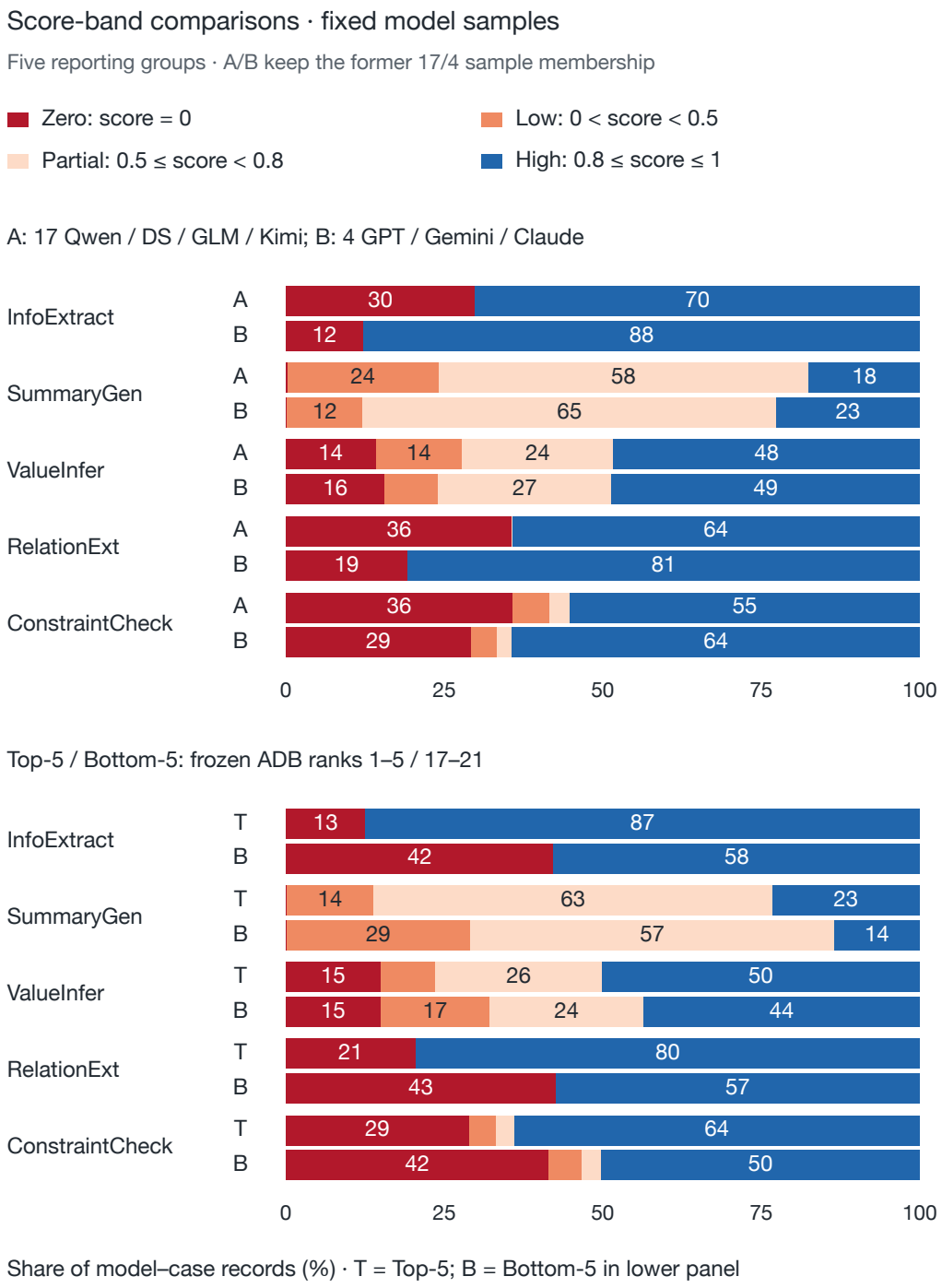

Figure 24: **Score-band composition for fixed model samples.** Upper panel: sample A (17 Qwen/DeepSeek/GLM/Kimi models) versus B (4 GPT/Gemini/Claude models). Lower panel: frozen ADB Top-5 (T) versus Bottom-5 (B). Each bar pools 160× member ATs × sampled models; labels show whole percentages for bands of at least 12%. Membership follows the five groups in Table 26; comparisons are descriptive.

Candidate catalogues are constructed from task inputs alone; prediction and reference candidates are mapped to unique, explicit identities after whitespace/case normalization. No aliases are derived from gold, name initials are not inferred as IDs, and neither semantic matching nor extra-candidate trimming is used. Ambiguous or incomplete mappings and parsing/reference issues exclude the whole pair. The 1,673 excluded records comprise 1,407 without reliable catalogues, 240 with catalogues but unknown candidates, and 26 with other parsing/reference issues.

Both means pool the same 1,687 eligible records; the full pool comprises $21 \times 160 = 3{,}360$ repeated model–task observations, not independent models. Coverage is bounded by input catalogues and candidate mappings: this is scoring sensitivity, not new model performance or evidence that semantic failures are generally repaired.

J CROSS-LINGUAL ANALYSIS

This appendix details the Chinese/English (ZH/EN) gap summarized in the Discussion. We split final case-level scores by the `zh`/`en` task labels, retaining $21 \times 16 \times 160 = 53{,}760$ model–case observations. Each AT has 80 cases per language, matched across models by task ID. The two language subsets contain different content, not paired translations; the differences therefore do not isolate a language effect. Language-specific model means weight the 16 ATs equally and exclude MTDA and ICDA.

J.1 SAMPLE IMBALANCE AND AGGREGATION

The existing developer-region grouping assigns 17 models to the domestic sample and four to the international sample. Comparisons first average the 16 AT language means within each model, then average equally across models in each sample. Error bars report the cross-model population standard deviation of these macro-means (divisor n, not $n-1$), not a confidence interval. The family distribution is Qwen 10, DeepSeek 4, Kimi 2, Google 2, GLM 1, OpenAI 1, Anthropic 1. The 17:4 groups are not matched for configurations or training, and most parameter counts are unknown; their differences describe this sample, not a geographic or size effect.

J.2 PER-MODEL CHINESE/ENGLISH GAP

Table 30 reports each model's ZH mean, EN mean, and signed EN$-$ZH gap. The gap is positive for 19 of the 21 models. Claude-Opus-5 has the largest positive gap ($+0.0627$); DeepSeek-V4-Pro has the smallest absolute overall gap ($|\text{gap}| = 0.0003$). Qwen3-Max (-0.0047) and DeepSeek-V4-Pro-0813 (-0.0012) have negative gaps. These macro-averages can hide opposing differences across ATs.

Table 30: **Per-model Chinese/English scores and gap** (EN$-$ZH), sorted by signed gap descending. Means equally weight the 16 ATs; 19 gaps are positive and two are negative in this model sample. The unpaired language subsets do not provide controlled estimates of language effects.

Model	Region	ZH	EN	EN$-$ZH
Claude-Opus-5	Foreign	0.7303	0.7930	+0.0627
Gemini-3.1-Pro	Foreign	0.7431	0.7956	+0.0526
GPT-5.6-Sol	Foreign	0.7501	0.7925	+0.0423
Kimi-K3	Domestic	0.7136	0.7512	+0.0376
Gemini-3.5-Flash	Foreign	0.7498	0.7810	+0.0312
Qwen3.5-35B-A3B	Domestic	0.5842	0.6153	+0.0310
Kimi-K2.5	Domestic	0.6769	0.7075	+0.0306
Qwen3.6-35B-A3B	Domestic	0.5657	0.5935	+0.0278
Qwen3.7-Plus	Domestic	0.7089	0.7352	+0.0263
Qwen3.5-27B	Domestic	0.6977	0.7239	+0.0262
DeepSeek-V3.2	Domestic	0.5928	0.6177	+0.0248
Qwen3.7-Max	Domestic	0.7494	0.7701	+0.0207
Qwen3.7-Flash	Domestic	0.5697	0.5870	+0.0173
DeepSeek-V4-Flash	Domestic	0.5405	0.5569	+0.0165
Qwen3.8-Flash	Domestic	0.6951	0.7115	+0.0164
Qwen3.8-27B	Domestic	0.6854	0.7012	+0.0157
GLM-5.2	Domestic	0.6933	0.7023	+0.0091
Qwen3.5-4B	Domestic	0.6156	0.6222	+0.0066
DeepSeek-V4-Pro	Domestic	0.6205	0.6208	+0.0003
DeepSeek-V4-Pro-0813	Domestic	0.6026	0.6013	−0.0012
Qwen3-Max	Domestic	0.6909	0.6862	−0.0047

J.3 PER-AT LANGUAGE SENSITIVITY

Table 31 sorts the 16 ATs by absolute mean EN$-$ZH gap. The largest absolute gaps are on AT-11 keyword extraction ($+0.1044$), AT-14 entity relation ($+0.0976$), and AT-08 attribute extraction ($+0.0596$); the smallest are on AT-12 causal judgment (-0.0030) and AT-16 candidate selection

(+0.0055). The descriptive subset AT-01–06 averages +0.0019 (SD 0.022); AT-07–10 averages +0.0209 (SD 0.034); attribute extraction/inference (AT-08/09) averages +0.0441 (SD 0.015). These three SDs describe dispersion across the member AT mean gaps, not across models. These legacy subsets do not define additional abilities, and the five formal groups remain reporting groups rather than independent latent dimensions. Signed differences can cancel, so a small mean gap does not demonstrate robustness; the different-content language samples also preclude attribution to language alone.

Table 31: **Per-AT language gap** (EN−ZH), sorted by |EN−ZH| descending. ZH and EN columns average the language-specific final scores equally over 21 models. SD is the population standard deviation of the 21 model-level signed gaps (divisor 21), not a confidence interval. Mean gap and cross-model dispersion need not rank tasks in the same order.

AT	Name	ZH	EN	EN−ZH	cross-model SD
AT-11	Keyword extraction	0.413	0.518	+0.1044	0.035
AT-14	Entity relation	0.613	0.711	+0.0976	0.101
AT-08	Attr. extraction	0.827	0.886	+0.0596	0.058
AT-15	Logic combination	0.604	0.656	+0.0518	0.063
AT-10	Summarization	0.692	0.657	−0.0348	0.008
AT-05	Label match	0.713	0.743	+0.0304	0.044
AT-07	Exemption judgment	0.580	0.610	+0.0300	0.037
AT-09	Attr. inference	0.659	0.688	+0.0287	0.047
AT-03	Intent match	0.735	0.708	−0.0268	0.048
AT-13	Temporal relation	0.675	0.697	+0.0220	0.041
AT-06	Condition sat.	0.752	0.773	+0.0202	0.050
AT-02	Emotion det.	0.729	0.710	−0.0185	0.057
AT-04	Mention det.	0.737	0.754	+0.0173	0.027
AT-01	Speech-act det.	0.736	0.725	−0.0113	0.062
AT-16	Candidate selection	0.496	0.501	+0.0055	0.051
AT-12	Causal judgment	0.689	0.686	−0.0030	0.048

J.4 Domestic versus International Families

Table 32 retains the two-stage aggregation for the existing samples. The four international models have a higher Chinese mean than the 17 domestic models (0.7433 vs. 0.6472, a 0.0961 gap) and a larger EN−ZH difference (+0.0472 vs. +0.0177). The Chinese-only contrast varies by AT: domestic minus international is +0.0417 on AT-11 keyword extraction and −0.2169 on AT-04 mention detection. These differences are descriptive; unequal sample composition and unmatched configurations and training prevent geographic or parameter-size attribution.

Table 32: **Region-labeled sample means by language**. Two-stage aggregation: equal-weight mean over 16 ATs per model, then equal-weight mean over the n models in each sample. These unmatched 17:4 samples do not estimate a regional effect.

Region	n	ZH	EN	EN−ZH
Domestic	17	0.6472	0.6649	+0.0177
International	4	0.7433	0.7905	+0.0472

J.5 AT-09 Hard-match Component Check

For AT-09 attribute inference, the recorded hard-score means are 0.401 (ZH) and 0.415 (EN); the judge-score means are 0.865 (ZH) and 0.844 (EN), while final-score means are 0.659 (ZH) and 0.688 (EN). These are marginal summaries of different scoring components, not a decomposition of the final language gap. In particular, the judge mean is lower in English while the final mean is higher. The `hybrid_judge_value` scoring path includes conditional fallback, so component marginal means alone cannot recover the final mean or identify the source of its difference. We retain the component values as descriptive scoring context, without attributing the final gap to hard matching, the judge, or language alone.

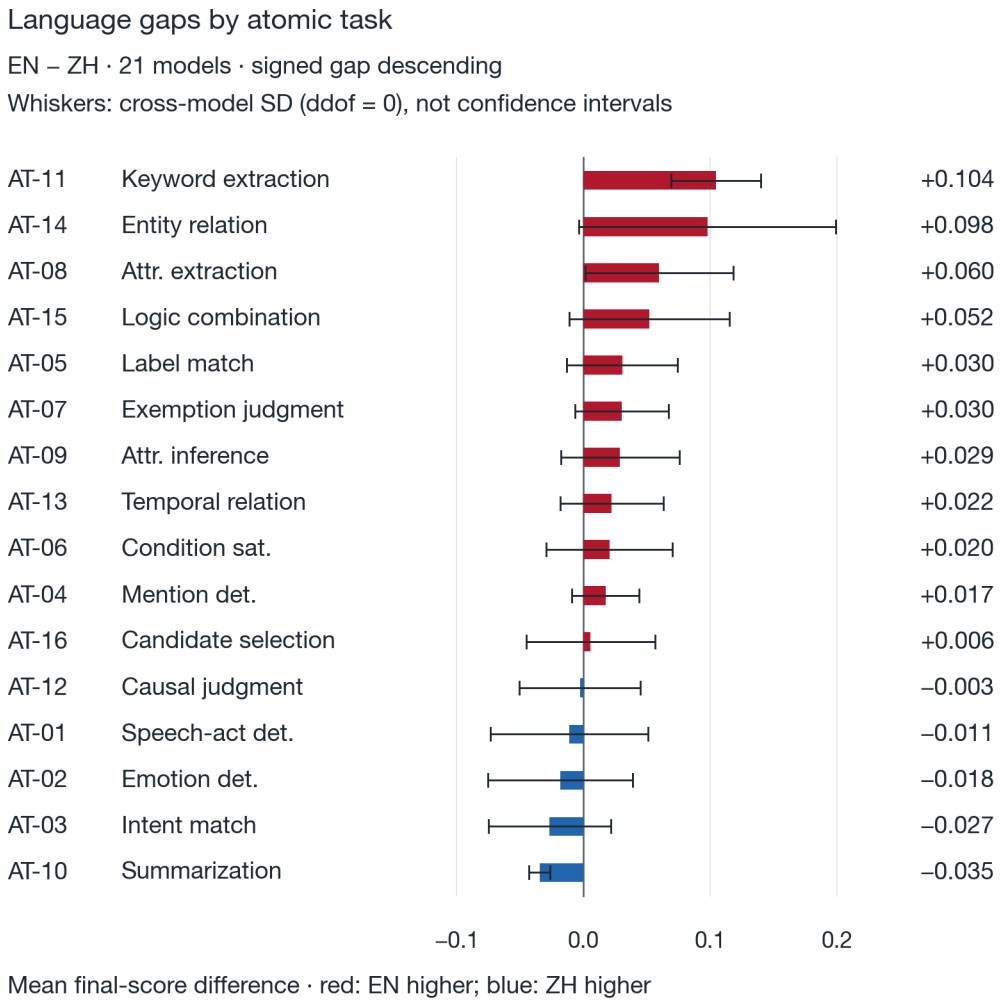

Figure 25: **Per-AT language-gap ranking.** Bars show mean final-score differences (EN−ZH), ordered by signed gap descending from top to bottom, unlike the absolute-gap ordering in Table 31. Red/right indicates a higher English mean; blue/left indicates a higher Chinese mean. Whiskers are cross-model SD with divisor 21, not confidence intervals. Both signs occur, and a near-zero mean can conceal model-level differences.

Ethics Details

Data Privacy and Licensing

The public-data construction pool retains 13,316 dialogues from 15 research datasets after filtering 24 candidate sources. It covers customer service, emotional support, task-oriented dialogue, knowledge-grounded conversation, and everyday conversation. Public availability does not itself establish redistribution rights. Source licenses and access conditions remain applicable to both original dialogues and derived instances; a release must distinguish redistributable material from source pointers and sampling metadata.

Dialogue-length and DCI filters select analytical material; they are not PII detectors or de-identification procedures. LLM synthesis may retain identifying details or introduce unsupported personal attributes. Neither filtering nor synthesis establishes that residual privacy risk is absent. Application-derived templates are described without proprietary business names or rules. The separate business-domain transfer material must not be treated as automatically covered by the public-

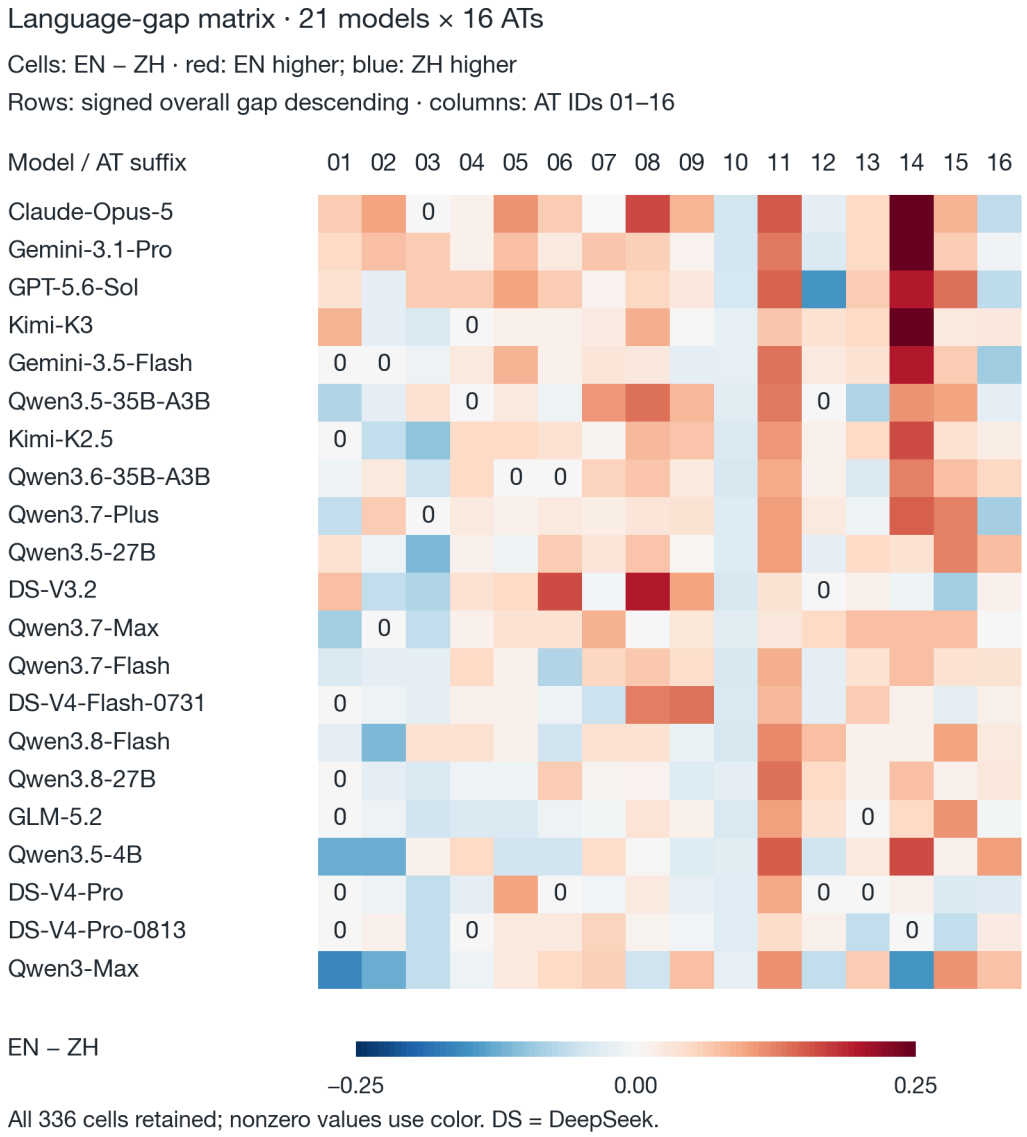

Figure 26: **Model × AT language-gap heatmap** (EN−ZH). All 21 × 16 cells are retained, with models ordered by signed overall gap descending and ATs in numerical order. Red indicates a higher English mean and blue a higher Chinese mean on the same symmetric scale. Exact zero cells are labeled; other values are encoded by color. The matrix preserves opposing model–task differences that can cancel in the aggregate; it does not isolate language from sample content. Task names are given in Table 31.

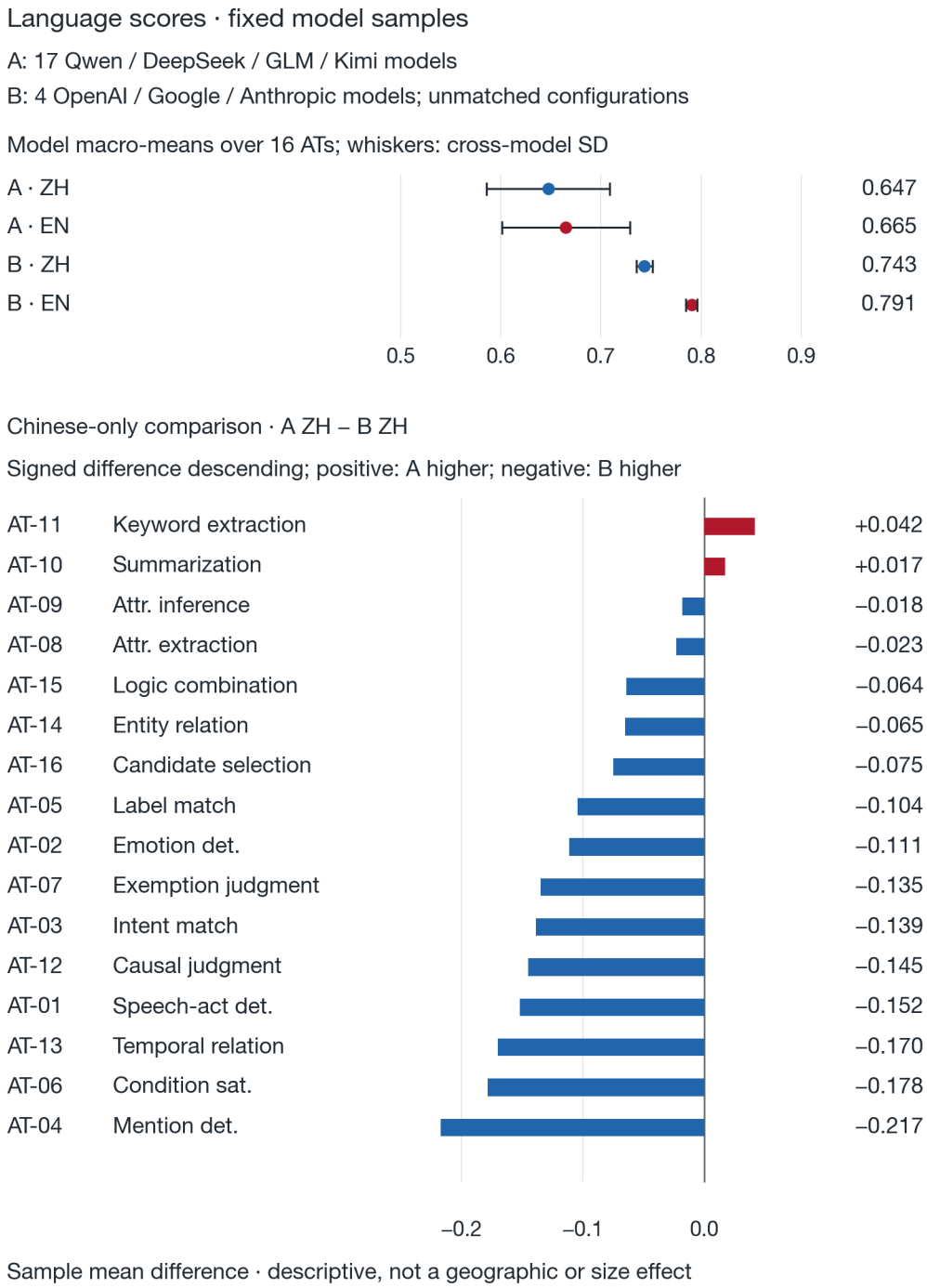

Figure 27: **Chinese/English comparison for fixed model samples.** A retains the 17 domestic models and B the four international models. Upper-panel points show each sample's ZH and EN means; whiskers are cross-model SD of the 16-AT model macro-means (divisor n), not confidence intervals. The lower panel shows A ZH minus B ZH by AT, ordered by signed difference descending, not an EN−ZH contrast. Sample composition, configurations, and training are unmatched; no geographic or size effect is identified.

data release permissions. Authorization and privacy review remain prerequisites for releasing such material.

BENCHMARK MISUSE RISKS

AtomDialBench is a diagnostic benchmark, not deployment certification. Its 18-unit average should be accompanied by task profiles; external dialogue and instruction-following scores remain separate. Neither a high overall rank nor a strong task score is sufficient for a safety-critical decision.

Results describe the evaluated API snapshots, which providers may subsequently change. Cached outputs support inspection of those snapshots but not a guarantee of identical future responses. The Chinese/English balance and hard/easy target quotas are sampling design choices, not a verified final difficulty split, a population model, or proof of maximum discrimination. Application-specific checks remain necessary, especially for uncovered domains and languages.

LLM-AS-JUDGE BIAS AND MITIGATION

Phase 3 of our evaluation pipeline employs an LLM as a judge to score model outputs on a 0–10 rubric. We acknowledge the following known biases and describe our mitigation strategies.

Known biases. *Self-enhancement bias*: an LLM judge may systematically favor outputs from models in its own family or matching its own generation style. *Position bias*: the judge may prefer the answer presented first (or last) regardless of quality. *Verbosity bias*: longer, more elaborately phrased answers may receive higher scores even when the additional content is not substantive. *Language/style preference*: the judge may have systematic preferences for certain languages, registers, or formatting conventions.

Mitigation measures. On applicable judge-scored paths, the protocol requests two evaluations with reference and candidate positions swapped. The prompt explicitly identifies their roles in both orders, and the implementation averages valid returned scores. It is therefore reference-aware and position-swapped, not blind. This aims to reduce position bias; it does not establish order independence or eliminate preference bias. Temperature zero reduces one source of variation without guaranteeing deterministic service behavior.

Deterministic-only tasks retain hard scores. Hybrid paths use equal hard/judge weighting when the hard score is positive and a calibrated judge-only fallback otherwise. MTDA combines per-AT scoring with judging; ICDA uses a separate professional-judge path. Qwen3.6-Flash is excluded from the main 21-model comparison. However, because Qwen3.6-Flash belongs to the same Qwen model family as 10 of the 21 evaluated models, same-family scoring preference has not been quantified and remains a validity limitation of the current evaluation.

Three computer-science researchers independently rated 900 stratified cases spanning the 18 units and were paid at the local average rate. The ratings used for human-agreement statistics were human-provided; this does not exclude model-generated instances or reference answers. Human–human ICC(2,1) is 0.918, while system–human case-level Spearman correlation is 0.697. By evaluation unit, correlations are 0.710 for AT-01–16 ($n = 800$), 0.557 for MTDA ($n = 50$), and 0.513 for ICDA ($n = 50$). The AT subset includes deterministic, hybrid, and judge-only fallback paths; these groups are not scoring-method categories. Appendix E reports the two-model sampling design. The differing subsets do not provide a matched comparison of scoring methods.

Synthesizer inclusion. Qwen3.7-Max synthesizes the AT cases and is also evaluated (rank 4 of 21), creating a potential synthesizer-specific advantage that has not been quantified. Standardized prompts and its observed rank do not bound that bias. Excluding Qwen3.6-Flash from the main comparison prevents direct self-scoring there, but does not eliminate possible same-family preference. A synthesizer-excluded validation is a necessary next step before drawing conclusions about Qwen3.7-Max's relative standing; readers should treat its rank as potentially inflated until such validation is completed.

Potential Societal Impact

Positive impact. AtomDialBench helps researchers and practitioners identify weaknesses in LLMs used for dialogue analysis and prioritize application-specific checks. Its 16 atomic tasks in five reporting groups, plus two composite units, provide task-level information alongside aggregate scores; neither replaces safety evaluation before deployment.

Dual-use considerations. Dialogue analysis can support useful review and summarization, but can also facilitate profiling, surveillance, or harmful model selection. Including analysis rather than interaction tasks does not remove these risks, and summary generation is itself part of the benchmark. Results should be used with context-specific authorization and human oversight.

Resource use. Evaluation involves API inference, selective judging, and local scoring. Coverage differs across model sets and scoring paths, so multiplying every case by every model and two judges does not yield a defensible total. We do not report an unaudited API-call, monetary-cost, or runtime estimate. Caching avoids some repeated inference; changed judge prompts or missing judge outputs still require additional calls.

Equity considerations. The self-constructed SynthAT tasks include balanced Chinese and English cases, while the external comparison benchmarks (external dialogue and IF benchmarks) span both languages as inherited from their original sources. However, the benchmark does not cover low-resource languages, and performance insights may not transfer to non-covered language communities.

Reproducibility Details

Reproducibility has three distinct levels here: identifying the frozen inputs and reported results, checking cached records against those results, and rerunning models and scorers. Existing artifacts support inspection, but their presence is not evidence that the full pipeline has been replayed successfully.

Code and Result Artifacts

The repository contains inference and scoring code under `dialogue_eval/`, task instances in its `tasks/` subdirectory, and cached inference, hard-match, judge, and final-result records in separate subdirectories. Analysis material is under `docs/research_at_benchmark/analysis/`. The frozen paper summary is `docs/research_at_benchmark/data/eval_results.json`; its metadata records the 2026-09-01 snapshot and source revision `f79f433`. That provenance should be distinguished from the current working implementation and any later verification result.

The paper compares 21 models on 18 ADB units. External coverage comprises 17 models on 12 dialogue and four instruction-following benchmarks, not 21 models on every unit. Code and permitted artifacts will be released upon acceptance; access and licensing must be documented in the actual release rather than inferred from a local path. Proprietary model weights and restricted source dialogues are not covered by a blanket code license.

Data Licensing and Access

The suite totals are 3,459 ADB instances, 4,471 external dialogue instances, and 3,353 instruction-following instances. ADB contains $16 \times 160 + 499 + 400$ instances. MTDA has 500 source records but one repeated `task_id`, yielding 499 evaluation cases. The IF sources contribute 300 (IFBench), 999 (CFBench), 1,012 (InverseIF), and 1,042 (GuideBench) cases. These are instance counts, not numbers of independent dialogues. Redistribution of raw or derived material remains subject to source terms; restricted material requires source pointers and selection metadata rather than an unrestricted-data promise.

Table 33: Artifact inspection and execution boundaries. None of these checks alone establishes full reproduction.

Operation	Required material	Boundary
Input integrity	Manifest and covered inputs	Does not cover output caches
Frozen-score audit	Final records and frozen summary	Read-only comparison, no rescoring
Offline scoring	Cached outputs and path-specific scorers	May write results; not a universal replay
New inference/judging	Model access and configuration	Paid calls and version drift possible

Environment and Execution Settings

Python dependencies are specified in the repository requirements. API-based evaluation uses local scoring alongside remote inference; the training experiments have separate accelerator and training configurations. Evaluated-model calls request temperature zero where supported. GPT-5.6-Sol and Claude-Opus-5 use provider defaults when explicit temperature is rejected; the main judge requests temperature zero. Saved sampling and bootstrap seeds describe those procedures, not a guarantee that concurrent API execution or future service responses are deterministic. Position swapping keeps reference roles visible and averages valid returns; it does not guarantee order-independent judging.

Cached Results and Verification Boundaries

A read-only audit should identify model and task IDs, compare case counts and final scores with the frozen summary, and record any discrepancy before changing artifacts. For AT-07, `aggregate_results.py` reads cached `score`, not `hard_score`. Existing records include hybrid and calibrated judge-only paths. An AT-07-only check used `compute_stats` on these final scores for the 21 paper models and, separately, `exp1_all_at_4B`. Case counts (160 each) and means (to four decimal places) matched the frozen `eval_results.json`, including Qwen3.7-Max's mean of 0.6144. This check did not cover other tasks, rerun inference or judging, or replay the full pipeline. A discrepancy would require a provenance investigation, not an automatic overwrite or silent replacement of the reported score.

There is no verified universal offline replay command. In `run_scoring.py`, `--tasks all` has no special all-task semantics, while `--models all` selects the registered model list, including training or judge models beyond the paper's 21. Selecting `SynthAT` as a prefix alone does not solve the remaining path differences. ICDA uses a separate professional-judge path; generic aggregation must not overwrite it. External dialogue benchmarks use native scoring, including CSDS `native_scorer_v2`. IFBench, CFBench, and InverseIF have their own result paths; GuideBench reads six sub-suites from the instruction-benchmark result directory. Unused external summary-judge records must not be counted as consumed scoring evidence.

The existing `data_check.py` still expects 500 MTDA records and uses a case-weighted average. It therefore does not directly verify the paper's 499-case MTDA and equal-weighted 18-unit ADB average. Likewise, checksum `--verify` covers the manifest's inputs, not inference or result caches. Input integrity, score agreement, and complete replay are separate claims.

Resource Requirements

Table 33 distinguishes operations rather than supplying an unaudited cost estimate. Counts of paid calls depend on actual coverage, consumed scorer outputs, retries, and path-specific judging. Neither a universal two-judge multiplier nor an estimate omitting IF scoring gives a defensible full total. Monetary costs and wall-clock times are therefore not estimated here.

Multi-Judge Ablation Reproducibility

The saved eight-judge ablation covers AT-07–10, 11 evaluated models, and the first 50 cases per AT, with self-scoring included. After excluding a few non-comparable pairs, 105 judge-pair–AT ICC(2,1) estimates average 0.696. The AT-09 value 0.964 is an eight-judge Spearman–Brown prediction. Per-judge records and analysis are under

`dialogue_eval/multi_judge_ablation/`; verification must preserve these sample and statistic definitions.

HUMAN EVALUATION REPRODUCIBILITY

The 3-annotator human verification study uses three independent human annotators (researchers with computer-science backgrounds) on a stratified sample of 900 cases covering all 18 evaluation units (inter-annotator ICC(2,1) $= 0.918$, ICC(2,k) $= 0.971$, mean Spearman $= 0.921$, Cohen's $\kappa = 0.867$). The ratings used in these agreement statistics were supplied by the three human researchers, paid at the local average rate; this statement does not exclude model-generated instances or reference answers. Documentation includes the sampled cases (`sample_cases_900.jsonl`), the individual human annotations (`annotator_{A,B,C}.jsonl`), and the agreement-computation script (`compute_agreement.py`), all under `dialogue_eval/human_verification/`. These files enable independent verification of our reported human–system correlation metrics.