From Static Reports to Interactive Dialogues: A Dual-Modality Benchmark and Distillation Framework for Engineering Capability Assessment

Anonymous ACL submission

Abstract

Engineering-education accreditation increasingly requires structured, evidence-grounded assessment of learner capability across the eleven Washington-Accord graduate-attribute dimensions, yet existing LLM-as-judge benchmarks return coarse holistic scores without dimension-level traceability and overlook the rapidly emerging modality of human–AI interaction transcripts. We introduce ENGREDUBENCH, a Chinese-language benchmark and methodology that addresses both gaps. The dataset contains 3,653 evidence sets paired with eleven-dimension capability scores spanning ten engineering domains and four life-cycle stages: 480 are dual-annotated by domain experts (mean quadratic-weighted Cohen's $\kappa = 0.706$, substantial agreement), and 3,173 are produced by an LLM-based distillation pipeline that selects per-dimension teachers from a pool of eight frontier models, pairs each with an independent supervisor, and applies a calibrated teacher–supervisor disagreement filter (post-filter sample-level keep rate 79.3%). A multi-agent evaluator pipeline provides fully traceable per-dimension rationales for every score. We release the dataset, prompts, evaluator code, agreement statistics, and rejection thresholds as a single audited substrate for downstream applications.

1 Introduction

Engineering education accreditation bodies—ABET in the United States (ABET Engineering Accreditation Commission, 2024), the Washington Accord internationally (International Engineering Alliance, 2021), and the China Engineering Education Accreditation in our setting—demand that graduates be assessed against a structured set of capability dimensions covering technical knowledge, design, ethics, communication and lifelong learning (Ho et al., 2024). In practice this assessment is performed by human instructors who manually read student deliverables and assign rubric-aligned scores. The process is labour-intensive, inconsistent across raters, and scales poorly as cohorts grow.

Recent work on *LLM-as-judge* (Liu et al., 2023; Zheng et al., 2023; Kim et al., 2024a) has demonstrated that frontier large language models can produce evaluations that correlate moderately well with human judgements across open-ended tasks. However, three obstacles block the transfer of these advances into engineering-education assessment.

Obstacle 1: Lack of traceability. Existing LLM-as-judge benchmarks return holistic scores or single-axis ratings; they rarely provide evidence-grounded rationales that an instructor can audit. In educational assessment, traceability is not a nice-to-have but a regulatory prerequisite: every score must be defensible against student appeal.

Obstacle 2: Static-only evaluation. Almost all prior engineering-education assessment infrastructure consumes *static* artefacts—reports, design documents, code submissions. With the rapid uptake of AI assistants, however, an increasing share of student engineering work is performed as a *dynamic dialogue* between the learner and one or more AI tools. The resulting human–AI interaction transcripts contain rich evidence of the learner's reasoning, trade-offs, and reflection that is invisible to a static-artefact evaluator. To our knowledge no public benchmark addresses this emerging modality.

Obstacle 3: Closed evaluators. The strongest LLM judges are proprietary frontier models whose per-call cost and rate limits make them unsuitable for routine institutional use. Bridging this gap requires distilling judge capability into open mid-size models, but doing so without inheriting a single teacher's biases is non-trivial.

Our work. We introduce ENGREDUBENCH, a Chinese-language benchmark and accompanying methodology for engineering-capability assessment that addresses all three obstacles. We contribute (i) a **dual-modality dataset** of 3,653 evidence sets (480 expert-annotated golden + 3,173 distilled) paired with eleven-dimension Washington-Accord capability scores, spanning ten engineering domains and four life-cycle stages; the 480 golden samples are dual-annotated and attain a mean quadratic-weighted Cohen's $\kappa = 0.706$ (substantial agreement) across the eleven dimensions; (ii) a **multi-agent evaluation pipeline** that decouples evidence organisation, per-dimension scoring, programmatic aggregation, summarisation, and post-hoc quality audit, producing fully traceable rationales for every score; (iii) a **three-role distillation framework** that selects the strongest single-dimension teacher from a pool of eight candidate frontier LLMs and pairs it with an independent supervisor for noisy-sample filtering, with a parallel summary teacher/supervisor for the two free-form fields; and (iv) a **public, audited resource**—dataset, prompts, agreement statistics, evaluator code, and rejection thresholds—released so that downstream applications can be reproduced and audited.

The remainder is organised as follows. Section 2 reviews related work, Section 3 formalises the task and rubric, Section 4 describes the data sources and annotation, Section 5 details the multi-agent pipeline and three-role distillation, Section 6 reports evaluator-selection, threshold-calibration, and distillation-quality experiments, and Section 7 discusses implications and limitations.

2 Related Work

2.1 LLM Judges and Educational Assessment

LLM-as-judge methods evaluate open-ended generation by prompting a strong LLM with a rubric and asking it to score or compare outputs (Liu et al., 2023; Zheng et al., 2023). Distilled open judges such as JudgeLM (Zhu et al., 2025), Prometheus (Kim et al., 2024a,b), ChatEval (Chan et al., 2024), and PandaLM (Wang et al., 2024) reduce inference cost relative to frontier proprietary judges, but their rubrics remain coarse (helpfulness, fluency, relevance) and they typically produce a single score per output, not the per-dimension evidence-grounded rationales required for educational accountability. Educational assessment provides the complementary foundation: the eleven-dimension Washington-Accord graduate-attribute framework (International Engineering Alliance, 2021; Ho et al., 2024; ABET Engineering Accreditation Commission, 2024) and automated essay scoring under Quadratic Weighted Kappa (Cohen, 1968; Yannakoudakis et al., 2011; Ramesh and Sanampudi, 2022). The closest prior dataset, EduBench (Xu et al., 2025), scores coarse-grained correctness in student question answering rather than holistic engineering capability across a structured rubric.

2.2 Data Distillation, Quality Control, and Engineering Benchmarks

Distillation that scales a small expert seed into a large synthetic corpus is by now standard, with quality-control variants including reward-model filtering (Cui et al., 2023), teacher-ensemble voting (Nvidia et al., 2024), and dataset-level documentation standards (Gebru et al., 2021; Soldaini et al., 2024). Our three-role framework (Section 5.2) differs in two ways: we select teachers *per dimension* rather than globally, and we explicitly separate *generation* (teacher) from *supervision* (supervisor) so that the strongest single-dimension model always generates while an independent supervisor flags noisy teacher outputs. A separate strand probes the engineering ability *of LLMs themselves*—SWE-Bench (Jimenez et al., 2024), LiveCodeBench (Jain et al., 2025), HumanEval (Chen et al., 2021)—which is orthogonal to our work: we use LLMs as the *evaluators* of human engineering capability, not as the subjects under evaluation.

3 Task Definition

3.1 Framework, Rubric, and Aggregation

We adopt the eleven-dimension engineering-capability framework codified by the Washington Accord (International Engineering Alliance, 2021; Ho et al., 2024) and the China Engineering Education Accreditation graduate-attribute system, conceptually parallel to the ABET Student Outcomes (ABET Engineering Accreditation Commission, 2024): (D1) Engineering Knowledge, (D2) Problem Analysis, (D3) Design and Development of Solutions, (D4) Investigation, (D5) Modern Tool Usage, (D6) Engineering and Sustainable Development, (D7) Engineering Ethics and Professional Conduct, (D8) Individual and Teamwork, (D9) Communication, (D10) Project Management, (D11) Lifelong Learning. Each dimension d is de-

composed into three *observation points* $o_{d,k}$ that operationalise the abstract construct into measurable behavioural facets (e.g. D3 into KPI quantification, multi-alternative trade-off, and prototype verification; full decomposition in Appendix A).

Each observation point is rated on a 0–5 ordinal scale binned into five bands—*Poor* ($[0, 1)$), *Weak* ($[1, 2)$), *Adequate* ($[2, 3)$), *Good* ($[3, 4)$), *Excellent* ($[4, 5]$)—with dimension-specific descriptors that condition scores on observable evidence (e.g. D5 above 3.5 requires direct evidence that the learner, not an AI assistant, operated the tool). The dimension score is the unweighted mean of its three observation-point scores, $s_d = \frac{1}{3}\sum_{k=1}^{3} s_{d,k}$, and the per-learner capability vector is $\mathbf{s} = (s_{D1}, \ldots, s_{D11}) \in [0, 5]^{11}$; downstream weighting is left to the consumer.

3.2 Task Formalisation

A defining feature of ENGREDUBENCH is that the same rubric is applied to two structurally distinct evidence modalities—multi-turn human–AI dialogues and static project reports—unified at the evidence layer (Section 4). Given an evidence set $E = \{e_1, \ldots, e_n\}$, the evaluation task produces a structured judgement

$$\mathcal{J}(E) = (\mathbf{s},\, \mathcal{R},\, \mathcal{V},\, \mathcal{F}), \tag{1}$$

where $\mathbf{s}$ is the score vector, $\mathcal{R} = \{r_d\}_{d=1}^{11}$ are dimension-level rationales citing evidence ids, $\mathcal{V}$ is the overall qualitative evaluation, and $\mathcal{F}$ is the set of dimension-targeted improvement suggestions. All four outputs must be *traceable*—every score and recommendation grounded in cited evidence—to support auditability by human instructors. This requirement distinguishes our task from generic LLM-as-judge benchmarks (Section 2).

4 Dataset

4.1 Data Sources

Dynamic interaction subset. Multi-turn human–AI dialogues collected from a Chinese engineering-education platform on which undergraduate and graduate learners conduct project-based engineering work with AI tools, under the platform's user agreement (*Ethics Statement*); personal identifiers and project-specific names were redacted prior to processing. Each project is tagged with one of ten engineering domains (FinTech, Smart Healthcare, Intelligent Manufacturing, Digital Agriculture, Information & Communications, Transportation & Logistics, Digital Education, Smart Energy, Environmental Technology, Digital Entertainment) and one of four life-cycle stages (Requirements & Strategy, Solution Design, Engineering Development, Operation & Evaluation), serving as the dynamic-subset stratification scheme.

Static document subset. Project proposal and closure reports authored by undergraduate teams as deliverables of three national innovation programmes (Da-Chuang, Shuang-Chuang, Chu-Yan). Each report is an authentic learner artefact whose internal structure follows the team's chapter-and-section organisation (e.g. *Chapter 1 Project Background*, *Chapter 2 Product Introduction*); micro-scene segmentation therefore proceeds along the report's native subsection boundaries (Section 4.3), preserving the author's intended discourse units.

4.2 Behaviour and Process Tagging

We unify the two modalities at the evidence layer through two parallel eight-label inventories. The *dynamic behaviour labels*—Thinking, Answering, Questioning, Multi-source Consultation, Summarisation, Trade-off Analysis, Generation/Revision, Reflection—capture the cognitive role of each dialogue utterance. The *static process labels*—Problem Framing, Solution Design, Experiment/Implementation, Data Analysis, Decision Justification, Validation & Evaluation, Summary & Reflection, Planning & Management—capture the methodological function of each report section. The two inventories are not collapsed into a single set: each is purpose-fit to its modality, with unification achieved at the next layer through a shared *micro-scene* structure (Section 4.3) and the eleven-dimension rubric (Section 3.1).

4.3 Micro-scene Decomposition

Both modalities are decomposed into a common unit we call a *micro-scene*: a coherent stretch of evidence with one identifiable interaction goal. A micro-scene is characterised by a four-element schema (interaction parties, task background, expert thread nodes, turn-counting rule) and a 70–140-character natural-language summary that an organiser agent uses for downstream evidence routing. For dynamic dialogues the micro-scene boundary is detected by topic-shift heuristics over consecutive utterances; for static reports it follows the report's native subsection boundary (one micro-scene per numbered subsection). The micro-scene unit en-

sures that the eleven-dimension rubric is applied to comparable evidence granularity across the two modalities.

4.4 Golden Standard Annotation

Annotation protocol. A subset of 480 evidence sets was expert-annotated against the eleven-dimension rubric, comprising 330 dynamic human–AI interaction samples and 150 static report samples. The dynamic subset was already annotated during evaluator development; the static subset was completed through the same rubric and adjudication procedure so that the final golden set uses a single capability framework across both modalities. To quantify annotation quality we dual-annotate the full 480-sample set by two independent expert annotators following the same rubric, and report per-dimension quadratic-weighted Cohen's κ (Cohen, 1968). The eleven dimensions yield a mean $\kappa = 0.706$ with 9/11 dimensions above 0.6 (substantial agreement under the Landis and Koch (1977) scale); the lowest two dimensions are *Sustainability* (κ=0.563) and *Personal & Teamwork* (κ=0.581), reflecting the inherent subjectivity of these rubrics on borderline cases. The complete per-dimension breakdown with 95% bootstrap CIs is reported in Appendix B, and we discuss the two moderate-agreement dimensions further in the limitations (Section 7). The first annotator's scores serve as the final golden labels; the second annotation is used solely for quality verification and is not folded into the released labels.

Modality split and distillation pipeline. To preserve the golden set's 2.2:1 dynamic-to-static ratio at scale, we sample 2,800 dynamic and 1,200 static candidate samples (4,000 total) from authentic learner activity, controlling for project type and student level. Each candidate is scored by the per-dimension teacher T_d and supervisor S_d (Section 5.2); a sample is discarded only when ≥ 3 dimensions are flagged by Eq. 3. Final composition is reported in Table 1.

4.5 Statistics

Domain and stage distribution. For the dynamic subset, the 4,000 candidate samples are stratified across the 10 engineering domains $\times$ 4 life-cycle stages, with a target cell size of 70 samples (actual range 59–77). Domain-level keep rates range from 68.7% to 81.5% (mean 76.4%), indicating that the per-dimension teacher–supervisor filter is approximately calibration-balanced rather than systematically biased toward any single domain (heatmap in Appendix F).

Subset	**Dyn.**	**Sta.**	**Total**
Golden (expert-annot.)	330	150	480
Candidate pool	2,800	1,200	4,000
Distilled (kept)	2,116	1,057	3,173
Filter-rejected	684	128	812
Background-only	0	15	15
Released total	**2,446**	**1,207**	**3,653**

Table 1: EngreduBench composition by modality. *Filter-rejected*: flagged on ≥ 3 of 11 dimensions (Eq. 4); *Background-only*: static sections routed entirely to the background bucket. Post-filter keep rate 79.3%.

Benchmark	Size	Modality	Annotation	Rubric
MMLU (Hendrycks et al., 2021)	15.9k	MCQ	human	—
BBH (Suzgun et al., 2023)	6.5k	prompt	human	—
EduBench (Xu et al., 2025)	4.0k	dialogue	LLM+human	12
Ours	3.7k	dyn+stat	hum+LLM	11

Table 2: EngreduBench vs. representative LLM benchmarks. Ours is the first to combine engineering-specific Washington-Accord dimensions with dual-modality annotation and a teacher–supervisor quality-control filter (κ + T/S; details in Section 6.1).

Static chapter-function distribution. The static subset is organised by chapter-function category rather than life-cycle stage. The distribution is bottom-heavy on *Experiment/Implementation* (728 candidates, 650 retained), reflecting the writing habits of national undergraduate innovation programmes (*Da-Chuang*, *Shuang-Chuang*, *Chu-Yan*); per-category counts in Appendix F. Table 2 situates EngreduBench against representative LLM benchmarks: to our knowledge it is the first to combine engineering-specific Washington-Accord capability dimensions with dual-modality annotation and a teacher–supervisor quality-control filter.

5 Methodology

We describe the two complementary components of EngreduBench: (i) a multi-agent LLM-as-judge pipeline that produces traceable evaluations from evidence (Section 5.1), and (ii) a three-role distillation framework that converts a small set of expert-annotated samples into a large training set for downstream applications while filtering noisy teacher outputs (Section 5.2).

5.1 Multi-Agent Scoring Pipeline

A single monolithic prompt that organises evidence, scores eleven dimensions, and synthesises an overall verdict suffers from prompt overload and arithmetic drift. We decompose the task into four LLM-driven agents and a program-side aggregation step. The **Evidence Organiser Agent** assigns each item e_i a stable identifier, stage / micro-scene labels, a behaviour label drawn from the modality-specific inventory (Section 4), and routing metadata: a single *primary dimension* $\mathrm{prim}(e_i) \in \{D1, \ldots, D11\}$ plus up to three *secondary dimensions* and a routing-confidence score; items that provide context but cannot independently support any dimension score are placed in a *background-only* bucket. The **Per-Dimension Scoring Agent** runs once per dimension with the dimension rubric (Section 3.1) and all evidence whose primary or (optionally) secondary routing includes d, outputting three observation-point scores $\{s_{d,k}\} \in [0, 5]^3$, a textual rationale citing evidence ids, a self-reported confidence, and a list of missing evidence types.

Aggregation, summary, and audit. The dimension score is computed deterministically by the program as the unweighted mean of the three observation-point scores (Eq. 1), removing the LLM from the arithmetic step and eliminating numerical drift. A **Summary Agent**, conditioned on the score vector and per-dimension rationales, produces $\mathcal{V}$ and $\mathcal{F}$ (with a deterministic template fallback on JSON-validation failure). An independent **Quality Audit Agent** re-reads the final $\mathcal{J}(E)$ together with the original evidence and flags four classes of failure (score–evidence inconsistency, ungrounded claims, omission of major evidence, rubric-band miscalibration); it may use a different LLM from the main pipeline.

5.2 Three-Role Distillation Framework

To scale beyond the 480 expert-annotated golden samples we distil capability scores onto a larger pool using LLM teachers, decoupling *generation* from *supervision* so that no single teacher's biases propagate across all dimensions. We define three data-driven roles drawn from a common pool of N models that have each scored the golden set: an **evaluator** E (highest overall consistency, used for selection and as a global second opinion), a per-dimension **teacher** T_d (highest single-dimension consistency on d, used to generate scores and rationales), and a per-dimension **supervisor** S_d (paired with T_d, used solely for dirty-sample filtering). Letting $\mathrm{Cons}(m, d)$ denote a composite per-dimension consistency between model m and the human golden labels, we select $E = \arg\max_m \frac{1}{11} \sum_d \mathrm{Cons}(m, d)$, $T_d = \arg\max_m \mathrm{Cons}(m, d)$, and $S_d = E$ when $T_d \neq E$ and $m_d^{(2)}$ (second-best on d) otherwise. Crucially, T_d is *never* demoted: the strongest model on d always generates, while the supervisor provides an independent second opinion whose disagreement with T_d flags potentially erroneous outputs—contrasting with prior pipelines that use a single global supervisor or self-consistency filtering (Cui et al., 2023).

Composite consistency. For each (model, dimension) pair we compute three complementary indicators against the human gold: Quadratic Weighted Kappa (QWK) on integer-rounded scores, Pearson r on raw scores, and Sentence-BERT cosine similarity on rationales (BGE-LARGE-ZH). The composite is a convex combination $\mathrm{Cons}(m, d) = w_1\,\mathrm{QWK} + w_2\, r + w_3 \cos$ with $\sum_i w_i = 1$ (Eq. 2); we report results under three weighting schemes (W-Equal, W-Score-Heavy, W-Text-Heavy) and treat ranking stability across the three as a robustness criterion.

$$\mathrm{Cons}(m, d) = w_1\,\mathrm{QWK} + w_2\, r + w_3 \cos. \quad (2)$$

Summary teacher and supervisor. The same methodology extends to the two free-form fields ($\mathcal{V}$, $\mathcal{F}$): each candidate is scored by BGE-large-zh cosine similarity to the human-gold reference on the 480 samples, averaged across the two fields, and the top two are designated T_Σ and S_Σ (Section 6.1). This selection runs offline and incurs no additional API cost.

5.3 Threshold Calibration and Filtering Rule

For each dimension d we calibrate τ_d and τ_{text} on the golden set, sweeping candidates to maximise post-filter T_d–human QWK under a minimum-keep-rate constraint. The two-stage filter flags a dimension when $|T_d(x) - S_d(x)| > \tau_d$ *or* $\cos(T_d.r,\, S_d.r) < \tau_{\text{text}}$ (Eq. 3), and rejects a sample only when at least three dimensions are flagged (Eq. 4)—preserving dimension-specific filtering while preventing excessive sample loss from iso-

Model	QWK	Pear.	Cos.	Comp.
Gemini 3.1 Pro (E)	0.822	0.884	0.846	**0.850**
Claude Opus 4.7	0.777	0.868	0.840	0.829
Qwen 3.6 Plus	0.760	0.845	0.837	0.814
GPT-5.4	0.712	0.840	0.827	0.793
Doubao Seed 2.0 Pro	0.732	0.825	0.809	0.789
Grok 4.2 Thinking	0.678	0.834	0.817	0.776
DeepSeek-V4-Pro	0.633	0.768	0.823	0.742
MiniMax M2.7	0.606	0.756	0.825	0.729

Table 3: Evaluator candidates ranked by composite consistency (Comp. = mean of QWK, Pearson r, CosSim) on the 480-sample golden set under W-Equal. The top-ranked model is selected as the global evaluator E.

Dim.	T_d	S_d	τ_d	τ_{text}
D1 EngrKnow	Claude	Gemini	0.70	0.84
D2 ProbAnal	Gemini	Claude	0.50	0.82
D3 DesgnSol	Gemini	Claude	1.00	0.82
D4 Research	Gemini	Claude	0.50	0.76
D5 ModTools	Gemini	Claude	0.70	0.84
D6 Sustain	Gemini	GPT-5.4	0.80	0.80
D7 Ethics	Gemini	Claude	0.60	0.84
D8 Team	Gemini	Claude	0.70	0.50
D9 Comm	Gemini	Qwen	0.60	0.68
D10 ProjMgmt	Gemini	Qwen	0.70	0.78
D11 LifeLearn	Gemini	Claude	0.80	0.86

Table 4: Per-dimension teacher T_d (strongest single-dim model) and supervisor S_d ($= E$ when $T_d \neq E$, else second-best on d). Models abbreviated: Gemini = Gemini 3.1 Pro, Claude = Claude Opus 4.7, Qwen = Qwen 3.6 Plus. Per-dim T_d/S_d QWK values in Figure 4.

lated cross-model disagreement.

$$\text{flag}_d(x) \iff |T_d(x) - S_d(x)| > \tau_d \vee \cos(T_d.r,\, S_d.r) < \tau_{\text{text}}, \quad (3)$$

$$\text{reject}(x) \iff \sum_{d=1}^{11} \mathbb{I}[\text{flag}_d(x)] \geq 3. \quad (4)$$

6 Experiments

We report three experiments. Section 6.1 selects the evaluator E, per-dimension teachers T_d, and summary teachers T_Σ/S_Σ; Section 6.2 calibrates the rejection thresholds; Section 6.3 reports the quality of the distilled expansion. The candidate pool is eight frontier LLMs released by 2026: CLAUDE OPUS 4.7, DEEPSEEK-V4-PRO, GPT-5.4, GEMINI 3.1 PRO, GROK 4.2 THINKING, QWEN 3.6 PLUS, MINIMAX M2.7, and DOUBAO SEED 2.0 PRO; per-(model, dimension) consistency is the convex combination of QWK on integer-rounded scores, Pearson r on raw scores, and BGE-LARGE-ZH cosine similarity on rationales (Eq. 2), reported under W-Equal, W-Score-Heavy, and W-Text-Heavy weightings.

6.1 Evaluator and Teacher Selection

Selection-stability heatmap. Per-dimension QWK against the human gold (Appendix G, Figure 4) reveals two patterns: (i) Gemini 3.1 Pro leads on 10 of the 11 dimensions, and (ii) Claude Opus 4.7 is strongest on D1 (Engineering Knowledge). Bootstrap resampling (1,000 replicates) confirms strong stability of this assignment.

Robustness and significance. The top-2 ranking is preserved across all three weighting schemes: Gemini 3.1 Pro remains first (Composite 0.850/0.845/0.846 for W-Equal/W-Score-Heavy/W-Text-Heavy) and Claude Opus 4.7 second (0.829/0.817/0.827). Under W-Equal, Gemini significantly outperforms Claude in paired MAE against human gold (Wilcoxon signed-rank, $p = 1.21 \times 10^{-4}$); bootstrap selection of E is perfectly stable (100% of 1,000 resamples).

Summary teacher selection. Applying the same role-selection methodology to the two free-form fields (overall `evaluation` and `suggestions`), we score each candidate by BGE-large-zh cosine similarity to the human-gold references, averaged across the two fields. The aggregate range across the eight candidates is narrow (0.778–0.825); the two complementary leaders—gemini-3.1-pro-preview (0.825 aggregate; strongest on `suggestions`, 0.802) and qwen3.6-plus (0.823 aggregate; strongest on `evaluation`, 0.858)—are designated T_Σ and S_Σ (full ranking in Appendix H).

6.2 Threshold Calibration

We calibrate dimension-specific thresholds under a 75% minimum-keep-rate constraint; the resulting τ_d (range 0.50–1.00) and τ_{text} (range 0.50–0.86) are reported in Table 4. Across the eleven dimensions, the average dimension-level rejection rate is 25.6% with an average teacher–human QWK improvement of +0.036 on the retained subset; the largest gains accrue to Individual & Teamwork (+0.071), Project Management (+0.066), and Modern Tool Usage (+0.045), while Design & Development receives much milder filtering (4.1% rejection; +0.008 QWK). The calibrated thresholds are applied only as dimension-level anomaly detectors, with sample-level rejection deferred to the ≥ 3-dimension aggregate veto (Eq. 4) to avoid the excessive sample loss of single-dimension poli-

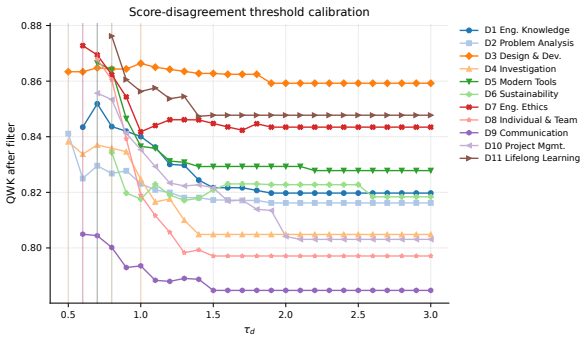

Figure 1: Threshold calibration curves for τ_d under the 75% minimum-keep-rate constraint. Each line is one dimension; analogous τ_{text} curves in Appendix I.

cies.

6.3 Distillation Quality

The calibrated sample-level filter rejects 20.7% of candidates (812 of 3,985 scored samples; an additional 15 background-only samples are programmatically dropped before scoring), yielding 3,173 retained instances (79.3% keep rate). Per-dimension teacher-score distributions cluster around mean 3.2–3.8 for eight of eleven dimensions, with *Project Management* the lowest (2.78) and *Engineering Ethics* the highest (3.77); the narrow standard deviation on *Sustainability & Society* (std 0.58) corroborates the moderate IAA on that dimension. Teacher–supervisor gaps concentrate at $|T_d - S_d| \leq 0.5$ across all dimensions, with heavy upper tails capturing precisely the samples flagged by Eq. 3. Full distributions and three illustrative borderline cases (T/S disagreement, aggregate-veto boundary, background-only drop) are reported in Appendices J and K.

7 Discussion

7.1 Implications

Our results suggest that human–AI interaction transcripts are a viable, and in some respects superior, source of evidence for engineering-capability assessment: compared with static project reports—the dominant artefact in current accreditation practice—transcripts expose the learner's reasoning trajectory, including discarded alternatives and revisions, and the eleven-dimension framework transfers cleanly to this dynamic modality once the evidence layer is unified. The three-role distillation framework is, in turn, a template for educational data curation more broadly: by treating teacher selection as an *empirical* per-dimension question and delegating noise floor to an independent supervisor, the design generalises to other multi-criterion assessment tasks (medical case-based learning, legal-reasoning training) where no single LLM dominates across all rubric dimensions. Practically, the open release of ENGREDUBENCH—3,653 evidence sets paired with eleven-dimension rationales, per-dimension rejection thresholds, and the full multi-agent prompt suite—lowers the marginal cost of formative engineering assessment and provides a single audited substrate on which downstream applications (instructor-facing decision support, fine-tuning of smaller open judges, longitudinal cohort analysis) can be built.

7.2 Limitations

Single platform, single language. The dynamic interaction subset is sourced from a single Chinese engineering-education platform, and all rubric descriptors are in Chinese. Population coverage, demographic representativeness, and cross-lingual transfer are therefore untested; cross-platform and cross-lingual replication remain future work.

Synthetic component. Approximately 87% of the released samples (3,173 of 3,653) are produced by the LLM-based distillation pipeline seeded from the 480 expert-annotated cases. Although we apply per-dimension threshold filtering and report post-filter distributions (Section 6.3), residual distributional drift cannot be fully ruled out and downstream users should re-validate the filter on their own targets.

Static chapter-function skew. The static-report subset is heavily skewed toward *Experiment/Implementation* sections (∼60% of static candidates), with sparse coverage of *Summary & Reflection* and *Decision Justification* (Appendix F). This skew is a property of the source corpora—national undergraduate innovation programmes whose reports concentrate on hands-on experimentation. The dataset is therefore best suited for engineering-practice capabilities; higher-order reflection is better served by the dynamic subset.

Moderate IAA on subjective and context-dependent dimensions. Two dimensions—*Sustainability & Engineering for Society* (κ=0.563) and *Personal & Teamwork* (κ=0.581)—attain moderate rather than substantial agreement on the dual-annotated golden set (Appendix B).

Sustainability claims are often implicit in engineering trade-offs, and teamwork signals are inferred from second-order cues. Engineering Ethics (D7) raises a related concern: our rubric caps scores when ethics terminology appears without substantive analysis, which may under-reward implicit ethical reasoning. We retain all dimensions in the released framework but flag them as candidates for rubric refinement and third-annotator adjudication. The eight-model evaluator pool is itself a 2026 snapshot and will become outdated; the selection methodology is model-agnostic and can be re-run as the pool evolves.

7.3 Future Work

Three directions warrant immediate follow-up: extending platform coverage to English-language engineering-education environments to test cross-lingual generality; integrating the multi-agent evaluator into an instructor-facing tool that surfaces evidence-grounded rationales and supports rapid acceptance/override, directly testing the workload-reduction hypothesis; and extending the three-role framework to a dynamic assignment in which T_d and S_d are re-selected periodically as new candidate models are released.

8 Conclusion

We presented ENGREDUBENCH, a dual-modality Chinese-language benchmark and methodology for engineering-capability assessment grounded in the eleven-dimension Washington Accord framework. Our contributions are a unified evidence-layer schema spanning dynamic transcripts and static reports, a multi-agent evaluator pipeline with traceable rationales, a three-role distillation framework that decouples per-dimension teaching from independent supervision, and a 3,653-sample resource with $\kappa = 0.706$ inter-annotator agreement on 480 expert-annotated anchors and a 79.3% post-filter sample-level keep rate. By open-sourcing the dataset, prompts, and evaluator code we lower the marginal cost of formative engineering-capability assessment and support the transition of educational evaluation infrastructure from static-artefact scoring to AI-era human–AI interaction analysis.

Ethics Statement

Data sources and consent. The dynamic interaction subset is drawn from a Chinese engineering-education platform whose user agreement permits research use of de-identified transcripts. The platform has accumulated more than 70 training scripts, 7 group-collaboration templates, nearly 30 capability-specific agents, and 15 rubric-aligned evaluation modules covering engineering capability, complex-problem solving, user insight, programming, and group collaboration; since July of the preceding year it has logged $79{,}621$ sign-ins and $6{,}750$ assignment submissions across more than 40 courses and teaching scenarios (introductions to information & communications, information theory & coding, engineering leadership, design thinking, electronic-circuit innovation, no-code engineering development, Python programming, intelligent communication, and in-service teacher training, among others), and has reached more than 100 partner institutions spanning research universities, vocational colleges, K–12 schools, science museums, education-management bodies, industrial research labs, and engineering-education accreditation associations. The static report subset comprises undergraduate-team project deliverables submitted to three national innovation programmes (Da-Chuang, Shuang-Chuang, Chu-Yan), under the respective programmes' submission terms. No samples were collected outside these existing agreements, and no recruitment of human subjects was conducted for this work.

Privacy and anonymisation. All evidence text was processed through a PII-redaction pass before annotation: personal identifiers (real names, student IDs, contact information), institution names, and project-specific names were replaced with neutral placeholders (e.g. *Student 1*). Source-of-evidence pointers refer to filename-level identifiers only; raw platform logs and original reports are not redistributed. The released dataset is provided under an anonymous, internal-use license during the review period; broader release will be re-evaluated post-acceptance in consultation with the providing institutions.

Annotator qualifications and compensation. Both expert annotators hold graduate degrees in engineering education and have prior rubric-based assessment experience. Annotation was performed under a structured protocol with periodic adjudication meetings to maintain consistency. Annotators were compensated above local industry rates for technical reviewing work, and time spent on rubric calibration was paid on the same basis as scoring time.

Dual-use considerations. An automated capability evaluator could in principle be misused for high-stakes summative assessment of individual students, a use the rubric was not designed for and which our IAA evidence (moderate agreement on subjective dimensions; Appendix B) does not justify. We therefore explicitly state that ENGREDUBENCH is intended for *formative* feedback, evaluator benchmarking, and methodology research, not for ranking or admissions decisions. We further note that the per-dimension traceable rationales make any deployed system *auditable* by human instructors—a property we view as a prerequisite for ethical use.

Limitations and mitigation. Section 7.2 documents the limitations of the released resource (single-platform, single-language, ∼87% LLM-distilled, two moderate-IAA dimensions, static-subset chapter-function skew). We open-source the rejection thresholds, supervisor outputs, and dropped samples so that downstream users can revalidate the filter on their own targets and audit borderline rejections. We do not claim that the dataset substitutes for human judgement in any high-stakes application.

References

ABET Engineering Accreditation Commission. 2024. Criteria for accrediting engineering programs, 2024–2025. ABET, Baltimore, MD, USA. https://www.abet.org/accreditation/accreditation-criteria/criteria-for-accrediting-engineering-programs-2024-2025/.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2024. Chateval: Towards better llm-based evaluators through multi-agent debate. In *International conference on learning representations*, volume 2024, pages 9079–9093.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, and 1 others. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

Jacob Cohen. 1968. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70(4):213–220.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Bingxiang He, Wei Zhu, Yuan Ni, Guotong Xie, Ruobing Xie, Yankai Lin, and 1 others. 2023. Ultrafeedback: Boosting language models with scaled ai feedback. *arXiv preprint arXiv:2310.01377*.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92.

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

June Ho, Viken Kortian, Nazmul Huda, and Ann Lee. 2024. Engineering management education: Washington accord accreditation programs. *Engineering Management Journal*, 36(4):353–365.

International Engineering Alliance. 2021. Graduate attributes and professional competencies. Version 4, IEA Washington Accord. https://www.ieagreements.org/.

Naman Jain, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. 2025. Livecodebench: Holistic and contamination free evaluation of large language models for code. In *International Conference on Learning Representations*, volume 2025, pages 58791–58831.

Carlos E Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. 2024. Swe-bench: Can language models resolve real-world github issues? In *International Conference on Learning Representations*, volume 2024, pages 54107–54157.

Seungone Kim, Jay Shin, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Ryan Shin, Sungdong Kim, James Thorne, Minjoon Seo, and 1 others. 2024a. Prometheus: Inducing fine-grained evaluation capability in language models. In *International Conference on Learning Representations*, volume 2024, pages 29927–29962.

Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. 2024b. Prometheus 2: An open source language model specialized in evaluating other language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4334–4353.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 2511–2522.

Bo Adler Nvidia, Niket Agarwal, Ashwath Aithal, Dong H Anh, Pallab Bhattacharya, Annika Brundyn, Jared Casper, Bryan Catanzaro, Sharon Clay, Jonathan Cohen, and 1 others. 2024. Nemotron-4 340b technical report. *arXiv preprint arXiv:2406.11704*, 6(8).

Dadi Ramesh and Suresh Kumar Sanampudi. 2022. An automated essay scoring systems: a systematic literature review. *Artificial intelligence review*, 55(3):2495–2527.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, and 1 others. 2024. Dolma: An open corpus of three trillion tokens for language model pretraining research. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15725–15788.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V Le, Ed H Chi, Denny Zhou, and 1 others. 2023. Challenging big-bench tasks and whether chain-of-thought can solve them. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13003–13051.

Yidong Wang, Zhuohao Yu, Wenjin Yao, Zhengran Zeng, Linyi Yang, Cunxiang Wang, Hao Chen, Chaoya Jiang, Rui Xie, Jindong Wang, and 1 others. 2024. Pandalm: An automatic evaluation benchmark for llm instruction tuning optimization. In *International Conference on Learning Representations*, volume 2024, pages 43573–43593.

Bin Xu, Yu Bai, Huashan Sun, Yiguan Lin, Siming Liu, Xinyue Liang, Yaolin Li, Zhuangzhi Dong, Jingren Zhang, Yufan Deng, and 1 others. 2025. Edubench: A comprehensive benchmarking dataset for evaluating large language models in diverse educational scenarios. *arXiv preprint arXiv:2505.16160*.

Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. 2011. A new dataset and method for automatically grading esol texts. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pages 180–189.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

Lianghui Zhu, Xinggang Wang, and Xinlong Wang. 2025. Judgelm: Fine-tuned large language models are scalable judges. In *International Conference on Learning Representations*, volume 2025, pages 51257–51296.

A Eleven-Dimension Rubric (Detailed)

Each of the eleven dimensions is operationalised by three *observation points* that anchor the rubric to concrete, verifiable behavioural facets. Table 5 lists the eleven dimensions and their three observation points. The same 0–5 ordinal scale, banded as in Table 6, applies to every observation point; the dimension score is the unweighted mean of its three observation-point scores.

Dimension	Three observation points
D1 EngrKnow	(i) cross-disciplinary integration; (ii) correctness & contextual fit; (iii) awareness of limits & uncertainty
D2 ProbAnal	(i) problem framing & decomposition; (ii) literature search & evidence synthesis; (iii) multi-factor & sustainability considerations
D3 DesgnSol	(i) requirement & KPI quantification; (ii) multi-alternative trade-off; (iii) prototype & verification closure
D4 Research	(i) testable hypotheses; (ii) experimental design; (iii) data analysis & conclusion validity
D5 ModTools	(i) tool selection & fit; (ii) modelling/coding & verification; (iii) reproducibility & provenance
D6 Sustain	(i) life-cycle & carbon footprint; (ii) HSE & legal compliance; (iii) social/environmental impact & stakeholders
D7 Ethics	(i) ethical risk identification; (ii) compliance & professional conduct; (iii) public-interest accountability
D8 Team	(i) role positioning & interdisciplinary cooperation; (ii) team-process management; (iii) deliverable quality & individual contribution
D9 Comm	(i) technical documentation; (ii) oral presentation & defence; (iii) cross-cultural / multilingual adaptation
D10 ProjMgmt	(i) scope/schedule/resource planning; (ii) risk management; (iii) budget & economic decisions
D11 LifeLearn	(i) goal-setting & gap identification; (ii) self-regulated learning & reflection; (iii) knowledge transfer & creative application

Table 5: Three observation points per dimension. Each observation point is rated 0–5 on the same band scale (Table 6). Dimension labels match Table 4.

Modality-aware caps. Three caps prevent the rubric from over-rewarding artefacts that do not entail the underlying capability: (1) for D1/D3/D5, evidence that consists only of planning, requirements or AI-generated artefacts without learner-driven verification is capped at 3.5; (2) for any dimension, evidence drawn from static-report *claims* (rather than verifiable process or result records) cannot

Range	Band	Generic descriptor
$[0, 1)$	Poor	Major errors or severe omissions; unable to support the criterion.
$[1, 2)$	Weak	Sporadic surface behaviour; method clearly insufficient.
$[2, 3)$	Adequate	Basic competence with notable gaps or insufficient rigour.
$[3, 4)$	Good	Substantive evidence with minor weaknesses; criterion broadly met.
$[4, 5]$	Excellent	Systematic, verifiable evidence with quantitative anchors and explicit boundary awareness.

Table 6: Five-level rubric bands. Dimension-specific descriptors (e.g. D5 requires direct evidence that the learner—not the AI assistant—operated the tool to exceed 3.5) refine these generic anchors; full descriptors are released with the dataset.

exceed 4.0 on implementation-related observation points; (3) when core implementation or testing is predominantly produced by an AI tool with negligible learner involvement, the affected observation points are capped at 3.0.

B Inter-Annotator Agreement (Per-Dimension)

Two expert annotators independently scored all 480 golden samples following the same eleven-dimension rubric. Because each sample is only scored on dimensions for which the underlying evidence supports a judgement, the per-dimension paired counts n vary across rows. Table 7 reports quadratic-weighted Cohen's κ (Cohen, 1968) together with 95% bootstrap confidence intervals (1,000 resamples per dimension) and exact / within-± 1 agreement percentages. Continuous scores in $[0, 5]$ are first rounded to integer categories $\{0, \ldots, 5\}$ before computing κ, following standard practice for ordinal rubrics.

The eleven dimensions show a mean $\kappa = 0.706$ (substantial agreement on the Landis and Koch (1977) scale, 0.61–0.80), with 9/11 dimensions above 0.6 and within-± 1 agreement averaging 96%. The two moderate-agreement dimensions—*Sustainability & Engineering for Society* (κ=0.563) and *Personal & Teamwork* (κ=0.581)—are discussed in the limitations section. The first annotator's scores are released as the final golden labels; the second annotation is used exclusively for quality verification.

Dim.	n	κ_q [CI]	Exact	± 1
D1 EngrKnow	299	.685 [.62,.74]	49.5%	89.6%
D2 ProbAnal	385	.725 [.66,.78]	62.3%	94.8%
D3 DesgnSol	414	.812 [.78,.84]	60.6%	98.8%
D4 Research	190	.768 [.71,.82]	57.4%	97.4%
D5 ModTools	215	.792 [.74,.84]	58.6%	97.7%
D6 Sustain	262	.563 [.48,.63]	47.7%	94.3%
D7 Ethics	213	.661 [.58,.73]	45.5%	85.9%
D8 Team	351	.581 [.52,.64]	43.3%	93.2%
D9 Comm	343	.665 [.61,.72]	45.8%	94.5%
D10 ProjMgmt	282	.726 [.67,.78]	55.3%	93.6%
D11 LifeLearn	198	.787 [.73,.84]	60.1%	96.0%
Mean	—	**.706**	53.4%	94.2%

Table 7: Per-dimension inter-annotator agreement on the 480-sample golden set. κ_q is the quadratic-weighted Cohen's κ with 95% bootstrap CIs (1,000 resamples). **Exact**: integer-bucketed exact match; ± 1: within one rubric band. n varies because annotators only score dimensions with supporting evidence. Dimension labels as in Table 4.

C Prompts

The four-agent pipeline is driven by four prompt templates (Table 8), translated and condensed below. The full Chinese-language prompts—including all routing-confidence calibration rules, dimension-specific anchor signals, and JSON schema—are released with the dataset.

Agent	Input	Output
Organiser	evidence E	ids, routing, conf.
Per-dim	rubric + evid.	3 obs. scores
Summary	scores + reasons	$\mathcal{V}$, $\mathcal{F}$
Audit	$\mathcal{J}(E)$ + evid.	failure flags

Table 8: The four agents in the multi-agent scoring pipeline (Section 5.1). Conf. = self-reported confidence.

Organiser Agent (condensed).

Role. Assign each evidence item a stable id, a stage label, a micro-scene label, a behaviour label, and routing metadata.
Inputs. The 11 dimensions and 8 modality-specific behaviour labels; raw evidence E.
Output (per item). `evidence_id`, `stage`, `micro_scene`, `behavior_label`, `primary_dimension`, ≤ 3 `secondary_dimensions`, `routing_confidence` $\in [0, 1]$. Items providing only context go to `background_only_evidence`.
Constraints. Single primary dimension; "engineering knowledge" kept separable from "problem analysis" and "design" to avoid absorption.

Per-Dimension Scoring Agent (condensed).

Role. Given dimension d, the dimension rubric (Appendix A), and all evidence routed to d,

output three observation-point scores.
Output. For each observation point $k \in \{1, 2, 3\}$: `score` $\in [0, 5]$, evidence-citing `reason`, `confidence`, `used_evidence_ids`, `missing_evidence_types`.
Anchors. Reach $\geq$ 4.0 when at least one of {systematic trade-off, boundary/limit awareness, weak quantification, in-scope verifiable progress, learner-led revision} is grounded in cited evidence; reach $\geq$ 4.5 when verification, quantitative, implementation, or traceability signals appear.
Caps. Pure-planning $\leq$ 4.0; AI-dominant artefacts without learner verification $\leq$ 3.0 on the affected observation points.

Summary Agent (condensed).

Role. Conditioned on the eleven-dimension score vector $\mathbf{s}$ and per-dimension rationales, write the overall qualitative evaluation $\mathcal{V}$ and dimension-targeted improvement suggestions $\mathcal{F}$.
Constraints. Cite dimension ids; do not introduce new evidence; on JSON-validation failure, the program substitutes a deterministic template covering the lowest-scoring dimensions.

Quality Audit Agent (condensed).

Role. Re-read the final $\mathcal{J}(E)$ together with the original evidence and flag four failure classes: (i) score–evidence inconsistency; (ii) ungrounded rationale claims; (iii) omission of major evidence; (iv) rubric-band miscalibration.
Output. Per-failure-class boolean flags with short justifications and references to evidence ids. May use a different LLM from the main pipeline.

D Computational Resources

The distillation pipeline processes 4,000 candidate samples through three LLM stages: organiser routing (1 call per sample), per-dimension teacher scoring (11 calls), and summary teacher generation on retained samples (1 call). Including occasional retries due to JSON-validation failures and API rate-limit backoff, total inference volume is approximately 5.7×10^4 LLM calls, $\sim$ 380M input tokens, and $\sim$ 50M output tokens. The total inference cost—running all teachers, supervisors, organiser, and summary teacher on the candidate pool—was approximately ¥9,400. The local components of the pipeline (BGE-LARGE-ZH embedding for rationale cosine, threshold calibration, distribution analysis, and inter-annotator-agreement computation) run on a single CPU workstation in under one hour combined and do not require GPU resources.

E Datasheets for Datasets

We follow the structure of Gebru et al. (2021).

Motivation. ENGREDUBENCH addresses the gap between the eleven-dimension Washington-Accord engineering-capability framework—used worldwide for undergraduate accreditation—and the absence of structured LLM-evaluatable benchmarks at this granularity. The dataset was created by the authors as the empirical substrate for an LLM-as-judge methodology aimed at lowering the human labour required for formative engineering-capability assessment.

Composition. 3,653 evidence sets paired with eleven-dimension capability scores, per-dimension rationales, and overall evaluation/suggestions texts. 480 samples are expert-annotated; 3,173 are distilled from an eight-frontier-model pool and filtered via teacher–supervisor disagreement (Section 5.2). All content is in Chinese.

Collection process. The dynamic subset is drawn from a Chinese engineering-education platform (under user agreement; personal identifiers and project-specific names redacted prior to processing). The static subset comprises undergraduate-team project reports submitted to national innovation programmes (Da-Chuang, Shuang-Chuang, Chu-Yan), under the respective programmes' submission terms.

Preprocessing. (1) Micro-scene segmentation along native discourse boundaries (Section 4.3); (2) eight-label behavioural / process tagging at the evidence level (Section 4.2); (3) PII redaction on all evidence text; (4) organiser-level routing to the eleven-dimension framework; (5) per-dimension scoring + supervisor filtering (Section 5.2).

Uses. The released dataset is intended for: (i) benchmarking LLM-judge consistency against an expert anchor; (ii) calibrating dimension-level filters for distillation pipelines; (iii) reproducibility of the three-role distillation methodology. It is *not* intended for high-stakes summative assessment of individual students.

Distribution. Access is provided under an anonymous, internal-use license during the review period. Personal identifiers, institution names, and project-specific names have been redacted. Source-of-evidence pointers refer to filename-level identifiers only.

Maintenance. The authors commit to maintaining the dataset and responding to reproduction

queries via the contact channel disclosed at the camera-ready stage. We expect to revise the rubric and rerun the distillation pipeline when frontier model snapshots materially shift.

F Dataset Distribution Figures

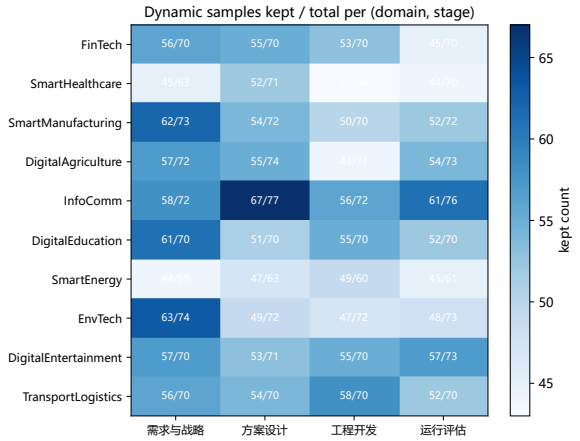

Figure 2: Dynamic-subset distribution: post-filter kept count over total candidate per (domain × life-cycle stage) cell. Labels are `kept/total`.

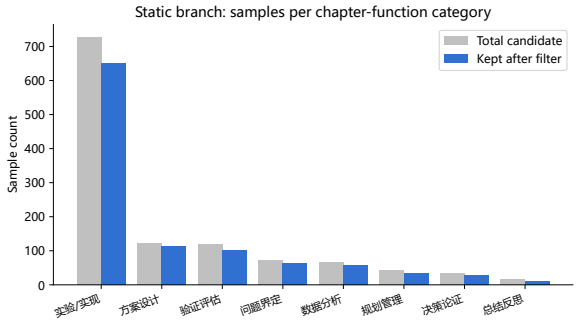

Figure 3: Static-subset candidate vs. retained sample counts per chapter-function category.

G Bootstrap Selection Stability and Overall Ranking

H Summary Teacher Ranking

I Threshold Calibration (τ_{text})

J Score and Disagreement Distributions

K Case Studies

We provide three qualitative cases illustrating how the three-role framework handles borderline samples.

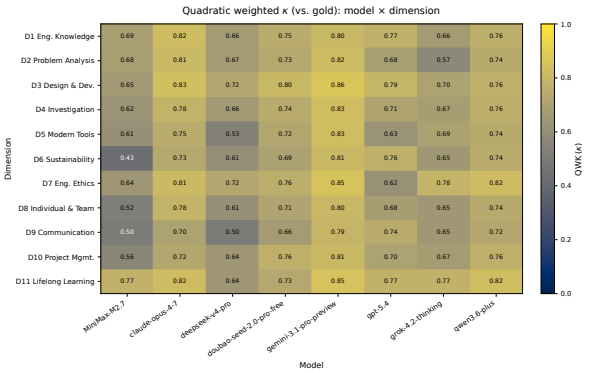

Figure 4: Per-dimension QWK of all eight evaluator candidates against the human golden labels. Rows follow the canonical D1–D11 dimension order.

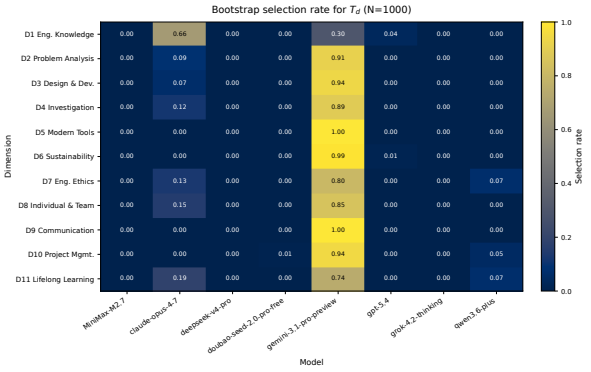

Figure 5: Bootstrap teacher-selection stability over 1,000 resamples. Each cell shows the probability that a model is selected as the dimension-level teacher T_d. Gemini 3.1 Pro dominates 10/11 dimensions; Claude Opus 4.7 dominates D1.

Case 1: Disagreement-driven rejection. On a sample from *Smart Manufacturing* (Operation & Evaluation stage), the teacher awarded 3.8 on *Problem Analysis* based on a chain of trade-off justifications, while the supervisor returned 1.9, noting that the trade-offs were not anchored to a quantitative target. The score gap of 1.9 exceeded τ_{ProbAnal}=0.50, and the rationale cosine of 0.41 fell below τ_{text}=0.82. The sample accumulated 4 dimension-level flags in total and was moved to `dropped`.

Case 2: Boundary kept by aggregate veto. On a *Digital Education* sample, only *Lifelong Learning* ($|T-S|$=1.2) and *Communication* ($|T-S|$=0.9) triggered the dimension-level flag. With $2 < 3$ flagged dimensions the sample survived under Eq. 4, validating the choice of a ≥ 3 aggregate veto over a single-dimension rule.

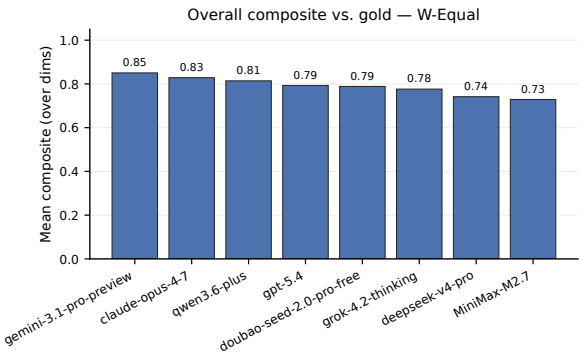

Figure 6: Overall evaluator consistency under the W-Equal scheme. Gemini 3.1 Pro ranks first, followed by Claude Opus 4.7 and Qwen 3.6 Plus.

Model	Eval.	Sugg.	Aggr.
gemini-3.1-pro-preview	0.849	**0.802**	**0.825**
qwen3.6-plus	**0.858**	0.789	0.823
claude-opus-4-7	0.842	0.788	0.815
MiniMax-M2.7	0.822	0.771	0.797
gpt-5.4	0.795	0.798	0.796
grok-4.2-thinking	0.826	0.752	0.789
doubao-seed-2.0-pro-free	0.783	0.778	0.781
deepseek-v4-pro	0.801	0.754	0.778

Table 9: Summary teacher / supervisor selection. BGE-large-zh cosine similarity to the human-gold `evaluation` (Eval.) and `suggestions` (Sugg.) fields, averaged across the 480 golden samples; Aggr. is the mean of the two field cosines. Top two are T_Σ (summary teacher) and S_Σ (summary supervisor).

Case 3: Background-only programmatic drop. A *static* sample drawn from a project's *Team Members* section contained only authorship metadata; the organiser routed all of its evidence to the `background_only` bucket. With no dimension receiving a scorable evidence set, the sample was moved to `dropped` with `reason=all_background` prior to summary generation. Across all 4,000 candidates, 15 samples fell into this category, all originating from administrative sections of static reports.

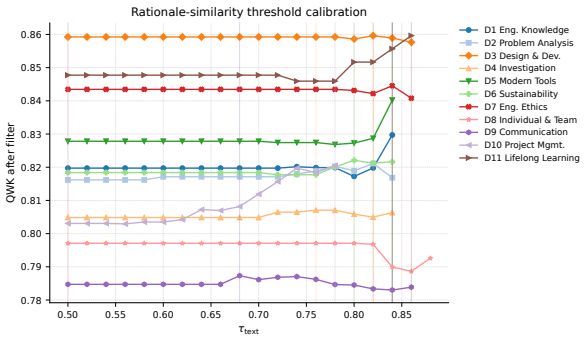

Figure 7: Rationale-similarity threshold τ_{text} vs. post-filter teacher–human QWK, under the 75% minimum-keep-rate constraint. Each line corresponds to one dimension. Companion to Figure 1 in the main text.

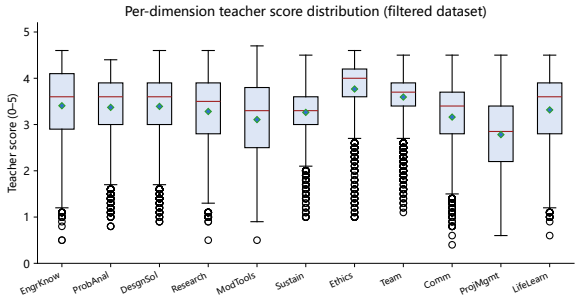

Figure 8: Per-dimension teacher-score distribution on the 3,173 retained samples. Box: IQR; whisker: 1.5×IQR; diamond: mean.

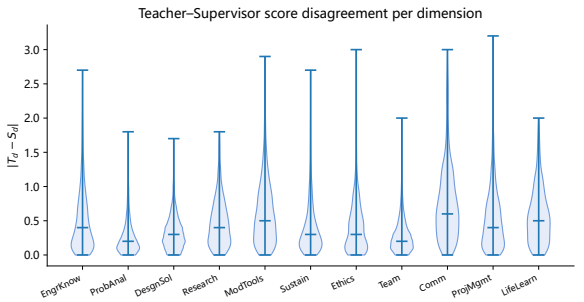

Figure 9: Teacher–supervisor score-gap distribution per dimension. Violin width is the kernel density estimate; the white line marks the median.