C^4-EDS: Planning Before Prompting for Educational Dialogue Synthesis

Anonymous authors
Paper under double-blind review

Abstract

Synthetic-data systems can now specify how individual examples are generated in considerable detail, while the data space formed by those examples may remain implicit. In educational dialogue synthesis, rich source, task, role, cognitive, and pedagogical conditions do not by themselves expose which semantic combinations are legal or how dataset composition should be planned. We formulate this gap as a generation-context representation problem and introduce C^4-EDS. Before a complete context is assembled, C^4-EDS decouples Content and Task into independently addressable semantic objects, makes their legal relations explicit, and thereby forms an enumerable and allocatable planning space. Selected relations are compiled into executable Jobs and integer candidate workloads. Representation, execution, and quality governance retain distinct responsibilities, while planning identities persist through synthesis and filtering so that retained data can be audited against the original plan. In K–12 Classical Chinese tutoring, 5,628 theoretical Content–Task combinations yield 2,807 legal Pairs and 5,614 executable Jobs. Under a matched workload, explicit reasoning improves structural realization without monotonically improving pedagogical quality; full Job-level planning reaches an automatic hard-pass rate of 88.28% and the highest mean continuous human rating. Downstream training further exhibits a target-dependent Process–Result trade-off. High-quality synthetic data therefore depends on how the generation space is represented, planned, and governed, as well as on how each example is generated.

1 Introduction

Synthetic-data generation is evolving from isolated model calls into a programmable process of data construction. Educational dialogue systems combine source evidence, learner state, role relations, diagnosis, questioning, scaffolding, and feedback (Graesser et al., 2005; Macina et al., 2023; Wang et al., 2024a; Kwon et al., 2024). General synthetic-data methods use instruction evolution, source corpora, task conditions, and multistage workflows to specify how individual examples are produced (Wang et al., 2023; Nayak et al., 2024; Ziegler et al., 2025; Patel et al., 2024). This example-level programmability can coexist with an implicit view of the dataset those examples are intended to form.

The distinction follows from the position of the generation context in data construction. Source material, task requirements, roles, cognitive requirements, and interaction constraints must eventually be assembled into an input the Generator can execute. That context is sufficient for one generation, while its semantic constituents, their admissible relations, and their organization across the dataset may remain embedded in assembled natural language. Individual examples can then be specified precisely even when feasible support, target composition, and post-filter composition lack a shared explicit representation.

We view this gap as a problem of generation-context representation. A generation context is commonly treated as the starting input to synthesis; from the perspective of data design, it is also the product of upstream decisions about which semantic constituents to select, which relations to permit, how to instantiate them, and how to allocate a finite generation budget.

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.

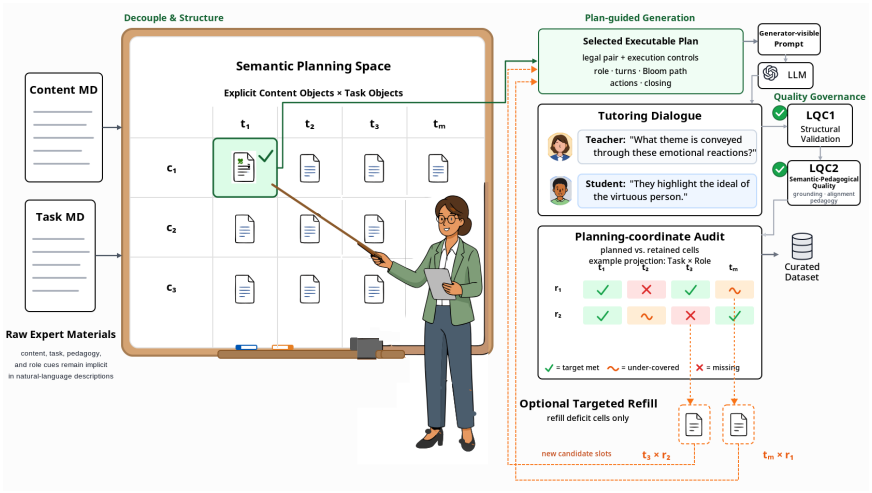

Figure 1: Representation, compilation, and governance in C^4-EDS. Content and Task objects define a legal planning space; selected relations compile into executable generation plans; quality outcomes map back to the same coordinates. The coordinates can also identify deficit cells for optional directed supplementation.

This leads to the central question: *what should be represented before a generation context is assembled?* Explicit semantic objects and relations make context assembly an addressable stage of dataset design.

C^4-EDS realizes this view by representing the semantic objects that constitute a context and the legal relations among them, compiling selected relations into executable plans and candidate workloads, and preserving their planning identities through generation and quality governance. Dataset composition can consequently be planned before generation and audited after filtering along the same coordinates. Figure 1 summarizes this representation-level view of data design and governance.

This representation raises three empirical questions: whether explicit context relations form an enumerable and executable data space; whether richer structured semantics or explicit reasoning suffice for higher quality under a shared workload, or whether their organization into a complete execution plan matters; and whether planning coordinates persist through quality governance and expose changes in retained composition and downstream behavior. The results reveal two non-monotonic patterns. Explicit reasoning improves structural realization without a corresponding gain in pedagogical quality, while the stricter training view is associated with stronger Process behavior and weaker Result behavior. Synthetic-data quality therefore depends on how the data space is represented, organized, and governed, as well as on the Generator.

We make three contributions:

- We formulate pre-generation data planning over independently addressable semantic objects and their legal relations. In the educational setting, Content and Task are separated from the complete context so that boundaries, attribute ownership, and admissible combinations can be enumerated and planned before generation.
- We compile legal Content–Task Pairs into executable Jobs that organize role, interaction, cognitive progression, pedagogical actions, and closing requirements. Integer Slot allocation converts target dataset composition into candidate workload.
- We preserve planning identities across the synthesis lifecycle, allowing retained coverage and composition to map back to the original planning space and supporting planned-to-retained composition audits and bucket-level deficit localization when target counts are specified.

2 Related Work

Educational dialogue and pedagogical context. Educational dialogue research has progressively expanded the content and structure of pedagogical contexts. Tutoring theory and intelligent tutoring systems model learning as adaptive interaction involving diagnosis, questioning, scaffolding, feedback, and adjusted support (Wood et al., 1976; Graesser et al., 2005; VanLehn, 2011). CIMA, MathDial, and Book2Dial bring language learning, problem solving, and textbook grounding into data-driven dialogue (Stasaski et al., 2020; Macina et al., 2023; Wang et al., 2024a). Recent work further represents pedagogical actions, error remediation, and multidimensional tutoring quality through planning, turn-level verification, open-ended tutor evaluation, and student simulation (Kwon et al., 2024; Li et al., 2026; Daheim et al., 2024; Wang et al., 2025; Maurya et al., 2025; Macina et al., 2025; Scarlatos et al., 2026). This trajectory establishes what a pedagogically meaningful context may contain and how tutoring behavior may unfold. C^4-EDS addresses the upstream semantic organization of these contexts by representing their constituents and admissible combinations as an addressable dataset space.

Synthetic-data conditioning and workflows. Synthetic-data research has steadily increased the programmability of generation conditions. Self-Instruct and Unnatural Instructions scale instruction construction; WizardLM and Magpie expand instruction complexity and elicit supervision from aligned models (Wang et al., 2023; Honovich et al., 2023; Xu et al., 2024; 2025). CodecLM, Bonito, and CRAFT condition task-specific synthesis on target distributions, domain corpora, or retrieved sources (Wang et al., 2024b; Nayak et al., 2024; Ziegler et al., 2025). DataDreamer, DataEnvGym, and reference-level feedback extend programmability to reproducible workflows, feedback-driven strategies, and multistage production (Patel et al., 2024; Khan et al., 2025; Mehri et al., 2026); controlled comparisons further show that generation strategies and dataset properties require direct evaluation (Kim et al., 2025). C^4-EDS asks whether the semantic composition of these increasingly rich conditions can form an enumerable and allocatable planning space.

Quality governance and structured generation. Quality governance and structured generation provide complementary forms of control. Model-based evaluation, data selection, and feedback-driven methods support ranking, filtering, diagnosis, and subsequent refinement; recent studies also characterize judge bias, quality–diversity trade-offs, and the effects of selection strategies (Liu et al., 2023; Zheng et al., 2023; Chen et al., 2024a; Lee et al., 2025; Chen et al., 2024b; Liu et al., 2024; Li et al., 2024; Bukharin et al., 2024; Iskander et al., 2024; He et al., 2023; Khan et al., 2025; Mehri et al., 2026). Query languages and grammar- or schema-constrained decoding control the output surface during generation (Beurer-Kellner et al., 2023; Scholak et al., 2021; Geng et al., 2023; Dong et al., 2025). C^4-EDS preserves pre-generation planning identities so that sample-level quality outcomes map back to Pair/Job coordinates and planned and retained compositions remain comparable; structured decoding continues to provide formal output control.

Together, these lines of work make generation contexts richer, more programmable, and easier to evaluate. C^4-EDS advances their semantic composition into an addressable dataset space: it represents context constituents and admissible relations before assembly, compiles them into executable workloads, and governs outcomes along the same coordinates.

3 Method

C^4-EDS organizes data synthesis around the representation of a generation context. Before a complete context is assembled, Content and Task semantics are represented as independent objects and legal cross-object relations are made explicit. Selected relations are instantiated as executable Jobs and assigned candidate workload through Slots. Each generated Candidate retains this planning identity, allowing pre-generation planning, staged quality governance, and post-generation composition auditing to share the same coordinates. Figure 2 summarizes this lifecycle.

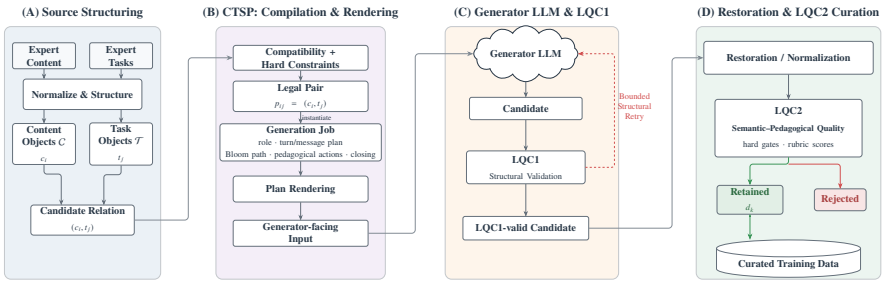

Figure 2: C^4-EDS lifecycle. Structured Content and Task objects form legal Pairs; CTSP compiles each selected Pair into a Job and rendered generator input; LQC1 validates structure, and restoration followed by LQC2 produces the curated training view while retaining planning identities.

3.1 From coupled generation contexts to an explicit planning space

Dataset-level planning requires semantic constituents that can be addressed before they are fused into a generation context. Let $\mathcal{C} = \{c_1, \ldots, c_{|\mathcal{C}|}\}$ be the Content registry and $\mathcal{T} = \{t_1, \ldots, t_{|\mathcal{T}|}\}$ the Task registry. A Content object records source material and stable curricular attributes. A Task object records the pedagogical objective, strategy, cognitive requirements, and applicable instructional attributes. Decoupling has a representational meaning here: Content- and Task-side semantics receive separate identities and ownership, enabling explicit evaluation of cross-object relations before a complete generation context is assembled. The present instantiation applies this domain-specific structuring to K–12 Classical Chinese content and teaching tasks.

Decoupling creates a candidate cross-object relation space. Compatibility and hard constraints determine which recombinations are admissible. A compatibility predicate $a(c,t) \in \{0,1\}$ applies type/module constraints, followed by a small hard-pruning set for relations ruled out by explicit domain evidence. The legal relation space is

$$\mathcal{P} = \{(c,t) : c \in \mathcal{C}, t \in \mathcal{T}, a(c,t) = 1\} \setminus \mathcal{H},$$

where each $p \in \mathcal{P}$ is a *Pair*: a legal, addressable Content–Task relation that provides the semantic coordinate for subsequent execution and composition planning.

3.2 Executable Jobs and candidate Slots

An admissible relation must become executable before it can carry dataset workload. A Job $j \in \mathcal{J}$ instantiates a legal Pair under an execution plan. The Pair states what may be generated; the Job states how that relation will be realized. Its plan may specify a role relation, turn and message structure, cognitive progression, pedagogical actions, and closing condition. These values are defined for each Job.

For a given construction run, $K_j \in \mathbb{Z}_{\geq 0}$ is the number of candidate Slots assigned to Job j. Slot (j,k) allocates one preplanned candidate opportunity. Structural or transient retries create additional runtime attempts for the same Slot, keeping dataset allocation separate from runtime calls. Expert attributes such as task frequency can serve as optional priors for a target policy; K_j follows the composition selected for the current construction run.

Job compilation determines which semantic relation and execution plan a Candidate realizes; Slot allocation determines how candidate opportunities are distributed across those executable plans. These define controllability at the levels of executable planning and dataset allocation.

The selected Pair semantics and Job-level execution requirements are then materialized as generator input. Each generated Candidate retains its Pair, Job, and Slot lineage.

3.3 Staged quality governance

Quality outcomes remain attached to their original planning coordinates through Candidate lineage. LQC1 checks output structure, message organization, and completeness, with limited retries inside the same Slot. Each E2 condition uses a structural contract matched to its declared output surface.

Restoration and normalization then stabilize authoritative references and the training surface. LQC2 evaluates four semantic hard gates (source grounding, task alignment, prompt non-leakage, and absence of unsupported hallucination) and five continuous pedagogical dimensions. Staging structural validation and semantic evaluation keeps retry outcomes and pedagogical quality distinct while preserving Pair, Job, and Slot lineage throughout.

3.4 Distribution planning and preservation

In C^4-EDS, dataset-level controllability means that a desired composition over explicit planning coordinates can be converted into integer candidate assignments before generation and audited in the same coordinates after quality filtering. Stable Job identities make this allocation a direct consequence of the representation. Choose an explicit planning dimension d with bucket space $\Omega_d = \{\omega_1, \ldots, \omega_m\}$ and a stable map $g_d : \mathcal{J} \to \Omega_d$. A usable dimension must be enumerable before generation, supported by legal Jobs, and recoverable from candidate lineage. The natural support of bucket ω is

$$S_\omega = |\{j \in \mathcal{J} : g_d(j) = \omega\}|.$$

S_ω describes feasible Job support. Target composition is specified separately.

Given total Slot budget B and target proportions $\pi_d^*(\omega)$ with $\sum_{\omega \in \Omega_d} \pi_d^*(\omega) = 1$, integer allocation produces bucket targets

$$N_\omega^* = \mathrm{Allocate}_{\mathbb{Z}}(B, \pi_d^*(\omega)), \qquad \sum_\omega N_\omega^* = B,$$

followed by Job allocations satisfying

$$N_\omega^* = \sum_{j : g_d(j) = \omega} K_j.$$

Finite budgets exactly realize the integer composition $\tilde{\pi}_d^*(\omega) = N_\omega^*/B$, which may differ from the continuous target through rounding.

After quality filtering, let R_ω be the retained Candidate count. Planned and retained compositions are

$$p_{\mathrm{plan}}^{(d)}(\omega) = \frac{N_\omega^*}{\sum_\nu N_\nu^*}, \qquad p_{\mathrm{ret}}^{(d)}(\omega) = \frac{R_\omega}{\sum_\nu R_\nu}.$$

We measure composition drift by

$$D_{\mathrm{pres}}^{(d)} = \mathrm{TVD}(p_{\mathrm{ret}}^{(d)}, p_{\mathrm{plan}}^{(d)}) = \frac{1}{2} \sum_{\omega \in \Omega_d} |p_{\mathrm{ret}}^{(d)}(\omega) - p_{\mathrm{plan}}^{(d)}(\omega)|.$$

TVD measures planned-to-retained composition drift; target-allocation accuracy is reported separately. If a use case specifies absolute target counts, $\Delta_\omega = \max(N_\omega^* - R_\omega, 0)$ identifies a bucket-level deficit. The same coordinates support optional directed supplementation by allocating new Slots to eligible Jobs.

4 Experiments

4.1 Questions and shared evaluation setting

E1 tests whether the representation forms an enumerable and executable planning space. E2 tests how progressively structured generation conditions differ under a matched workload. E3 tests how complete curated training views are associated with downstream Result

and Process behavior. Human evaluation and calibration protocols appear in Appendix A; matched inputs and outputs in Appendix B; and distribution, training-view, and diagnostic records in Appendices C–E.

The E2 comparison uses 128 legal Pairs expanded into 256 role-conditioned Jobs, with $K_j = 2$ and 512 candidates per condition. All conditions share the same Jobs, candidate budget, generator, processing stages, and LQC2. Their LQC1 validators match the declared output surface. The conditions are

$$B0 = (P_0, J_0), \quad B1 = (P_1, J_0), \quad B1\text{-CoT} = (P_1, J_{\text{CoT}}), \quad \text{CTSP} = (P_1, J_{\text{full}}).$$

P_0 contains basic source and task conditions. P_1 contains structured Pair/Task semantics. B1-CoT adds an explicit reasoning contract; its structured Pair/Task payload is identical to B1 for all 256 Jobs. CTSP preserves the core P_1 semantics and materializes a complete Job-level execution plan. This is a progressive block-level comparison. Only B1 to B1-CoT approaches a single-factor contrast.

The adjacent contrasts isolate progressively richer requirement blocks. B0 to B1 compares basic source/task conditions with structured Pair/Task semantics and is a semantic-block comparison. B1 to B1-CoT holds those structured semantics fixed and adds an explicit reasoning requirement, making it the closest comparison to a single-factor contrast. B1-CoT to CTSP compares reasoning-only execution with full Job-level planning over role, turn/message structure, cognitive progression, pedagogical actions, and closing requirements. This final transition supports interpretation at the level of the complete execution block; individual Job-field effects remain outside the design.

4.2 E1: an enumerable, executable planning space

The registries contain 134 Content and 42 Task objects, yielding 5,628 theoretical relations. Compatibility retains 2,814 relations and hard pruning removes seven, leaving 2,807 legal Pairs. Expansion over two role spaces produces 5,614 Jobs. The important result is operational: every legal relation and Job has a stable identity that can be counted, addressed, and allocated before generation.

We separately test target-to-workload materialization. Three target families, three generator-facing representations, and three seeds produce 27 settings, each with a 512-Slot budget. All 27 materialize exactly: bucket-level L_1 allocation error is zero, and the audit finds no missing, extra, duplicate, or orphan Slots. The result establishes exact realization of the integer Slot plan; continuous targets are represented by their finite-budget integerization.

4.3 E2: structural realization, semantic quality, and preserved coordinates

Panels A and B of Table 1 separate structural realization from semantic–pedagogical retention. B0 has the highest first-pass structural closure; CTSP has the highest LQC2 hard-pass rate and mean quality. The two stages therefore describe different outcome axes. Relative to B0, B1 has lower Pass@1 (63.67% versus 87.30%) and substantially more structural retry calls (333 versus 72), while its hard-pass rate rises from 64.65% to 75.98% and mean quality from 2.892 to 3.780. Structured Pair/Task semantics are associated with higher semantic–pedagogical retention alongside harder first-pass structural realization, yielding a mixed profile.

B1 to B1-CoT is the closest progressive contrast. Across Pass@1/2/3, structural closure changes from 63.67/87.50/93.55% to 67.97/92.38/97.27%, while structural retry calls fall from 333 to 220. In Panel B, hard pass falls from 75.98% to 48.44% and mean quality from 3.780 to 3.037. The B1-CoT condition exhibits more grounding failures (68 versus 58), alignment failures (120 versus 45), and leakage failures (188 versus 4), while hallucination changes from 106 to 101. Explicit reasoning is associated with better early structural realization and lower semantic–pedagogical retention; the quality difference is concentrated in alignment and leakage, with little change in hallucination.

Table 1: Matched E2 outcomes under the four generation conditions. Panel A reports cumulative LQC1 structural closure: Pass@1, Pass@2, and Pass@3 denote validity within zero, one, and two structural retries, respectively. All conditions eventually reach 512/512 LQC1-valid terminal records under retry-until-valid governance; final closure serves as a sanity check, while early Pass@k and retry calls provide the comparison. Panel B reports the LQC2 hard-pass rate, mean quality, and four multi-label hard-gate failure counts. Bold marks the best value in each column according to metric direction; ties are bolded jointly.

Panel A: LQC1 structural realization

Condition	Pass@1	Pass@2	Pass@3	Retry calls
B0	**87.30**	**98.63**	**100.00**	**72**
B1	63.67	87.50	93.55	333
B1-CoT	67.97	92.38	97.27	220
CTSP	72.46	87.89	94.92	253

Panel B: LQC2 semantic–pedagogical quality

Condition	Hard pass	Mean quality	Ground.	Align.	Leakage	Halluc.
B0	64.65	2.892	51	86	**4**	124
B1	75.98	3.780	58	45	**4**	106
B1-CoT	48.44	3.037	68	120	188	101
CTSP	**88.28**	**4.200**	**23**	**8**	22	**36**

Relative to B1-CoT, CTSP has higher Pass@1 (72.46% versus 67.97%) and lower Pass@2/3 (87.89/94.92% versus 92.38/97.27%), with more retry calls (253 versus 220). Its semantic–pedagogical outcomes are stronger: hard pass is 88.28% versus 48.44%, mean quality is 4.200 versus 3.037, and grounding, alignment, leakage, and hallucination failures are 23/8/22/36 versus 68/120/188/101. CTSP changes role, turn planning, cognitive progression, pedagogical actions, and closing requirements jointly, so this comparison supports a block-level interpretation; isolated field effects remain outside the design.

Because every retained Candidate keeps its planning identity, each condition can be compared against the same planned composition. On the representative Task×Role axis, TVD is 0.0861 for B0, 0.1458 for B1, 0.1794 for B1-CoT, and 0.0489 for CTSP. All four conditions retain all 32 Task×Role cells; the differences arise from relative retention across cells. Across the seven audited axes, CTSP has the lowest TVD on six, with Grade as the exception. Appendix C reports the complete audit.

Independent human evaluation. Two independent evaluators assessed the four condition outputs for 64 matched cases (256 responses). They applied the same four binary gates and five continuous pedagogical dimensions without using the automatic labels. Hard-pass agreement was 73.05% ($\kappa = 0.441$). Their continuous scores had Spearman $\rho = 0.636$ and Pearson $r = 0.627$, indicating shared ordering signal alongside substantial scale sensitivity. We use $H_i = (H_{i,A} + H_{i,B})/2$ as the primary continuous human reference and retain evaluator-specific results as sensitivity analyses.

Panel means were 2.879 for B0, 3.723 for B1, 2.955 for B1-CoT, and 3.932 for CTSP. The matched CTSP–B1 mean difference was +0.209 (Holm-adjusted paired Wilcoxon $p = 0.0215$). Binary hard-pass judgments do not support a claim that CTSP exceeds B1 on every human measure: evaluator A passed 40 versus 48 responses, while evaluator B passed 45 for both conditions.

Automatic–human calibration. All 256 responses join uniquely to the primary automatic LQC2 records. Automatic and panel quality have Spearman $\rho = 0.685$ with cluster-bootstrap 95% CI $[0.601, 0.752]$. Lin's concordance correlation coefficient is 0.634 $[0.541, 0.713]$, MAE is 0.594, and automatic scores have mean bias +0.191 $[0.101, 0.282]$. The evaluator captures meaningful ranking signal but has incomplete absolute calibration.

We use it for scalable screening, ranking, and diagnosis, with human assessment retained as the pedagogical reference.

4.4 E3: DOWNSTREAM BEHAVIOR UNDER COMPLETE TRAINING VIEWS

E3 compares a Base model with models trained on a 512-record CTSP-LQC1 view and a 145-record strict CTSP-LQC2 view. The views retain the sample count, tokens, coverage, and composition formed at their respective governance stages, defining a pipeline-level comparison of complete training views.

Table 2: Downstream scores for the three complete training views. Bold marks the highest value in each metric column.

Training view	Final	Result	Process
Base	62.1	**67.8**	56.4
CTSP-LQC1-SFT	59.4	60.8	58.0
CTSP-LQC2-SFT	**62.3**	63.0	**61.5**

Relative to Base, the strict view changes Final by +0.2, Result by −4.8, and Process by +5.1 (Table 2). The association is a Result–Process trade-off. It is consistent with curation changing which behavior is most available for learning, while the view-level confounds preclude attribution to LQC2 alone.

5 DISCUSSION

The central role of C^4-EDS is to change the representational unit of data synthesis. Once Content, Task, and their legal relations have stable identities before a complete generation context is formed, the data space can be enumerated, allocated, and audited directly. E1 shows that these coordinates can be materialized as candidate workloads, making dataset composition an actionable design object before generation.

E2 reveals a non-monotonic relation between generation conditions and pedagogical quality. Under the same structured Pair/Task semantics, explicit reasoning improves early structural realization while semantic–pedagogical quality declines, with the largest differences in task alignment and prompt leakage. Full Job-level planning organizes role relations, cognitive progression, pedagogical actions, interaction structure, and closing requirements into an execution plan and is associated with higher overall pedagogical quality. The organization of execution requirements therefore matters alongside the presence of explicit reasoning.

Stable planning identities also connect quality governance to the same data space. Retained Candidates map back to their original Pair, Job, and planning region, so selection changes both sample count and relative dataset composition. The TVD audit shows different retained compositions under the shared pre-generation plan, with lower drift for CTSP on most audited axes. Automatic LQC2 and dual-human evaluation show stable ranking correspondence alongside a measurable scale offset, consistent with evidence that model-based judges remain distinct from human assessment (Zheng et al., 2023; Chen et al., 2024a; Lee et al., 2025).

E3 indicates that data quality is target-dependent. The stricter training view is associated with stronger Process behavior and lower Result performance, forming a target-specific trade-off. Related work likewise finds that selection, validation, and quality–diversity choices alter downstream outcomes (Li et al., 2024; Bukharin et al., 2024; Iskander et al., 2024). For educational data, quality must therefore be interpreted in relation to the behavior the model is intended to learn.

6 LIMITATIONS

The empirical study centers on K–12 Classical Chinese literature. The representation principle can be instantiated elsewhere, while each domain requires its own Content/Task ontol-

ogy, compatibility rules, and attributes. E2 compares blocks of Job-level execution requirements and supports block-level interpretation. Human evaluation covers two evaluators and 64 matched cases; automatic–human correspondence remains specific to the present rubric, judge, and generation setting. E3 compares complete training views that differ in size, tokens, coverage, and composition, so its evidence concerns view-level associations with downstream behavior. Composition planning also requires stable, enumerable dimensions recoverable from Candidate lineage. The present work defines optional directed supplementation as an operational mechanism.

7 Conclusion

This work reconsiders synthetic-data construction through the representation of generation contexts. C^4-EDS organizes Content and Task as independently addressable semantic objects, makes their legal relations explicit, compiles those relations into executable Jobs and integer Slot workloads, and preserves planning identities through generation and quality governance. The resulting data space can be planned before generation and audited along the same coordinates after selection.

The experiments show that this representation can carry dataset-level workloads and expose non-monotonic relations between generation conditions and data quality: explicit reasoning can improve structural realization without a corresponding gain in pedagogical quality, while training views formed under different quality criteria emphasize different Result and Process behavior. More generally, synthetic-data construction requires decisions about individual examples together with an explicit representation of the data space they occupy and mechanisms for planning and governing data within that space.

AI use statement

Generative AI was used in the reported pipeline to produce synthetic dialogue candidates and automatic quality judgments. It was also used during manuscript preparation for language editing, structural consistency checks, evidence tracing, coding support, and LaTeX assistance. The authors verified AI-assisted text and analysis against the frozen scientific note and archived experimental records and take responsibility for the final claims, code, and artifacts.

Ethics statement

The system prepares tutoring data from expert-authored educational materials. Certification of learning outcomes and instructional decisions remain the responsibility of educators. Generated records may contain grounding, leakage, or pedagogical errors; learner-facing use requires teacher oversight, privacy safeguards, and separate effectiveness evaluation.

References

Luca Beurer-Kellner, Marc Fischer, and Martin T. Vechev. Prompting is programming: A query language for large language models. *Proceedings of the ACM on Programming Languages*, 7(PLDI):1946–1969, 2023. doi: 10.1145/3591300. URL https://doi.org/10.1145/3591300.

Alexander Bukharin, Shiyang Li, Zhengyang Wang, Jingfeng Yang, Bing Yin, Xian Li, Chao Zhang, Tuo Zhao, and Haoming Jiang. Data diversity matters for robust instruction tuning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 3411–3425. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-emnlp.195. URL https://aclanthology.org/2024.findings-emnlp.195/.

Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. Humans or LLMs as the judge? a study on judgement bias. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8301–8327. Association for Computational Linguistics, 2024a. doi: 10.18653/v1/2024.emnlp-main.474. URL https://aclanthology.org/2024.emnlp-main.474/.

Lichang Chen, Shiyang Li, Jun Yan, Hai Wang, Kalpa Gunaratna, Vikas Yadav, Zheng Tang, Vijay Srinivasan, Tianyi Zhou, Heng Huang, and Hongxia Jin. AlpaGasus: Training a better Alpaca with fewer data. In *The Twelfth International Conference on Learning Representations*, 2024b. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/9543942c237ded1b39b1fd37259ff88e-Abstract-Conference.html.

Nico Daheim, Jakub Macina, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. Stepwise verification and remediation of student reasoning errors with large language model tutors. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8386–8411. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.478. URL https://aclanthology.org/2024.emnlp-main.478/.

Yixin Dong, Charlie F. Ruan, Yaxing Cai, Ziyi Xu, Yilong Zhao, Ruihang Lai, and Tianqi Chen. XGrammar: Flexible and efficient structured generation engine for large language models. In *Proceedings of Machine Learning and Systems*, volume 7, 2025. URL https://proceedings.mlsys.org/paper_files/paper/2025/hash/5c20ca4b0b20b0bd2f1d839dc605e70f-Abstract-Conference.html.

Saibo Geng, Martin Josifoski, Maxime Peyrard, and Robert West. Grammar-constrained decoding for structured NLP tasks without finetuning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 10932–10952. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.674. URL https://aclanthology.org/2023.emnlp-main.674/.

Arthur C. Graesser, Patrick Chipman, Brian C. Haynes, and Andrew Olney. AutoTutor: An intelligent tutoring system with mixed-initiative dialogue. *IEEE Transactions on Education*, 48(4):612–618, 2005. doi: 10.1109/TE.2005.856149. URL https://doi.org/10.1109/TE.2005.856149.

Zexue He, Marco Tulio Ribeiro, and Fereshte Khani. Targeted data generation: Finding and fixing model weaknesses. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8506–8520. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-long.474. URL https://aclanthology.org/2023.acl-long.474/.

Or Honovich, Thomas Scialom, Omer Levy, and Timo Schick. Unnatural instructions: Tuning language models with (almost) no human labor. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14409–14428. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-long.806. URL https://aclanthology.org/2023.acl-long.806/.

Shadi Iskander, Sofia Tolmach, Ori Shapira, Nachshon Cohen, and Zohar Karnin. Quality matters: Evaluating synthetic data for tool-using LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 4958–4976. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.285. URL https://aclanthology.org/2024.emnlp-main.285/.

Zaid Khan, Elias Stengel-Eskin, Jaemin Cho, and Mohit Bansal. DataEnvGym: Data generation agents in teacher environments with student feedback. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/846d3ef94d8c1a833a62cff05569c851-Abstract-Conference.html.

Seungone Kim, Juyoung Suk, Xiang Yue, Vijay Viswanathan, Seongyun Lee, Yizhong Wang, Kiril Gashteovski, Carolin Lawrence, Sean Welleck, and Graham Neubig. Evaluating language models as synthetic data generators. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6385–6403. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.320. URL https://aclanthology.org/2025.acl-long.320/.

Soonwoo Kwon, Sojung Kim, Minju Park, Seunghyun Lee, and Kyuseok Kim. BIPED: Pedagogically informed tutoring system for ESL education. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3389–3414. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.186. URL https://aclanthology.org/2024.acl-long.186/.

Yukyung Lee, JoongHoon Kim, Jaehee Kim, Hyowon Cho, Jaewook Kang, Pilsung Kang, and Najoung Kim. CheckEval: A reliable LLM-as-a-judge framework for evaluating text generation using checklists. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 15771–15798. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.796. URL https://aclanthology.org/2025.emnlp-main.796/.

Ming Li, Yong Zhang, Zhitao Li, Jiuhai Chen, Lichang Chen, Ning Cheng, Jianzong Wang, Tianyi Zhou, and Jing Xiao. From quantity to quality: Boosting LLM performance with self-guided data selection for instruction tuning. 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. 7602–7635. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.naacl-long.421. URL https://aclanthology.org/2024.naacl-long.421/.

Zechen Li, Qiannan Zhu, Mei Wang, Jia Li, and Hua Huang. Planning-guided tutoring with assessment-driven memory for pedagogical LLM tutors. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7165–7188. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.325. URL https://aclanthology.org/2026.acl-long.325/.

Wei Liu, Weihao Zeng, Keqing He, Yong Jiang, and Junxian He. What makes good data for alignment? a comprehensive study of automatic data selection in instruction tuning. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/6091f2bb355e960600f62566ac0e2862-Abstract-Conference.html.

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. doi: 10.18653/v1/2023.emnlp-main.153. URL https://aclanthology.org/2023.emnlp-main.153/.

Jakub Macina, Nico Daheim, Sankalan Chowdhury, Tanmay Sinha, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. MathDial: A dialogue tutoring dataset with rich pedagogical properties grounded in math reasoning problems. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 5602–5621. Association

for Computational Linguistics, 2023. doi: 10.18653/v1/2023.findings-emnlp.372. URL https://aclanthology.org/2023.findings-emnlp.372/.

Jakub Macina, Nico Daheim, Ido Hakimi, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. MathTutorBench: A benchmark for measuring open-ended pedagogical capabilities of LLM tutors. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 204–221. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.11. URL https://aclanthology.org/2025.emnlp-main.11/.

Kaushal Kumar Maurya, K. V. Aditya Srivatsa, Kseniia Petukhova, and Ekaterina Kochmar. Unifying ai tutor evaluation: An evaluation taxonomy for pedagogical ability assessment of LLM-powered ai tutors. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 1234–1251. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.naacl-long.57. URL https://aclanthology.org/2025.naacl-long.57/.

Shuhaib Mehri, Xiusi Chen, Heng Ji, and Dilek Hakkani-Tür. Beyond sample-level feedback: Using reference-level feedback to guide data synthesis. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 141–164. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.eacl-long.7. URL https://aclanthology.org/2026.eacl-long.7/.

Nihal Nayak, Yiyang Nan, Avi Trost, and Stephen H. Bach. Learning to generate instruction tuning datasets for zero-shot task adaptation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 12585–12611. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-acl.748. URL https://aclanthology.org/2024.findings-acl.748/.

Ajay Patel, Colin Raffel, and Chris Callison-Burch. DataDreamer: A tool for synthetic data generation and reproducible LLM workflows. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3781–3799. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.208. URL https://aclanthology.org/2024.acl-long.208/.

Alexander Scarlatos, Jaewook Lee, Simon Woodhead, and Andrew Lan. Simulated students in tutoring dialogues: Substance or illusion? In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 42349–42385. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.1960. URL https://aclanthology.org/2026.acl-long.1960/.

Torsten Scholak, Nathan Schucher, and Dzmitry Bahdanau. PICARD: Parsing incrementally for constrained auto-regressive decoding from language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 9895–9901. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.emnlp-main.779. URL https://aclanthology.org/2021.emnlp-main.779/.

Katherine Stasaski, Kimberly Kao, and Marti A. Hearst. CIMA: A large open access dialogue dataset for tutoring. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 52–64. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.bea-1.5. URL https://aclanthology.org/2020.bea-1.5/.

Kurt VanLehn. The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist*, 46(4):197–221, 2011. doi: 10.1080/00461520.2011.611369. URL https://doi.org/10.1080/00461520.2011.611369.

Jian Wang, Yinpei Dai, Yichi Zhang, Ziqiao Ma, Wenjie Li, and Joyce Chai. Training turn-by-turn verifiers for dialogue tutoring agents: The curious case of LLMs as your coding tutors. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 12416–12436. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-acl.642. URL https://aclanthology.org/2025.findings-acl.642/.

Junling Wang, Jakub Macina, Nico Daheim, Sankalan Pal Chowdhury, and Mrinmaya Sachan. Book2Dial: Generating teacher student interactions from textbooks for cost-effective development of educational chatbots. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 9707–9731. Association for Computational Linguistics, 2024a. doi: 10.18653/v1/2024.findings-acl.578. URL https://aclanthology.org/2024.findings-acl.578/.

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. doi: 10.18653/v1/2023.acl-long.754. URL https://aclanthology.org/2023.acl-long.754/.

Zifeng Wang, Chun-Liang Li, Vincent Perot, Long Le, Jin Miao, Zizhao Zhang, Chen-Yu Lee, and Tomas Pfister. CodecLM: Aligning language models with tailored synthetic data. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 3712–3729. Association for Computational Linguistics, 2024b. doi: 10.18653/v1/2024.findings-naacl.235. URL https://aclanthology.org/2024.findings-naacl.235/.

David Wood, Jerome S. Bruner, and Gail Ross. The role of tutoring in problem solving. *Journal of Child Psychology and Psychiatry*, 17(2):89–100, 1976. doi: 10.1111/j.1469-7610.1976.tb00381.x. URL https://doi.org/10.1111/j.1469-7610.1976.tb00381.x.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, Qingwei Lin, and Daxin Jiang. WizardLM: Empowering large pre-trained language models to follow complex instructions. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/82eec786fdfbbfa53450c5feb7d1ac92-Abstract-Conference.html.

Zhangchen Xu, Fengqing Jiang, Luyao Niu, Yuntian Deng, Radha Poovendran, Yejin Choi, and Bill Yuchen Lin. Magpie: Alignment data synthesis from scratch by prompting aligned LLMs with nothing. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/be06e3802e9411381feece79b4d960c1-Abstract-Conference.html.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-2020. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/91f18a1287b398d378ef22505bf41832-Abstract-Datasets_and_Benchmarks.html. Datasets and Benchmarks Track.

Ingo Ziegler, Abdullatif Köksal, Desmond Elliott, and Hinrich Schütze. CRAFT your dataset: Task-specific synthetic dataset generation through corpus retrieval and augmentation. *Transactions of the Association for Computational Linguistics*, 13:1693–1721, 2025. doi: 10.1162/tacl.a.56. URL https://aclanthology.org/2025.tacl-1.76/.

A Human Evaluation and Automatic–Human Calibration

A.1 Protocol and rubric

The human study sampled 64 matched case identities and evaluated all four E2 conditions, yielding 256 response-level units and 512 independent rating events. Both evaluators were trained graduate students in Chinese language and literature. They rated the archived outputs independently. No third-rater adjudication was used.

The rubric separates acceptability from continuous pedagogical quality. For evaluator r, the binary label is

$$Y_{i,r}^{\text{hard}} = \prod_{k=1}^{4} g_{i,r,k},$$

where the gates assess source grounding, task alignment, prompt non-leakage, and absence of unsupported hallucination. Five 1–5 dimensions assess instructional-intent realization, contextual-affordance use, scaffolded progression, dialogic adaptivity, and imitation-signal quality. The overall score is their unweighted mean, keeping the aggregation rule transparent and each dimension traceable.

The automatic evaluator uses aligned constructs so that machine and human judgments can be compared dimension by dimension. Human evaluators score the generated responses independently and without access to automatic results; the subsequent calibration analysis measures correspondence between the two judgments.

A.2 Inter-evaluator results and paired method comparisons

Hard-pass agreement was 187/256 (73.05%), with Cohen's $\kappa = 0.441$. Continuous overall scores had Spearman $\rho = 0.636$ and Pearson $r = 0.627$. Evaluator A used a higher score scale on average; we therefore define the primary continuous reference as $H_i = (H_{i,A} + H_{i,B})/2$ and report individual evaluators as sensitivity analyses.

Table 3: Dual-human results on 64 matched cases per condition.

Condition	A hard pass	B hard pass	Panel mean
B0	34/64	37/64	2.879
B1	48/64	45/64	3.723
B1-CoT	29/64	27/64	2.955
CTSP	40/64	45/64	3.932

For the panel score, matched CTSP differences were +1.053 versus B0, +0.209 versus B1, and +0.977 versus B1-CoT; Holm-adjusted paired Wilcoxon tests gave $p < 0.001$, $p = 0.0215$, and $p < 0.001$, respectively. Binary outcomes show greater evaluator sensitivity, so continuous pedagogical scores provide the primary human comparison.

A.3 Automatic–human calibration

Human units were joined by case identity and condition to the canonical primary LQC2 pool; 256/256 matched, with no duplicate or silently removed analysis rows. We used a 10,000-replicate percentile cluster bootstrap over the 64 case identities, retaining all four conditions for a sampled identity (seed 20260922).

On the 187 cases where the two evaluators agreed on hard pass, automatic agreement was 74.9%, $\kappa = 0.416$, and MCC = 0.438. This consensus subset covers 73.0% of responses. Across the five continuous dimensions, Spearman correlations range from 0.601 to 0.693. Condition- and evaluator-specific results show scale sensitivity; the automatic evaluator serves as a screening and diagnostic signal, with the human panel as the pedagogical reference.

Table 4: Primary continuous calibration against the dual-human panel.

Metric	Estimate	95% cluster-bootstrap CI
Spearman ρ	0.685	[0.601, 0.752]
Pearson r	0.653	[0.564, 0.731]
Lin's CCC	0.634	[0.541, 0.713]
MAE	0.594	[0.527, 0.665]
Bias (automatic − panel)	+0.191	[0.101, 0.282]

B Matched-case Inputs and Outputs

B.1 Matched Case and Comparability

The case uses the archived *Night Tour of Chengtian Temple* sentence-translation task. A fixed matched identity determines the selection. The four records share the upstream Content–Task semantic anchor, pre-specified candidate position, generator, processing stages, and LQC2 protocol, while their execution requirements and output-matched LQC1 validators follow the corresponding conditions. The case provides a qualitative view of matched inputs and outputs; aggregate E2 results provide condition-level evidence.

B.2 Progressive Comparison Design

B0 provides the source, task, and shared constraints. B1 exposes structured Pair/Task semantics. B1-CoT retains the same structured semantics and adds an explicit reasoning requirement. CTSP exposes the core structured semantics within a complete Job-level execution plan. The comparison is progressive and block-level; only B1 to B1-CoT approaches a single-factor contrast.

B.3 Generator-visible Input Comparison

Table 5: Semantic and task information explicitly visible to the generator. A checkmark records an input-design fact, not quality.

Information	B0	B1	B1-CoT	CTSP
Source text and task identity	✓	✓	✓	✓
Structured task semantics	–	✓	✓	✓
Teaching strategy and principles	–	✓	✓	✓
Cognitive and exam attributes	–	✓	✓	✓

Table 6: Generator-visible interaction and execution planning.

Planning element	B0	B1	B1-CoT	CTSP
Job-specific role relation	–	–	–	✓
Turn/message plan	–	–	–	✓
Job-specific cognitive progression	–	–	–	✓
Per-turn pedagogical actions	–	–	–	✓
Closing condition	–	–	–	✓

B1 and B1-CoT contain identical structured Pair/Task payloads for all 256 E2 Jobs. The three tables record generator-visible input-design facts only; they carry no quality judgment.

B.4 Verbatim Generator-facing Inputs

The following four inputs reproduce the archived generator-facing text for the matched case.

B.4.1 B0

Table 7: Generator-visible reasoning and output requirements.

Requirement	B0	B1	B1-CoT	CTSP
Explicit reasoning contract	–	–	✓	✓
Reasoning–answer alignment	–	–	✓	✓
Structured messages output	✓	✓	✓	✓
Planned message count	–	–	–	✓

System prompt

你是中文学习助手。只输出一个 JSON 对象，顶层只包含 messages 字段。messages 使用 user/assistant 交替，每条包含 role 与 content。

User prompt

请围绕下面文本和学习任务生成一段适合课堂使用的教学问答。问答要清楚、贴合原文，不编造题目没有给出的材料。

质量约束：assistant 不能扮演学生替答，不能在 content 中写 “学生回答/老师追问” 等角色标签；必须以教学支持者身份回应，可以追问、点拨、诊断、检查或布置练习。每个解释、判断和例子都要扣住原文词句；不能编造原文没有的字词、出处、情节或题目材料。翻译任务默认是古诗或文言文译成现代汉语，不要把古诗或文言文翻译成英文，除非任务明确要求英译。不能只给完整答案，至少体现一个引导问题、诊断判断、支架步骤、检查点或巩固练习。对话要自然，少用模板化编号、加粗堆叠和通用赏析套话。

文本：记承天寺夜游（苏轼）
原文：元丰六年十月十二日夜，解衣欲睡，月色入户，欣然起行。念无与为乐者，遂至承天寺寻张怀民。怀民亦未寝，相与步於中庭。庭下如积水空明，水中藻、荇交横，盖竹柏影也。

任务：句子翻译

Response format used for generation

1. type: json_schema
2. schema name: sft_messages
3. top-level required field: messages
4. minimum messages: 2
5. each message contains only role and content

B.4.2 B1

System prompt

你是 K12 中文教育对话生成器。只输出一个 JSON 对象，顶层只包含 messages 字段。messages 使用 user/assistant 交替，每条消息只包含 role 与 content。

User prompt

根据以下 metadata 生成一个 tutoring dialogue JSON。metadata 只作为生成依据，不要把字段名、schema 名或内部标签写进对话内容。

质量约束：assistant 不能扮演学生替答，不能在 content 中写 “学生回答/老师追问” 等角色标签；必须以教学支持者身份回应，可以追问、点拨、诊断、检查或布置练习。每个解释、判断和例子都要扣住原文词句；不能编造原文没有的字词、出处、情节或题目材料。翻译任务默认是古诗或文言文译成现代汉语，不要把古诗或文言文翻译成英文，除非任务明确要求英译。不能只给完整答案，至少体现一个引导问题、诊断判断、支架步骤、检查点或巩固练习。对话要自然，少用模板化编号、加粗堆叠和通用赏析套话。必须把 metadata 中的 strategy.steps 和 principles 转成自然课堂动作，不要把字段名、schema 名、内部标签或 metadata 结构写进对话。

metadata={“base_ref”:{“title”:“记 承 天 寺 夜 游”,“author”:“苏 轼”,“type”:“文 言 文”,“grade”:“7-9”},“task_ref”:{“phase_name”:“第 二 环 节： 疏 通——文 言 语 言 与

句式解析”,“task_name”:“句子翻译”,“cognitive_requirements”:[“理解”,“应用”,“分析”,“评估”],“cognitive_level_max”:“评估”,“positioning”:“扫除阅读障碍，精准把握字词句义，落实“语言建构与应用”。“},”source_fields“:{”base.content“:” 元丰六年十月十二日夜，解衣欲睡，月色入户，欣然起行。念无与为乐者，遂至承天寺寻张怀民。怀民亦未寝，相与步於中庭。庭下如积水空明，水中藻、荇交横，盖竹柏影也。“,”task.strategy.method_name“:” 三步译句法 + 五法（文言）“,”task.strategy.steps“:[{”id“:1,”head“:” 标 “,”detail“:” 圈出关键词义、活用、句式特点等“},{”id“:2,”head“:” 析 “,”detail“:” 分析句子成分和结构 “},{”id“:3,”head“:” 译 “,”detail“:” 用留/替/调/补/删完成翻译，体现句式特点 “}],”task.strategy.principles“:[{”id“:1,”text“:” 翻译必须体现关键词和句式特征 “},{”id“:2,”text“:” 建立稳定翻译流程，方便教学与命题 “}],”task.exam_features“:{”frequency_level“:” 极高频 “,”focus“:[” 文言翻译 “],”question_templates“:[” 用现代汉语翻译文中画横线的句子。“],”item_types“:[” 翻译题 “]},”task.labels“:{”text_type“:[” 文言文 “],”knowledge_points“:[” 文言翻译 “,” 关键词 “,” 句式特征 “],”exam_tags“:[” 翻译 “],”scenario“:[” 疏通环节 “,” 核心考点 “],”elements“:[” 关键词 “,” 句式还原 “,” 语义连贯”]}}}

Response format used for generation

1. type: json_schema
2. schema name: sft_messages
3. top-level required field: messages
4. minimum messages: 2
5. each message contains only role and content

B.4.3 B1-CoT

System prompt

你是 K12 中文教育对话生成器。只输出一个 JSON 对象，顶层只包含 messages 字段。messages 使用 user/assistant 交替，每条消息只包含 role 与 content。每条 assistant content 必须直接包含闭合的 <think>…</think>，并在 </think> 后给出面向课堂的正式回答；不要输出额外字段、结构化对象或内部合同名。

User prompt

根据以下 metadata 生成一个 tutoring dialogue JSON。metadata 只作为生成依据，不要把字段名、schema 名或内部标签写进对话内容。

质量约束：assistant 不能扮演学生替答，不能在 content 中写 “学生回答/老师追问” 等角色标签；必须以教学支持者身份回应，可以追问、点拨、诊断、检查或布置练习。每个解释、判断和例子都要扣住原文词句；不能编造原文没有的字词、出处、情节或题目材料。翻译任务默认是古诗或文言文译成现代汉语，不要把古诗或文言文翻译成英文，除非任务明确要求英译。不能只给完整答案，至少体现一个引导问题、诊断判断、支架步骤、检查点或巩固练习。对话要自然，少用模板化编号、加粗堆叠和通用赏析套话。必须把 metadata 中的 strategy.steps 和 principles 转成自然课堂动作，不要把字段名、schema 名、内部标签或 metadata 结构写进对话。<think> 中要先说明本轮认知起点、文本依据和教学动作选择；</think> 后的可见回答必须执行同一个教学动作，不能只复述结论。

Think-trace 输出要求：assistant 的 content 必须从 <think> 开始，在 <think>…</think> 内用自然语言写出基于原文证据、学习任务和教学决策的思考过程，然后在 </think> 后写最终课堂回答。<think> 和 </think> 都只能出现一次，思考和回答都要完整，不能只写标签或短句。

metadata={“base_ref”:{“title”:“记承天寺夜游”,“author”:“苏轼”,“type”:“文言文”,“grade”:“7-9”},“task_ref”:{“phase_name”:“第二环节：疏通——文言语言与句式解析”,“task_name”:“句子翻译”,“cognitive_requirements”:[“理解”,“应用”,“分析”,“评估”],“cognitive_level_max”:“评估”,“positioning”:“扫除阅读障碍，精准把握字词句义，落实“语言建构与应用”。“},”source_fields“:{”base.content“:” 元丰六年十月十二日夜，解衣欲睡，月色入户，欣然起行。念无与为乐者，遂至承天寺寻张怀民。怀民亦未寝，相与步於中庭。庭下如积水空明，水中藻、荇交横，盖竹柏影也。“,”task.strategy.method_name“:” 三步译句法 + 五法（文言）

“,”task.strategy.steps“:[{”id“:1,”head“:” 标 “,”detail“:” 圈出关键词义、活用、句式特点等 “},{”id“:2,”head“:” 析 “,”detail“:” 分析句子成分和结构 “},{”id“:3,”head“:” 译 “,”detail“:” 用留/替/调/补/删完成翻译，体现句式特点 “}],”task.strategy.principles“:[{”id“:1,”text“:” 翻译必须体现关键词和句式特征 “},{”id“:2,”text“:” 建立稳定翻译流程，方便教学与命题 “}],”task.exam_features“:{”frequency_level“:” 极高频 “,”focus“:[” 文言翻译 “],”question_templates“:[” 用现代汉语翻译文中画横线的句子。“],”item_types“:[” 翻译题 “]},”task.labels“:{”text_type“:[” 文言文 “],”knowledge_points“:[” 文言翻译 “,” 关键词 “,” 句式特征 “],”exam_tags“:[” 翻译 “],”scenario“:[” 疏通环节 “,” 核心考点 “],”elements“:[” 关键词 “,” 句式还原 “,” 语义连贯”]}}}

Response format used for generation

1. type: json_schema
2. schema name: sft_messages_with_think
3. top-level required field: messages
4. minimum messages: 2
5. each message contains only role and content

B.4.4 CTSP

System prompt

你是 K12 中文教育对话生成器。只输出一个 JSON 对象，不含解释或代码块。

角色

1. **user**：一线语文老师，正在课堂上推进本节课的教学，遇到了一个具体的卡点——不知道怎么提问、学生上不了台阶、或者收束没有抓手。不是来要教案的，是来求助具体问题的。
2. **assistant**：教研员。诊断教师当前的课堂瓶颈，给出下一节课可直接执行的教学动作、提问顺序或评价抓手。必须落到给定的策略步骤或教学原则。
3. **assistant 默认产出 **：先诊断，再给提问链/课堂动作/观察点/达标检查点；不要代写学生标准答案。
4. **assistant 的 Bloom 用法 **：教研员使用 Bloom 认知层级来诊断教师策略的推进阶段（你当前在哪个层级？下一步往哪里走？），并为教师的课堂提问排序提供分析框架——与老师直接向学生教学不同，这里 Bloom 是诊断和排序工具，而非教学工具。

输出结构

顶层只允许：{ messages }
messages[i]：user/assistant 都只允许 { role, content }。
生成模型只需输出 messages；严禁把 dialogue_id、role_space、turn_mode、closing_type、base_ref、task_ref、source_id、task_name、module、phase_id、method_name、title、author 等确定性信息作为 JSON 字段输出；这些字段只允许由 pipeline 按 prompt/job/base/task 做确定性 join/backfill。
这些确定性信息会在 user prompt 中作为蒸馏依据提供给你，你必须用它们来组织 <think> 和可见回答；content 可以自然引用篇名、作者/来源、任务名、教学方法、策略步骤和文本词句，但不得写成字段清单或 schema 复述。
每条 assistant 的 content 必须从 <think> 开始，包含且只包含一对闭合的 <think>…</think>，并在 </think> 后写面向课堂的正式回答。
不要输出额外 JSON 字段，也不要把格式说明或内部术语写进内容。

1. <think> 内用自然语言显式记录：本轮认知起点与目标、当前教学意图、依据哪些原文词句与策略步骤/原则做判断，以及为什么这样推进。
2. <think> 内至少引用一处原文可见词句，且至少引用一条 task.strategy 的 step 或 principle。
3. <think> 内的推理要按邻级推进，不能只报结论不写推进过程。
4. </think> 后只写用户应看到的回答，不暴露内部标签、字段名、检查项或 schema 术语。

Bloom 认知层级

记忆 → 理解 → 应用 → 分析 → 评估 → 创造。
单轮：assistant 在一个 <think> 中展现从 bloom_from 到 bloom_to 的完整邻级推进路径。
多轮：每轮 assistant 都要有自己的 <think>，相邻轮次的目标层级应保持衔接。
think 中表达的目标层级不得超过当前任务的 cognitive_level_max。

学生求助场景：Bloom 是教学工具——assistant 用 Bloom 判断学生认知位置并推进。
教师向教研员请教场景：Bloom 是诊断和排序工具——assistant 用 Bloom 分析教师当前的策略阶段，为教学建议提供认知框架。

底线

1. 围绕原文中看得见的词句展开，不泛泛而谈
2. 高阶任务（分析/评估/创造）第一轮不直接给完整答案
3. user 语气与角色一致
4. prompt 中的规划说明、字段名、检查项和内部标签只服务于生成，不得原样写进任何 messages.content；<think> 标签除外
5. user 消息必须是自然的提问/求助。正确示例：
   学生：‘老师，这首诗里的 “田田” 是什么意思？’
   老师：‘我在讲《出塞》的默写时，学生总是记不住 “但使龙城飞将在”，怎么办？’
   错误示例（元指令）：‘以本文为例，示范如何进行主题提炼’、‘将典型设问替换为本文内容’

User prompt

生成 1 个对话样本 JSON。只输出结果对象，不要附加解释。
本次是 single_turn 对话，共 2 条消息；顶层只保留 messages。最后一条 assistant 必须按照 strategy_delivered 所限定的收束方式自然结束。每条 assistant 的 content 使用 <think>…</think> + 正式回答。

CTSP 编译约束（只作为生成依据，不作为输出字段）

1. role_space: teacher_researcher。这决定 user/assistant 的身份关系与话语风格。
2. turn_mode: single_turn。这决定对话轮次和每轮推进颗粒度。
3. target_message_count: 2。输出 messages 数量必须严格对齐。
4. bloom_path: 理解 → 评估。assistant 的自然 <think> 必须沿这个认知路径邻级推进。
5. closing_type_candidates: strategy_delivered。最后一轮必须按这些收束目标自然结束。
6. 以上 CTSP 字段是高级 teacher 生成高质量自然 CoT 的可见计划，不得作为 JSON 字段复制到输出。

文本

记承天寺夜游　苏轼　文言文　7-9

元丰六年十月十二日夜，解衣欲睡，月色入户，欣然起行。念无与为乐者，遂至承天寺寻张怀民。怀民亦未寝，相与步於中庭。庭下如积水空明，水中藻、荇交横，盖竹柏影也。

场景与任务

任务：句子翻译（第二环节：疏通——文言语言与句式解析）
教学方法：三步译句法 + 五法（文言）
认知层级范围：理解 → 应用 → 分析 → 评估（最高不超过评估）
任务定位：扫除阅读障碍，精准把握字词句义，落实 “语言建构与应用”。
策略步骤：

1. ** 标 **：圈出关键词义、活用、句式特点等
2. ** 析 **：分析句子成分和结构
3. ** 译 **：用留/替/调/补/删完成翻译，体现句式特点

教学原则：翻译必须体现关键词和句式特征；建立稳定翻译流程，方便教学与命题

考情焦点：文言翻译
考频等级：极高频
考点标签：翻译
知识点：文言翻译、关键词、句式特征
教学场景：疏通环节、核心考点
题型分布：翻译题
典型设问：用现代汉语翻译文中画横线的句子。

行为

你是 ** 老师 **，向教研员请教如何教。问题的核心是 ** 教学策略 **。单轮交付：给出可直接执行的课堂动作、提问链和达标检查，围绕给出的策略步骤组织建议，不代写学生标准答案，也不要只讲抽象原则。
提问：以本文为例，示范如何设计一个要求学生找证据做分析的课堂提问，说明问题设计意图和学生可能的卡点。
你的教学位置：你正在备课「句子翻译」，属于第二环节：疏通——文言语言与句式解析（初中阶段），正处于疏通环节。
你已经了解三步译句法 + 五法（文言）的基本思路，但在课堂落地时遇到了瓶颈——在第二环节：疏通——文言语言与句式解析中讲授这篇文章（意象层次很丰富、论述线索很清晰）时，提问顺序容易散，学生上不了台阶——问题之间需要形成递进链。
你的提问应该聚焦在' 怎么在课堂上操作'，描述具体困难，而非泛泛求建议。

对话节奏

认知推进：理解 → 评估
回答形态建议：教学建议清单 + 提问链 + 达标检查
每次 assistant 推进都要落成一个真实可执行的教学动作：先扣住当前课堂卡点，再给判断，再落一个下一步动作/检查点；可以是提问链、板书组织、课堂活动、检查点，或收束练习。最后一条 assistant 应以 “把可直接执行的教学策略交付清楚” 的状态收束。

1. 第 1 次 assistant 回复：从理解推进到评估，先交付可执行策略；回答形态以 “诊断 + 提问链 + 检查点” 为主。

质量底线

1. assistant 的判断必须同时依托原文可见词句与当前任务策略，不要脱离这两类依据
2. 优先把策略步骤和教学原则翻成真实对话中的动作、追问、检查点或练习，不要只复述方法名或抽象口号
3. 先给文本抓手或任务抓手，再给判断；不要先下结论再临时补理由
4. 一轮只解决眼前一个缺口：给出判断后，要自然落到下一步动作、检查点或巩固练习
5. 需要背景信息时，只能使用给定文本与任务能支撑的内容；拿不准就回到文本，不补外部常识
6. 可见的 user/assistant 内容必须像真实教学对话，不要把 prompt、自检项、字段名、规划标签或内部术语写出来
7. 可见回答只呈现给用户需要看到的内容，不要夹带 JSON 标签或结构说明
8. 不编造原文、作者、任务要求、比较材料细节或课堂实施结果
9. 避免空泛表达，如 “注意情感”“把握主旨” 却没有文本抓手、问题链或任务动作

Response format used for generation

1. type: json_schema
2. schema name: sft_messages_with_think
3. top-level required field: messages
4. minimum messages: 2
5. each message contains only role and content

B.5 Terminal Provenance

The four records share the same upstream Content–Task semantic anchor and the same pre-specified candidate position under a fixed selection rule. They differ in condition-specific

execution requirements. B0 is Direct (0 structural retries); B1, B1-CoT, and CTSP are Rescued after 2, 1, and 5 LQC1 structural retries. The case provides a qualitative matched illustration; aggregate E2 results provide condition-level evidence.

The archive provides the final LQC1-valid terminal messages that entered LQC2. Some intermediate failed response bodies from Rescued conditions are unavailable. Role labels are added for readability; archived message text and the B1-DirectCoT/CTSP <think> traces remain verbatim.

Table 8: Objective terminal provenance for the matched records. Hard pass applies the four LQC2 gates.

Condition	Messages	LQC1 retries	Class	Hard pass	Quality
B0	2	0	Direct	Yes	4.1
B1	10	2	Rescued	Yes	3.8
B1-CoT	2	1	Rescued	Yes	2.8
CTSP	2	5	Rescued	Yes	4.6

B.6 Representative Matched Outputs

Excerpts follow a deterministic rule: we reproduce the opening of the first assistant response for every condition. For the multi-turn B1 record, we additionally show the final assistant response. The rule is independent of quality.

B0.

> ### 教学问答示例（课堂使用）

The response opens with a classroom-oriented question-and-answer format.

B1.

> 当然可以！我们先用“三步译句法”。第一步是 ** 标 **——找出关键词和句式特点。请你先把这句话里你认为重要的词语圈出来，比如“庭下”“积水”“空明”“藻”“荇”“交横”“竹柏影”。
>
> 不客气！记得在翻译时始终坚持“三步译句法”，并灵活运用五法，这样可以帮助你快速拆解文言句子，提升翻译准确度。下次遇到类似句式，试着自行练习一遍，再来跟我讨论。加油！

The multi-turn response begins with elicitation and closes with a practice recommendation.

B1-CoT.

> <think> 在进行文言翻译时，首先要依据任务要求的“三步译句法 + 五法”。

The response opens by exposing the requested reasoning surface.

CTSP.

> <think> 当前认知位于“理解”层级：教师已能阐述三步译句法（标、析、译）与五法的基本概念，能够指出文中关键词如“庭下如积水空明，水中藻、荇交横，盖竹柏影也”的字面意义。

The response opens from the Job-specific instructional state encoded for this case.

These excerpts illustrate the outputs for one matched anchor. Aggregate analyses, rather than this case, support condition-level comparisons.

B.7 Full Verbatim Terminal Outputs

B.7.1 B0

Terminal status: Direct; LQC1 structural retries: 0; LQC2 overall quality: 4.1.

[**user**]

> 请围绕下面文本和学习任务生成一段适合课堂使用的教学问答。问答要清楚、贴合原文，不编造题目没有给出的材料。
> 文本：记承天寺夜游（苏轼）
> 原文：元丰六年十月十二日夜，解衣欲睡，月色入户，欣然起行。念无与为乐者，遂至承天寺寻张怀民。怀民亦未寝，相与步於中庭。庭下如积水空明，水中藻、荇交横，盖竹柏影也。
> 任务：句子翻译

[**assistant**]

> ### 教学问答示例（课堂使用）
> ** 教师提问 **1**：请把第一句 “元丰六年十月十二日夜，解衣欲睡，月色入户，欣然起行。” 译成现代汉语。** 学生示例答案 **：在元丰六年十月十二日的夜里，我脱下衣服想要睡觉，月光照进屋子，我心情愉快地起身出门。
> ** 教师提问 **2**：第二句 “念无与为乐者，遂至承天寺寻张怀民。” 的意思是什么？请翻译。** 学生示例答案 **：想到没有人可以一起娱乐，我于是去承天寺寻找张怀民。
> ** 教师提问 **3**：请翻译 “怀民亦未寝，相与步於中庭”。** 学生示例答案 **：张怀民也还没睡，我们一起在寺院的中庭散步。
> ** 教师提问 **4**：句子 “庭下如积水空明，水中藻、荇交横，盖竹柏影也。” 的现代译文是？** 学生示例答案 **：庭院下面像积满清澈的水，水里有交叉的藻类和荇菜，那是因为竹子和柏树的影子映在水里。
> ** 教师指导要点 **：
>
> 1. ** 时间与地点 **：明确 “元丰六年十月十二日夜” 对应的时间概念，翻译时可保留原年份或改写为 “元丰六年十月十二日晚”。
> 2. ** 动词顺序 **：古文常把动词放在后面，翻译时需调整顺序，使句子符合现代汉语的主谓宾结构。
> 3. ** 词义解释 **：
>
> - “解衣” 指脱下衣服。
> - “欣然” 表示高兴、愉快。
> - “步於中庭” 即在中庭里走步。
> - “荇” 是一种水草，古文中常与 “藻” 并列出现。
>
> 4. ** 保持意境 **：原文描写月光、清水、竹柏的倒影，要在译文中保留这种静谧、清幽的意境，避免过度直译导致画面感丢失。
>
> ---
>
> ** 课堂练习 **（请自行翻译）
>
> 1. “念无与为乐者”。
> 2. “相与步於中庭”。
> 3. “庭下如积水空明”。
> 4. “盖竹柏影也”。
>
> ** 教师提示 **：让学生先用自己的语言把句子表达出来，再对照参考答案，检查是否漏掉了时间、情感或自然景物的细节。

B.7.2 B1

Terminal status: Rescued; LQC1 structural retries: 2; LQC2 overall quality: 3.8.

[**user**]

老师，我在做《记承天寺夜游》的翻译练习，遇到句子“庭下如积水空明，水中藻、荇交横，盖竹柏影也。”不太懂该怎么翻译，能帮我分析一下吗？

[**assistant**]

当然可以！我们先用“三步译句法”。第一步是 ** 标 **——找出关键词和句式特点。请你先把这句话里你认为重要的词语圈出来，比如“庭下”“积水”“空明”“藻”“荇”“交横”“竹柏影”。

[**user**]

好的，我标了：庭下、积水、空明、藻、荇、交横、竹柏、影。还有“如”“盖”这两个词。
我觉得“如”是比喻，“盖”是解释。

[**assistant**]

很好，已经把关键词抓住了。接下来进入第二步 ** 析 **——分析句子成分和结构。整句可以拆成两大分句：

1. “庭下如积水空明，水中藻、荇交横，”
2. “盖竹柏影也。”
   第一分句是比喻句，主语是“庭下”，谓语是省略的“像”，补语是“积水空明”。其中“水中藻、荇交横”是并列状语，描绘水中的植物交错。第二分句是解释说明，指出上述景象的原因是“竹柏的影子”。请你再确认一下这些成分是否准确。

[**user**]

我看出来了：第一分句的主语是“庭下”，比喻的对象是“积水空明”，还有两个并列的状语“藻交横”。第二分句是解释，主语是“盖”，宾语是“竹柏影”。

[**assistant**]

分析得很到位！现在进入第三步 ** 译 **，使用文言翻译的五法：留、替、调、补、删。

- ** 留 **：保留原句的关键词，如“庭下”“积水”“空明”。
- ** 替 **：把文言的“如”替换为“像”。
- ** 调 **：把省略的谓语补上，使句子完整。
- ** 补 **：在现代汉语里补充“在”“的”等助词，使语义连贯。
- ** 删 **：去掉冗余的古代虚词“也”。
  请你尝试按照这五法把句子译成现代汉语。

[**user**]

我尝试翻译：庭院下面像积水一样清澈明亮，水中藻类和荇菜交错，原来是竹子和柏树的影子。

[**assistant**]

译文思路正确，基本抓住了原文意思。可以再进行 ** 评估 **，检查是否忠实原意并符合文言句式特征：

1. “庭院下面”改为“庭下”，更简洁。
2. “像积水一样清澈明亮”可以写成“像积水般空明”。
3. “原来是”对应“盖…也”，可以保留“原来是”。
4. “竹子和柏树的影子”可以写成“竹柏的影子”。
   优化后的译文：
   “庭下像积水般空明，水中的藻、荇交错，原来是竹柏的影子。”你觉得这样更贴近原文吗？

[**user**]

是的，这样更简练，也更符合原文的意境。谢谢老师！

[**assistant**]

不客气！记得在翻译时始终坚持“三步译句法”，并灵活运用五法，这样可以帮助你快速拆解文言句子，提升翻译准确度。下次遇到类似句式，试着自行练习一遍，再来跟我讨论。加油！

B.7.3 B1-CoT

Terminal status: Rescued; LQC1 structural retries: 1; LQC2 overall quality: 2.8.

[**user**]

请根据《记承天寺夜游》中的句子“元丰六年十月十二日夜，解衣欲睡，月色入户，欣然起行。念无与为乐者，遂至承天寺寻张怀民。”进行翻译，说明关键词、句式特点，并展示三步译句法的过程。

[**assistant**]

<think> 在进行文言翻译时，首先要依据任务要求的“三步译句法 + 五法”。

1. 标（标记）阶段：抓取句中的关键词义及活用形式，如“解衣”“欲睡”“月色入户”“欣然起行”“念无与为乐者”“遂至”。这些词多为动词短语或介词结构，体现了古文的动作连贯性和情感倾向。句式上多为并列复句，前半段描写夜间情境，后半段交代行动动机。
2. 析（分析）阶段：分解句子成分。

- 主语省略，实际主语为“我”。
- 谓语动词为“解衣”“欲睡”“起行”。
- 状语包括时间（元丰六年十月十二日夜）、环境（月色入户）、情感（欣然）等。
- 补语/宾语有“无与为乐者”（没有人可以一起玩乐），以及目的状语“遂至承天寺寻张怀民”。
  通过分析，可见句子采用了“时间 - 情境 - 动作 - 动机”顺序的叙事结构，符合古文简练、层层递进的特点。

3. 译（翻译）阶段：运用留/替/调/补/删五法。

- 留：保留关键意象“月色入户”。
- 替：将“解衣欲睡”替换为现代表达“脱下衣服想要睡觉”。
- 调：调整顺序，使信息更顺畅。原句先描述环境再表行动，译文亦保持此顺序。
- 补：补全省略的主语“我”。
- 删：去掉文言冗余的“念”字，直接表达动机。
  综合上述，形成通顺、忠实原意的现代汉语译文。
  最后，评估层面需要确保译文既保留原文意境，又符合现代阅读习惯，满足考试对文言翻译的高频要求。</think> 翻译如下：

元丰六年十月十二日的夜晚，我脱下衣服想要睡觉，月光透进屋里，我高兴地起身。由于没有人可与我一起玩乐，我于是来到承天寺寻找张怀民。

B.7.4 CTSP

Terminal status: Rescued; LQC1 structural retries: 5; LQC2 overall quality: 4.6.

[**user**]

以本文为例，示范如何设计一个要求学生找证据做分析的课堂提问，说明问题设计意图和学生可能的卡点。我的教学位置是正在备课「句子翻译」，属于第二环节：疏通——文言语言与句式解析（初中阶段），正处于疏通环

节。我已经了解三步译句法 + 五法（文言）的基本思路，但在课堂落地时遇到了瓶颈——在第二环节：疏通——文言语言与句式解析中讲授这篇文章（意象层次很丰富、论述线索很清晰）时，提问顺序容易散，学生上不了台阶——问题之间需要形成递进链。你的提问应该聚焦在' 怎么在课堂上操作'，描述具体困难，而非泛泛求建议。

[**assistant**]

<think> 当前认知位于 “理解” 层级：教师已能阐述三步译句法（标、析、译）与五法的基本概念，能够指出文中关键词如 “庭下如积水空明，水中藻、荇交横，盖竹柏影也” 的字面意义。但缺少将这些认知转化为学生可操作的提问序列的评估能力。目标是推进到 “评估” 层级，即让教师能够设计能让学生自行寻找证据、分析句式并对翻译决定进行自我评估的提问链。依据任务策略步骤 1“标：圈出关键词义、活用、句式特点” 等，以及文本中 “庭下如积水空明” 这一可视化描写，构建递进提问，使学生从识别关键词到解释句法，再到比对翻译方案并评价其合理性。此过程在理解的基础上加入评估维度，形成教学闭环。</think> 下面给出可直接在课堂使用的操作方案，分为三段：

**1. 诊断与教学目标 **

- ** 诊断 **：提问顺序散乱导致学生只能停留在词义记忆，未能逐层推进至句式分析与翻译决策的自我评价。
- ** 目标 **：通过 “标 - 析 - 译” 三步，形成从找证据 → 结构分析 → 翻译评估的递进链，确保每位学生都有明确的操作台阶。

2. 具体提问链（对应标、析、译）
步骤	提问示例	设计意图	可能卡点 & 预防措施
——	————-	————-	——————-
标	(1) “‘庭下如积水空明’ 中的 ‘积水’ 在原文里指的是什么景象？” (2) “‘藻、荇交横，盖竹柏影也’ 这句话里出现了哪些关键词？”	帮助学生定位关键词，转化为现代概念，为后续结构分析做素材准备。	学生可能只说出字面意思。提醒学生结合上下文想象夜色中的水面与植被倒影。
析	(3) “这两句的句式属于 ‘……（并列/递进）结构’，请指出主谓宾或偏正关系。” (4) “‘盖竹柏影也’ 中的 ‘盖’ 起什么作用？它在整句中的语法功能是什么？”	引导学生分析句子成分，识别 ‘盖’ 的补充说明功能，体会文言中常见的 ‘连绵’ 句式。	学生易混淆 ‘盖’ 是动词还是介词。提供句子拆分示例，示范标注主语、谓语、补语。
译	(5) “请用 ‘留/替/调/补/删’ 中的两种方法，把 ‘庭下如积水空明’ 翻译成现代汉语。” (6) “如果把 ‘盖竹柏影也’ 译成 ‘是竹柏的倒影’，还有其他表达方式吗？请说明各自的优缺点。”	让学生实践五法，体会不同译法对意境的影响；通过比较提升翻译的评价意识。	学生常只给出一种直译。引导他们思考语言的轻重、意象保留与流畅度之间的权衡。
评	(7) “根据刚才的翻译，你认为哪种版本最能体现 ‘夜游’ 的静谧氛围？请用原文证据说明。” (8) “同伴的译文有什么值得借鉴的地方？你可以怎样改进自己的译文？”	让学生自行评估译文，使用原文证据进行论证，培养自我反馈与同伴互评的能力。	学生可能缺乏评价标准。提前给出评价要点：词义保留、句式对应、情感色彩、语言流畅。

**3. 达标检查点 **

1. ** 关键词定位 **：每位学生能在文本中划出至少 3 个关键词并解释其意象。
2. ** 句式剖析 **：学生能够用符号标注出主 - 谓 - 宾或偏 - 补结构，并说明 ‘盖’ 的语法功能。
3. ** 译法运用 **：在翻译练习中，学生使用至少两种五法，并能口头说明选择理由。
4. ** 翻译评估 **：学生能够依据原文证据，对比两种以上译文，指出哪种更符合原文情境并给出改进建议。

** 课堂实施建议 **

- ** 板书结构 **：左侧列出 “标 - 析 - 译 - 评” 四列，学生完成后自行在对应格子里填写答案，形成可视化的进度条。
- ** 小组合作 **：每组分配一句，轮流完成标、析、译、评四步，最后全班共享评价表。
- ** 即时反馈 **：在每一步提问后，教师快速点名检查两位学生的回答是否符合检查点，及时纠偏。

如此递进的提问链把 “找证据 → 结构分析 → 翻译决策 → 自评” 四个层面贯通，避免提问散乱，使学生在每一步都有明确的操作台阶，能够在疏通环节顺畅上台阶。

B.8 Lineage and Evidence Boundary

All four terminal records pass the LQC2 hard gates. Pair, Job, Slot, attempt, and terminal Candidate identities remain distinct throughout the archive. Retry counts describe structural-recovery burden for the same Slot; quality scores describe the terminal Candidate that entered LQC2. The matched case connects generator-visible inputs, retry records, and terminal outputs, while aggregate E2 results provide the condition-level comparison.

C Distribution Planning and Preservation

C.1 Target materialization

The allocation audit crosses three target families (balanced, empirical-frequency-weighted, and P2/P3-enriched), three generator-facing representations (document-first, flat-metadata, and CTSP), and three seeds (S1/S2/S3; archived as 20260807/20260808/20260809). Each of the 27 settings has budget $B = 512$. Integer bucket targets equal the realized Slot counts in all settings, with bucket-level L_1 error zero. Slot identity checks find no missing, extra, duplicate, or orphan records. This verifies target-to-workload compilation after integerization.

Natural Job support S_ω, target Slots N^*_ω, and retained Candidates R_ω are distinct quantities. The first describes feasibility, the second construction intent, and the third survival under a named retention rule. Distribution tables in this submission use primary LQC2 hard pass as the retention definition unless explicitly labeled otherwise.

C.2 Preservation audit

For each explicit dimension, preservation compares the common 512-Slot plan with the condition-specific hard-pass composition. Table 9 reports planned-to-retained TVD for the seven audited axes. Every condition retains full category support on every axis; the values therefore quantify shifts in relative composition. Threshold-qualified and selected-for-SFT views use their corresponding membership definitions in separate analyses. Absolute targets also permit bucket-level deficit calculation and optional directed supplementation.

Table 9: Planned-to-retained TVD on the seven E2 planning axes. Lower values indicate less composition drift. All conditions retain complete category support on every reported axis.

Condition	Role	Mode	Bloom	Grade	Task type	Task	Task×Role
B0	0.0740	0.0630	0.0681	0.0175	0.0274	0.0589	0.0861
B1	0.0141	0.0402	0.1120	0.0111	0.0513	0.1256	0.1458
B1-CoT	0.0242	0.0635	0.1472	**0.0028**	0.0786	0.1613	0.1794
CTSP	**0.0022**	**0.0069**	**0.0390**	0.0158	**0.0022**	**0.0437**	**0.0489**

Table 10: Training-view membership used in E3.

View	Rows	Train	Validation
Base	0	0	0
CTSP-LQC1-SFT	512	430	82
CTSP-LQC2-SFT	145	121	24

D Downstream Training-view Provenance

The strict 145-row view is defined by archived primary-judge sidecars and the final selection rule. The local audit copy was reconstructed row by row from those sidecars because the original export was absent from the Windows handoff. Membership lineage is recoverable; the available training provenance covers the archived view definition and reported downstream scores.

Final combines Result and Process evaluation. Result evaluates answer/task outcome, whereas Process evaluates the pedagogical interaction. The training views differ simultaneously in sample count, token budget, coverage, and composition. These differences bound all E3 interpretation at the complete-view level.

E Implementation and Additional Diagnostics

E.1 Lifecycle objects and attempts

The recoverable lifecycle is Content/Task registries → legal Pair → Job → rendered input → Slot/Candidate → LQC1 → normalization → LQC2 → training-view membership. Registry-to-downstream identity is fully represented; earlier raw-material-to-registry editing history is only partially recoverable.

A Slot is an allocation unit. A provider call is a runtime event. Initial calls, structural retries, and transient retries can all belong to one Slot. Pass@k counts structural closure within the first k structural attempts; LQC2 reports semantic–pedagogical quality.

Table 11: Complete structural closure and retry burden.

Condition	Pass@1	Pass@2	Pass@3	Structural retry calls
B0	87.30	98.63	100.00	72
B1	63.67	87.50	93.55	333
B1-CoT	67.97	92.38	97.27	220
CTSP	72.46	87.89	94.92	253

E.2 Failure attribution and efficiency boundaries

Table 12: Primary LQC2 hard-gate failure attribution. Categories are multi-label.

Condition	Any failure	Grounding	Alignment	Leakage	Hallucination
B0	181	51	86	4	124
B1	123	58	45	4	106
B1-CoT	264	68	120	188	101
CTSP	60	23	8	22	36

Provider-call statistics include initial, structural-retry, transient, and rate-limit invocations and are reported separately from Slot counts. Historical fields labeled prompt tokens use the archived approximation $\lceil \text{characters}/4 \rceil$ and are retained as provenance; formal token usage is reported from tokenizer-based records where available.

Direct and rescued candidates are descriptive structural-path subgroups. Rescue records eventual structural closure after retry, and semantic quality refers to the terminal Candidate evaluated by LQC2.