Measuring Optimal Challenge: Trajectory-Based Difficulty Alignment in Open-Ended Language Tutoring

Ziqi Shu
Stanford University
ziqishu@stanford.edu

Shuman Wang
Stanford University
shuman@stanford.edu

Michael Hardy
Stanford University
hardym@stanford.edu

Abstract

Conversational English as a Foreign Language (EFL) tutoring relies on dynamically generated exercises rather than fixed item banks, so traditional difficulty estimation cannot verify whether a task is appropriately calibrated to a learner. We propose a framework that measures difficulty alignment directly from observable interactional behavior, classifying each exercise into one of three states (Under-Challenged, Optimally Challenged, or Over-Challenged) based on turn-level sequences of student attempts, errors, confusion, and tutor scaffolding. Using 1,566 exercises from the Teacher-Student Chatroom Corpus, we validate the classification against human annotation (Cohen's $\kappa = 0.79$ at the state level) and show that a learner's cumulative trajectory of these states predicts success on subsequent exercises. Aggregating these predictions into a within-session capability-shift proxy, we find that sessions with higher proportions of over-challenging exercises systematically yield lower estimated shifts, while optimally challenging interactions are significantly associated with greater improvement than under-challenging ones—patterns consistent with Krashen's Input Hypothesis.

1 Introduction

How can we determine computationally whether a tutoring exercise is appropriately calibrated to an individual learner? In conversational English as a Foreign Language (EFL) tutoring, exercises are not drawn from a fixed item bank but are selected or generated dynamically or organically in response to evolving learner in-context needs. Whether delivered by a tutor drawing on experience or by an LLM system generating exercises on the fly, we know of no computationally tractable method for accurately evaluating whether the tutoring content is at an optimal level of productive engagement for a learner, leaving us without a measure of instructional calibration.

Multiple traditions in learning theory converge on optimal productive engagement: from desirable difficulties (Bjork and Bjork, 2011) and productive failure (Kapur, 2008) to Vygotsky's Zone of Proximal Development (Vygotsky, 1978). In second language acquisition, Krashen's Input Hypothesis (Krashen, 1977) formalizes this as $i + 1$: optimal acquisition occurs when input slightly exceeds the learner's current mastery level, i. The Interaction Hypothesis (Long, 1996) further predicts that scaffolding moves during task completion facilitate learning, highlighting the role of the tutor. These frameworks yield an operationalizable expectation : appropriately calibrated exercises should be observable in the learner's interactional behavior in conversation with their tutor.

Existing computational approaches to difficulty estimation, however, are not designed to operate on dynamically generated exercises in open-ended conversational settings. Item Response Theory (Hambleton et al., 1991) estimates item difficulty and learner ability from response patterns over a calibrated item bank, supporting adaptive assessment rather than instructional calibration in dynamic tutoring. LLM-based difficulty estimation (Brant et al., 2026; Liu et al., 2025) sidesteps the item-bank requirement but infers difficulty from the model's own capabilities rather than from learner-specific evidence. Knowledge tracing (Corbett and Anderson, 1994; Piech et al., 2015), including recent LLM-based variants (Neshaei et al., 2024), addresses a related but distinct problem: estimating a learner's mastery of a fixed taxonomy of skills from their history of correct and incorrect responses, in order to guide which skills to practice next. It is well suited to domains where competencies decompose into discrete, repeatedly practiced units, but open-ended conversational practice offers neither a fixed skill taxonomy nor sufficient per-skill repetition, and even where it could be applied, it would tell us what a learner knows rather than whether a

Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications, pages 651–667

given exercise was appropriately challenging at the moment of delivery. Crucially, none of these approaches leverage the rich behavioral signal available in the interaction itself—treating difficulty as either an intrinsic property of the item or an aggregate of historical response patterns, rather than as something that manifests in real time through the learner's interactional behavior.

We argue that the interaction surrounding an exercise—the sequence of student attempts, errors, confusion signals, and tutor scaffolding moves—provides a theoretically grounded and reproducible operationalization of Krashen's optimal challenge. The core insight is that while "difficulty" and "productive struggle" are latent constructs without direct ground truth, their constituent behaviors are concrete and observable: a student producing an incorrect response, asking what a word means, or requesting clarification are all explicit, annotatable events. By mapping these turn-level behavioral sequences to optimal challenge alignment states—*Optimally Challenged* ($i + 1$; struggle followed by success), *Under-Challenged* (i; correct without prior struggle), and *Over-Challenged* ($i + \mathbb{Z}_{>1}$ unresolved difficulty)—we obtain a computational measure grounded directly in the observable indicators that the Input and Interaction Hypotheses identify as markers of productive learning.

However, classifying individual exercises is only the initial step; isolated difficulty states are momentary snapshots. To test the underlying learning theories, we must understand how these states accumulate to shape learning trajectories. We achieve this through a two-stage framework. At the micro-level, we examine whether a learner's cumulative trajectory of difficulty states, combined with performance and scaffolding features, predicts success on subsequent exercises. At the macro-level, we leverage this predictive model to construct an in-situ *capability-shift proxy*, the shift in a student's expected success probability across a session, and test whether sessions with higher proportions of optimally challenged interactions yield greater estimated improvement. Together, this framework motivates three research questions:

- **RQ1 (Annotation & Classification):** Can turn-level interactional trajectories be reliably annotated and systematically mapped to difficulty alignment states that reflect the relational difficulty of tutoring exercises?
- **RQ2 (Micro-Level Prediction):** To what extent do session-level interaction histories—including prior performance, scaffolding behavior, and difficulty alignment states—predict a learner's probability of success on subsequent exercises?
- **RQ3 (Macro-Level Inference):** How does the overall composition of difficulty alignment states within a tutoring session relate to estimated learner improvement across the session, controlling for structural confounders?

2 Data

The dataset used in this study is the **Teacher-Student Chatroom Corpus (TSCC) Version 2** (Caines et al., 2020), which contains message-based English Language tutoring sessions. The dataset includes a total of 260 sessions, involving 2 teachers and 12 students.

Each session is stored as a TSV file, where each row represents an anonymized message sent by either a student or a tutor. The dataset provides the following attributes for each message: `user.id` (unique identifier for the sender), `role` (student/teacher), and `seq.type` (the message type, where the label *exercise* indicates that the teacher is providing an exercise in the current sequence.) Additionally, the dataset includes a metadata file, which provides annotations for each session, including: Student's CEFR (Common European Framework of Reference for Languages, (Council of Europe, 2001)) level, age, and native language (See Table 1).

ID	CEFR	Age	L1	#Sess.	#Ex.
002	B1	22	Japanese	3	27
003	B2	32	Japanese	4	26
004	C1	40	Spanish	28	53
005	B2	20	Italian	6	34
006	B1	23	Thai	4	29
007	B2	22	Mandarin Chinese	32	297
009	B2	12	Ukrainian	48	518
010	B2	13	Ukrainian	14	138
011	B2	26	Russian, Ukrainian	51	217
012	C2	33	Russian	26	47
014	C2	30	Italian	26	72
015	B2	30	Mandarin Chinese	18	108

Table 1: Student Demographic Summary in TSCC v2. # Sess. indicates the number of sessions, and # Ex. indicates the number of exercises.

3 Methods

Our methodology proceeds in two analytical stages. The *micro-level* stage classifies each exercise into a difficulty alignment state (Sections 3.1–3.3) and fits a Generalized Linear Mixed-Effects Model that predicts next-exercise success from cumulative trajectory features (Section 3.4), establishing that these features carry predictive signal. The *macro-level* stage then aggregates the micro model's out-of-fold predictions into a session-level capability-shift proxy (Section 3.5) and regresses it on cumulative difficulty proportions, controlling for exercise-length changes, tutor identity, and student-level variance (Section 3.6). This two-stage design isolates the pedagogically meaningful contribution of difficulty composition from co-varying surface confounds that the micro model alone cannot disentangle.

3.1 Data Preprocessing and Exercise Identification

Starting from 2,527 sequences labeled as containing exercises in the original corpus, we applied a filtering pipeline to isolate exercises that are self-contained within the chat log, contain an explicit exercise body, and require productive English practice. Following established methodologies that validate LLMs as reliable data annotators against human-labeled subsets (Gilardi et al., 2023; Wang et al., 2023; Törnberg, 2023), we employed a hybrid annotation process: one author manually judged eligibility for a 10% random sample (n=252), then validated an automated classifier (GPT-4.1-2025-04-14; prompt documented in Appendix C.1) against these judgments. We required precision > 0.85 on the verification set before applying the classifier at scale, prioritizing precision over recall to keep the final exercise set conservative; the threshold was met (full validation metrics reported in Section 4.1). The automated pipeline was then applied to the remaining sequences, yielding a final dataset of **1,566** tutor-led exercises. Full eligibility criteria and pipeline details are provided in Appendix A.

3.2 Turn-Level Behavioral Annotation

To capture the interactional trajectory following each exercise, we developed a turn-level coding scheme (Table 2) classifying each post-exercise turn by speaker role and pedagogical function. Annotation proceeds sequentially and terminates at a stop condition (Student Correct Response or Teacher Direct Answer). We annotated a random sample of 157 exercises (10% of the filtered corpus).

To ensure robustness, we employed a two-stage validation process.

Human Inter-Rater Reliability Two authors independently annotated a pilot subset of 30 exercises (approximately 150 turns). Inter-rater agreement was calculated using Cohen's κ to account for chance agreement. Full annotation proceeded only after achieving $\kappa > 0.70$ (true $\kappa = 0.96$), indicating substantial agreement. Disagreements were resolved through discussion, particularly clarifying the distinction between *Content* and *Procedure* confusion and the three types of student response.

LLM-Human Agreement To scale annotation, we evaluated GPT-5.2 (GPT-5.2-2025-12-11, non-reasoning) on the fully manually annotated set (157 exercises, 1,570 turns). The prompt is documented in Appendix C.2. The model was prompted using the same definitions and contextual constraints described above. Following the same threshold-gate criterion used for human IRR but with a stricter cutoff, we required GPT-5.2 to achieve Cohen's $\kappa > 0.80$ ("almost perfect agreement" under the Landis and Koch scale (Landis and Koch, 1977)) against human annotations before deploying the pipeline on the full corpus; the threshold was met (full per-label results in Section 4.2).

3.3 Trajectory-to-Challenge Mapping

We map each exercise's post-exercise turn sequence (up to $L_{max} = 10$ turns) into one of three mutually exclusive difficulty alignment states. The tripartite partition follows directly from Krashen's Input Hypothesis (Krashen, 1977), which distinguishes input at i (under-challenging), at $i + 1$ (optimally challenging), and beyond $i + 1$ (over-challenging) relative to the learner's current mastery i. In the human-annotated set, 93.6% of exercises reached a stop condition within this window, confirming that the threshold captures the natural resolution boundary of most interactions. The three states are:

- **Under-Challenged (i):** The student produces a correct response without prior evidence of content-level struggle (no "Response Attempt – Incorrect" or "Content Question/Confusion" in the trajectory).

Speaker	Label	Definition
Teacher	Hint/Instruction	Scaffolding, elicitation (e.g., "Tell me more"), or clarifying instructions related to the exercise.
	Direct Answer	Explicit provision of the correct answer or full solution.
Student	Response Attempt – Correct	Grammatically and semantically complete answer.
	Response Attempt – Incorrect	Attempt containing linguistic errors or incorrect content.
	Response Attempt – Partial	Correct but incomplete response (e.g., filling only one of multiple blanks).
	Content Question/Confusion	Questions or statements reflecting gaps in linguistic knowledge (e.g., "I don't know this word").
	Procedure Question/Confusion	Questions about task mechanics rather than language content (e.g., "Should I type it?").
Any	NA	Turn unrelated to the specific exercise body (off-topic chat, meta-talk, references to other exercises).

Table 2: Turn-level tutor-student interaction taxonomy.

- **Optimally Challenged ($i+1$):** The student eventually succeeds, but the preceding trajectory contains explicit markers of struggle (an incorrect attempt or content confusion).
- **Over-Challenged ($i+\mathbb{Z}_{>1}$):** The tutor provides the answer via "Direct Answer," or the interaction reaches L_{max} without a correct student response.

Figure 1 illustrates each state with example trajectories from the corpus.

This classification operates exclusively on student-driven evidence: tutor scaffolding alone does not trigger the Optimally Challenged state, and procedural questions about task mechanics (e.g., "Should I type it?") do not count as evidence of content-level difficulty—if a student asks such a question and subsequently solves the task, the trajectory remains Under-Challenged.

3.4 Session-Level Next-Exercise Success Prediction

Having classified each exercise into a difficulty alignment state, we next ask whether a learner's cumulative session history, including their trajectory of the difficulty alignment state and interaction features, provides a predictive signal for success on subsequent exercises. We formalize this as a binary classification task: given the first t exercises in a session, predict whether the student will respond correctly to exercise $t+1$.

3.4.1 Trajectory Prefix Construction and Feature Selection

For a session containing T exercises (ordered chronologically by turn number), we construct $T-1$ prefix rows, where each row at position $t \in \{1, \ldots, T-1\}$ uses the history of exercises 1 through t to predict the outcome of exercise $t+1$.

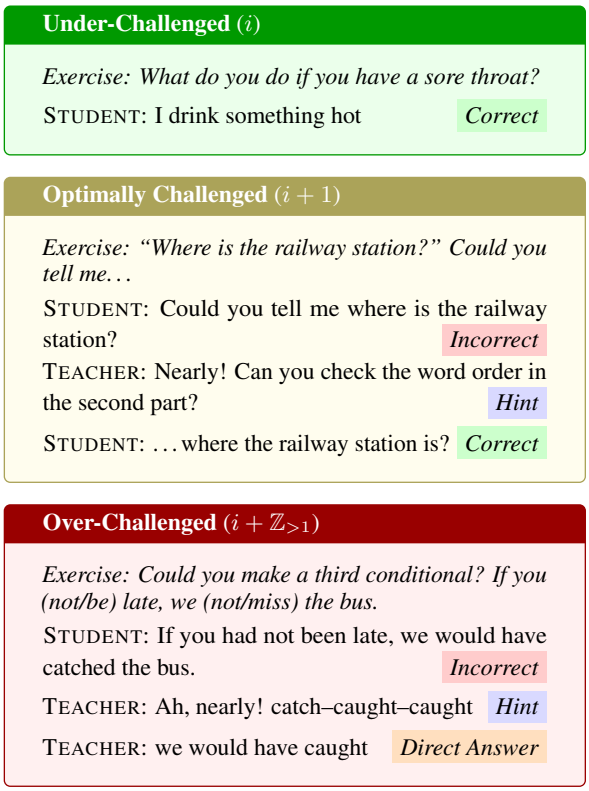

Figure 1: Illustrative trajectories for each difficulty alignment state, drawn from the TSCC corpus. Turn labels are color-coded: *Correct*, *Incorrect*, *Hint*, *Direct Answer*.

Sessions with only a single exercise are excluded. We initially extracted 24 features across four categories: historical performance, difficulty alignment trajectories, scaffolding interactions, and exercise surface properties. A comprehensive table defining all initial features is provided in Appendix D.

To prevent information leakage across related observations, we employ **session-level** 5-fold cross-validation. All prefix rows from a given session are assigned to the same fold, ensuring that no session contributes to both training and evaluation within

any fold.

To objectively identify the most robust and generalizable subset of predictors from this initial pool and to prevent overfitting, we applied L1-regularized logistic regression (Lasso) across 5-fold cross-validation. We established a strict inclusion criterion: only features selected by the Lasso penalty in at least 0.60 of the folds (i.e., ≥ 3 out of 5 folds) were retained for the final predictive model.

This data-driven selection process retained 13 features. The key retained features central to our pedagogical analysis include:

- **Cumulative Proportions of Difficulty Alignment States:** The cumulative fraction of exercises up to turn t that were classified as Over-Challenged and Under-Challenged, respectively. The proportion of Optimally-Challenged was dropped by Lasso to avoid collinearity with the other two. By definition, these three proportions represent the complete composition of the session history, summing to 1.
- **Consecutive Optimally Challenged States:** The count of uninterrupted Optimally Challenged states immediately preceding $t + 1$.
- **Consecutive Successes:** The count of consecutive correct responses immediately preceding $t + 1$.

Other retained features include scaffolding metrics (e.g., average hints per exercise), upcoming exercise character and word lengths, and the student's baseline CEFR proficiency. A complete feature dictionary and the detailed L1 selection frequencies are provided in Appendix D.

After excluding 50 sessions containing none or single exercise, this construction yielded 1,324 prediction rows from 210 sessions. The target variable (next-exercise success) has a positive rate of 70.7%, indicating moderate class imbalance toward correct responses. Prefix lengths (the number of prior exercises available as history) range from 1 to 34 (mean ≈ 5.8), reflecting substantial variation in the amount of session context available.

3.4.2 Generalized Linear Mixed-Effects Model

We model next-exercise success using a generalized linear mixed-effects model (GLMM; Breslow and Clayton, 1993) with a binomial response and a logit link function. Standard logistic regression assumes that all observations are independent; however, exercises are structurally dependent as they are nested within specific students and tutors. To account for this nested structure, we include a student-level random intercept to control for baseline differences in learner proficiency, and a fixed effect for teacher identity to account for unobserved pedagogical differences between the two tutors in the dataset.

Formally, for the student-random + teacher-fixed specification, let $y_{ijs} \in \{0, 1\}$ denote correctness for prediction instance i from student j with teacher s. We model

$$\text{logit}(P(y_{ijs} = 1)) = \mathbf{z}_{ij}^{\top}\boldsymbol{\beta} + \boldsymbol{\alpha}_s^{\top}\mathbf{1}\{\text{teach.} = s\} + u_j \quad (1)$$

where $\mathbf{z}_{ij}$ contains the selected covariates (numeric covariates are standardized within each training fold), $\boldsymbol{\alpha}_s$ denotes teacher fixed effects (with a reference category), and $u_j \sim \mathcal{N}(0, \sigma_u^2)$ is a student-level random intercept.

To ensure our findings regarding key behavioral predictors are not artifacts of a specific model structure, we also tested alternative specifications, including models without random effects and models with varying control structures. These robustness checks are reported in Appendix E, demonstrating consistent feature effects across different specifications.

3.5 In-Situ Capability-Shift Proxy

Observational tutoring corpora such as TSCC typically lack formal pre- and post-assessments, precluding direct measurement of learning outcomes. We therefore construct an *in-situ capability-shift proxy*—a within-session shift in model-estimated success probability, not a measured change in language acquisition—by leveraging the out-of-fold predicted probabilities from the next-exercise success model described in Section 3.4.

The predictive model established in Section 3.4 serves as a noisy but informative indicator of a student's real-time capability. The core intuition is that if a student's model-estimated probability of success increases over the course of a session, the model has detected a positive shift in the student's interactional profile toward patterns associated with correct responses—providing an operational proxy for *model-estimated* within-session capability shifts, rather than a direct measure of

language acquisition.

3.5.1 Definition

For each session trajectory, the next-exercise success model produces a sequence of out-of-fold predicted probabilities $\hat{p}_2, \hat{p}_3, \ldots, \hat{p}_T$, where $\hat{p}_t$ denotes the predicted probability that the student answers exercise t correctly, conditioned on the session history through exercise $t-1$. Note that no prediction is generated for the first exercise ($t = 1$), as it has no prior history.

We define the **capability shift** for a trajectory as the direct difference between the final predicted probability and the initial predicted probability within the session. Specifically, let $n = T - 1$ denote the number of predictions in a trajectory. We distinguish two cases:

- $n < 3$ (fewer than 4 exercises / 3 predictions): The trajectory is **discarded**, as it provides insufficient intermediate history to form a valid pedagogical composition. (57 trajectories discarded)
- $n \geq 3$: CS $= \hat{p}_T - \hat{p}_2$. (153 trajectories)

3.6 Associating Difficulty Alignment with Capability Shift

Having constructed the in-situ capability-shift proxy, we transition from micro-level, exercise-by-exercise prediction to session-level attribution. Because the micro-level model focuses on isolated exercise outcomes rather than cumulative improvement, we aggregate its predictions across full session trajectories to operationalize and observe capability shifts. This analysis evaluates how the model-estimated capability shifts are explained by the session's overall composition—specifically the proportions of Under-Challenged, Optimally Challenged, and Over-Challenged exercises. By regressing the capability-shift proxy against these macro-level proportions, we interpret the micro-level interactional dynamics through the pedagogical framework of the Input Hypothesis.

3.6.1 Trajectory-Level Variables

For each valid session-level trajectory, we extract the following variables:

- **Difficulty alignment proportions.** The cumulative proportion of exercises classified as Under-Challenged, Optimally Challenged, and Over-Challenged across the full session.
- **Capability shift.** The in-situ capability-shift proxy as defined in Section 3.5.
- **Exercise-length deltas (control variables).** To account for the possibility that shifts in exercise surface complexity, rather than actual capability improvements, drive changes in the predicted success probability, we compute control variables capturing how exercise length changes between the early and late windows. Using the same early/late window indices as the capability-shift computation, we calculate $\Delta_{\text{chars}} = \overline{\ell}^{\text{chars}}_{\text{late}} - \overline{\ell}^{\text{chars}}_{\text{early}}$ and $\Delta_{\text{words}} = \overline{\ell}^{\text{words}}_{\text{late}} - \overline{\ell}^{\text{words}}_{\text{early}}$, where $\overline{\ell}$ denotes the mean exercise body length (in characters or words) within the respective window.
- **Teacher identity (control variable).** It is a binary indicator (dummy variable) encoding which of the two tutors in the dataset conducted the session. This fixed effect is included to systematically control for unobserved, tutor-specific baseline differences in pedagogical style, hint generosity, or exercise formulation.

3.6.2 Controlled Linear Mixed-Effects Regression

To evaluate the relationship between session composition and estimated capability shift while accounting for the nested data structure, we regressed the capability-shift proxy on the cumulative difficulty proportions using a Linear Mixed-Effects Model (LMM). This model controls for shifts in exercise surface complexity (Δ_{chars} and Δ_{words}) and tutor identity as a fixed effect, while including a random intercept for student identity to account for multiple sessions per learner.

Because the three difficulty alignment proportions are compositionally constrained (they sum to one within each trajectory), entering all three as predictors would introduce perfect collinearity. We therefore fit the model by including two of the three proportions, omitting the third as the reference category.

Formally, omitting Under-Challenged as the ref-

erence, the model is specified as:

$$\begin{aligned} \text{CS}_{js} = \beta_0 &+ \beta_1 \cdot \text{PropOptim}_{js} \\ &+ \beta_2 \cdot \text{PropOver}_{js} \\ &+ \beta_3 \cdot \Delta_{\text{chars},js} \\ &+ \beta_4 \cdot \Delta_{\text{words},js} \\ &+ \gamma \cdot \text{Tutor}_s + u_j + \varepsilon_{js} \end{aligned} \tag{2}$$

where j indexes the student and s indexes the tutor. Under this parameterization, β_1 estimates the expected change in capability shift when shifting proportion from Under-Challenged to Optimally Challenged exercises (holding Over-Challenged and all controls constant), and β_2 estimates the corresponding contrast from Under-Challenged to Over-Challenged.

To recover the third pairwise contrast (Optimally Challenged versus Over-Challenged), we equivalently re-parameterize the same model with Over-Challenged as the reference category. This controlled specification quantifies the theoretical variables' impact on the expected capability shifts while mathematically isolating them from structural confounders.

4 Results

4.1 Exercise Identification

From the initial 2,527 sequences labeled as containing tutor exercises, we evaluated the filtering pipeline described in Section 3.1 on a 10% human-annotated verification set ($n = 252$), in which one author manually judged exercise eligibility by reviewing each target turn alongside a context window of 15 surrounding turns.

GPT-4.1 (prompted at temperature 0) was evaluated against these human judgments on the same verification set, achieving an accuracy of 0.84, precision of 0.88, recall of 0.85, and Cohen's $\kappa = 0.66$. Under the Landis and Koch scale (Landis and Koch, 1977), this represents substantial agreement. The moderate κ relative to the high accuracy reflects the inherent ambiguity of boundary cases—particularly turns involving partially self-contained exercises or where the distinction between scaffolding for a prior exercise and initiation of a new one was unclear. Given the strong overall validation, especially the high precision, we applied the automated pipeline to the remaining sequences, yielding a final dataset of 1,566 eligible exercises.

Label	Supp.	Acc.	Prec.	Rec.	F1
Hint / Instruction	124	.975	.824	.871	.847
Direct Answer	26	.993	.800	.769	.784
Proc. Q. / Confusion	6	.999	.857	1.000	.923
Content Q. / Confusion	24	.993	.783	.750	.766
Response Attempt:					
Correct	121	.985	.930	.876	.902
Incorrect	61	.989	.891	.803	.845
Partial	23	.986	.513	.870	.645
NA	1185	.963	.992	.959	.975
Overall	**1570**	**.932**	**.932**	**.932**	**.932**

Table 3: LLM–human agreement on turn-level behavioral annotation ($n = 1{,}570$ turns from 157 exercises). Accuracy is computed in a one-vs-rest setting per label.

4.2 Turn-Level Behavioral Annotation

To scale annotation to the full corpus, we evaluated GPT-5.2 against reconciled human annotations (157 exercises, 1,570 turns). Table 3 reports per-label performance. Overall accuracy was 0.93 (Cohen's $\kappa = 0.84$), indicating strong alignment with human judgment across the taxonomy. Performance was highest for structurally unambiguous labels such as Response Attempt – Correct (F1 = 0.90). The primary error mode was Response Attempt – Partial (F1 = 0.65), where the model frequently conflated partial responses with fully correct or incorrect ones: a distinction that requires evaluating completeness relative to the exercise prompt. Given the strong overall agreement, we applied the automated pipeline to annotate the remaining corpus.

We further validated the LLM annotations at the trajectory-state level, where they enter the downstream model. Applying the mapping rule of Section 3.3 to both human and GPT-5.2 turn labels yields a difficulty alignment state for each of the 157 human-annotated exercises. Agreement between human-derived and LLM-derived states is substantial (accuracy = 0.885, Cohen's $\kappa = 0.786$). Per-class F1 ranges from 0.81 (Over-Challenged) to 0.92 (Under-Challenged), with Optimally Challenged at 0.85. The full state-level confusion matrix and per-class breakdown appear in Appendix B.

4.3 Difficulty Alignment State Distribution

Applying the trajectory-to-challenge mapping (Section 3.3) to all 1,566 eligible exercises yields a markedly skewed distribution: 920 exercises (58.7%) were classified as Under-Challenged, 165 (10.5%) as Optimally Challenged, and 481 (30.7%) as Over-Challenged.

The predominance of Under-Challenged exer-

cises indicates that, across the corpus, students most frequently answered correctly without exhibiting prior evidence of content-level struggle. The relatively small proportion of Optimally Challenged exercises, where students demonstrated initial difficulty but ultimately succeeded, is consistent with the theoretical expectation that the zone of productive struggle is narrow relative to tasks that are either too easy or too difficult. The substantial Over-Challenged proportion (30.7%) reflects cases where the tutor ultimately provided the answer or the interaction reached the turn limit without a correct student response.

4.4 Next-Exercise Success Prediction

The generalized linear mixed-effects model achieved a mean AUC of 0.61 and mean classification accuracy of 0.70 in the 5-fold cross-validation. The above-chance AUC confirms that the retained trajectory features carry a discriminative signal for next-exercise outcomes beyond baseline prevalence. To ensure these results are robust to model specification, we tested alternative modeling structures (detailed in Appendix E), which yielded near-identical predictive performance.

To identify which features drive the model's predictions, we report standardized fixed-effect coefficients in Table 4. The strongest pedagogical predictor was the count of consecutive optimally challenged exercises immediately preceding the target turn ($\beta = 0.233$, $p = 0.029$): an uninterrupted sequence of productive difficulty is associated with higher expected learner success on the subsequent task. Among controls, average hints per exercise showed a positive association ($\beta = 0.201$, $p = 0.049$), consistent with the hypothesis that scaffolding facilitates subsequent independent performance. The CEFR proficiency ordinal yielded a negative coefficient ($\beta = -0.184$, $p = 0.020$), reflecting that higher-proficiency learners might face more demanding exercises in this corpus rather than indicating lower capability. Cumulative difficulty proportions were not significant predictors of immediate turn-level success ($p > 0.30$); their role in shaping macro-level session trajectories is examined in Section 3.6.

4.5 Associating Difficulty Alignment with Capability Shift

To test whether the session-level composition of difficulty explains the capability-shift proxy, we analyzed the 153 valid session-level trajectories ($n \geq 4$ exercises) using the controlled LMM described in Section 3.6. To evaluate all pairwise contrasts among difficulty states, the model was parameterized twice, rotating the reference category. Complete estimates are presented in Table 5.

Fixed Effects	β	**SE**	p
Consecutive Optimal States	0.233	0.107	**.029**
Avg. Hints per Exercise	0.201	0.102	**.049**
CEFR Level Ordinal	−0.184	0.079	**.020**
Prop. Under-Challenged	0.148	0.155	.340
Prop. Over-Challenged	−0.135	0.164	.411
Next Exercise # Words	−0.111	0.183	.543
Consecutive Successes	0.110	0.096	.250
Total Teacher Hints	−0.103	0.097	.288
Next Exercise # Chars	−0.075	0.184	.683
Consecutive Failures	−0.055	0.094	.555
Total Procedural Qs	−0.038	0.070	.583
Last 3 Success Rate	0.031	0.187	.870
Total Content Qs	−0.003	0.071	.961
Intercept	0.623	0.158	$<$ **.001**
Tutor Fixed Effect	0.439	0.208	**.035**

Table 4: Standardized fixed-effect coefficients from the GLMM predicting next-exercise success. All continuous predictors are z-scored per fold. Student random intercept variance: 0.003.

The pairwise contrasts show two main findings:

The Detrimental Effect of Over-Challenge. The data strongly indicates that excessive difficulty hinders capability improvement. When specifying Over-Challenged as the baseline reference, shifting the session composition toward either Under-Challenged ($\beta = 0.136$, $p < 0.001$) or Optimally Challenged ($\beta = 0.219$, $p < 0.001$) yields a substantial and highly significant increase in the capability-shift proxy.

The Superiority of Optimal Challenge. The critical theoretical test lies in the contrast between Optimally Challenged and Under-Challenged states. When holding Over-Challenged constant and utilizing Under-Challenged as the baseline, increasing the proportion of Optimally Challenged exercises is associated with a significant positive increase in the expected capability shift ($\beta = 0.083$, $p = 0.041$). This pattern is consistent with sustained productive difficulty ($i + 1$) producing greater capability improvement than simple independent mastery (i), supporting the trajectory-based difficulty alignment classification.

5 Discussion

Our results provide evidence consistent with the Input Hypothesis by establishing the hierarchy: Optimally Challenged > Under-Challenged > Over-

Table 5: Controlled LMM predicting the capability-shift proxy from difficulty alignment proportions ($n = 153$ trajectories). Panels (a) and (b) are re-parameterizations of the same model, differing only in the omitted reference category to display all pairwise contrasts. The model includes a student-level random intercept.

(a) Reference Group: Under-Challenged

Predictor	β	**SE**	p
Intercept	-0.020	0.023	.386
Prop. Optimal $i+1$	0.083	0.041	**.041**
Prop. Over $i+\mathbb{Z}_{>1}$	-0.136	0.032	$<$ **.001**
Structural Controls			
Δ Characters (Δ_{chars})	-0.0005	0.0006	.378
Δ Words (Δ_{words})	0.0004	0.0031	.888
Tutor Fixed Effect	0.056	0.019	**.002**

(b) Reference Group: Over-Challenged

Predictor	β	**SE**	p
Intercept	-0.155	0.026	$<$ **.001**
Prop. Optimal $i+1$	0.219	0.041	$<$ **.001**
Prop. Under i	0.136	0.032	$<$ **.001**
Structural Controls			
Δ Characters (Δ_{chars})	-0.0005	0.0006	.378
Δ Words (Δ_{words})	0.0004	0.0031	.888
Tutor Fixed Effect	0.056	0.019	**.002**

Challenged. Specifically, the robust negative association for the Over-Challenged proportion suggests excessive difficulty hinders development, while the positive contrast between Optimally and Under-Challenged states indicates that sessions with greater productive struggle are associated with greater capability shifts ($p = 0.041$). Crucially, these associations remain robust after controlling for student baseline variance, tutor identity, and exercise surface complexity. The near-zero coefficients for Δ_{chars} and Δ_{words} confirm the classification captures substantive interactional demand rather than superficial prompt changes. Together, these findings provide quantitative support for the value of optimal challenge in dynamic tutoring.

Beyond these findings, this framework provides a highly scalable approach for educational measurement. Instead of requiring the LLM to directly infer the abstract, latent difficulty of a task, the framework assigns the model a much simpler objective: classifying explicit, observable interactional behaviors. Because this specific classification task is straightforward for the LLM, the resulting turn-level annotations are reliable and amenable to automation at scale. Consequently, this automatic behavior-based classification framework is in principle transferable to other flexible, open-ended, and dynamically generated on-the-fly educational contexts.

This reliable, real-time annotation enables two direct system applications. First, the trajectory-to-challenge mapping can serve as a continuous evaluation metric for both human-led and LLM-generated exercises, providing computational feedback on whether the current pedagogical strategy maintains the learner's optimal challenge zone. Second, these alignment states can function as a theoretically grounded reward signal for reinforcement learning frameworks; rather than simply rewarding correct answers, LLMs can be optimized to generate interaction trajectories that sustain productive struggle. Crucially, because the classification operates on the immediate post-exercise interaction rather than extensive prior history, deployed systems can utilize this framework from the first exercise of a session without facing a cold-start problem.

However, these applications must be implemented with strict caution regarding their current empirical validation. The associations identified in this study are observational, and we do not demonstrate causal effects of optimal challenge on learning outcomes, nor can we guarantee generalizability across different demographics or learning domains. Translating these within-session associations based on a model-estimated proxy into generalized claims about learning, or deploying them as definitive training targets in live systems, requires future experimental manipulation and external outcome measurement.

Limitations

Aggregated Proportions versus Sequential Dynamics. Our macro-level model aggregates each session's difficulty history into cumulative proportions, discarding the chronological ordering of states. Yet sequence likely matters: an Over-Challenged opening followed by Optimally Challenged exercises may yield different gains than the reverse, even at identical proportions. Future work should apply sequence-aware models to capture how specific state transitions shape learning trajectories.

Annotation Visibility. Our classification rests on observable interaction in the chat transcript: typed turns, errors, and explicit questions. The transcript necessarily misses cognitive effort that does not surface in writing. A student who silently translates a word, consults an external resource, or works through a difficult construction in their head before

typing a correct answer will be classified as Under-Challenged on the visible evidence, even though the exercise may have exceeded the student's independent capability. This systematically biases the classification toward Under-Challenged for learners who externalize less of their reasoning, and would correspondingly understate the proportions of Optimally Challenged and (where external help is used) Over-Challenged exercises. Richer multimodal signals, such as typing latencies, screen capture, and eye tracking, could close this visibility gap.

Proxy Validity. Because the TSCC lacks pre- and post-assessments, our dependent variable in RQ3 is an in-situ proxy derived from the micro-level predictive model's out-of-fold probabilities. This creates an inferential dependency: the sensitivity of RQ3's findings is bounded by the predictive model's discriminative capacity (AUC = 0.61). The proxy captures short-term behavioral shifts within tightly bounded sessions (mean $\approx$ 5.8 prior exercises) rather than durable language acquisition. RQ3's results should therefore be interpreted as associations detected through the lens of this model, not as independent evidence of learning outcomes.

Unmeasured Confounders. Our analyses are observational, and the reported coefficients identify associations rather than causal effects. The student-level random intercept absorbs stable between-learner differences and the tutor fixed effect absorbs stable between-tutor differences in baseline, but several plausible confounders remain incompletely controlled. First, within-learner ability can fluctuate across sessions in ways the random intercept does not capture. Second, tutor adaptation is dynamic. A tutor may select harder exercises for students who appear to be progressing well, which would induce reverse causation between session composition and the capability-shift proxy. Third, task type (e.g., grammar drills vs. open-ended writing prompts) likely affects both how difficulty manifests in the trajectory and how the proxy responds, beyond what exercise-length controls absorb. Establishing causal effects would require experimental manipulation of session composition or substantially richer logs of tutor decision-making.

Data Scope. The analysis relies on 12 students and 2 teachers from the TSCC Version 2. This limited sample constrains the generalizability of both the predictive model and the session-level inferences. In particular, the student random intercepts are estimated from few sessions per learner, and the tutor fixed effect captures only a single pairwise contrast. Replication on larger, longitudinal corpora with standardized external assessments is necessary to establish whether these associations generalize across populations and translate into lasting acquisition gains.

Acknowledgments

This work was supported in part by the Stanford Education Data Science MS Fund.

References

Elizabeth L. Bjork and Robert A. Bjork. 2011. Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. In Morton Ann Gernsbacher, Richard W. Pew, Leaetta M. Hough, and James R. Pomerantz, editors, *Psychology and the Real World: Essays Illustrating Fundamental Contributions to Society*, volume 2, pages 56–64. Worth Publishers.

Thiago Brant, Julien Kühn, and Jun Pang. 2026. Estimating exam item difficulty with llms: A benchmark on brazil's enem corpus. *Preprint*, arXiv:2602.06631.

Norman E. Breslow and David G. Clayton. 1993. Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88(421):9–25.

Andrew Caines, Helen Yannakoudakis, Helena Edmondson, Helen Allen, Pascual Pérez-Paredes, Bill Byrne, and Paula Buttery. 2020. The teacher-student chatroom corpus. *Preprint*, arXiv:2011.07109.

Albert T. Corbett and John R. Anderson. 1994. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction*, 4(4):253–278.

Council of Europe. 2001. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Cambridge University Press.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120.

Ronald K. Hambleton, Hariharan Swaminathan, and H. Jane Rogers. 1991. *Fundamentals of Item Response Theory*. Sage Publications, Newbury Park, CA.

Manu Kapur. 2008. Productive failure. *Cognition and Instruction*, 26(3):379–424.

Stephen Krashen. 1977. Some issues relating to the monitor model. In H. Douglas Brown, Carlos Alfredo Yorio, and Ruth H. Crymes, editors, *Teaching and Learning English as a Second Language: Trends in Research and Practice: On TESOL '77: Selected Papers from the Eleventh Annual Convention of Teachers of English to Speakers of Other Languages, Miami, Florida, April 26–May 1, 1977*, pages 144–158. Teachers of English to Speakers of Other Languages, Washington, DC.

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

Yunting Liu, Shreya Bhandari, and Zachary A. Pardos. 2025. Leveraging llm respondents for item evaluation: A psychometric analysis. *British Journal of Educational Technology*, 56(3):1028–1052.

Michael H. Long. 1996. The role of the linguistic environment in second language acquisition. In William C. Ritchie and Tej K. Bhatia, editors, *Handbook of Second Language Acquisition*, pages 413–468. Academic Press, San Diego.

Seyed Parsa Neshaei, Richard Lee Davis, Adam Hazimeh, Bojan Lazarevski, Pierre Dillenbourg, and Tanja Käser. 2024. Towards modeling learner performance with large language models. *Preprint*, arXiv:2403.14661.

Chris Piech, Jonathan Spencer, Jonathan Huang, Surya Ganguli, Mehran Sahami, Leonidas Guibas, and Jascha Sohl-Dickstein. 2015. Deep knowledge tracing. *Preprint*, arXiv:1506.05908.

Petter Törnberg. 2023. Chatgpt-4 outperforms experts and crowd workers in annotating political twitter messages with zero-shot learning. *Preprint*, arXiv:2304.06588.

Lev S. Vygotsky. 1978. *Mind in Society: The Development of Higher Psychological Processes*. Harvard University Press, Cambridge, MA.

Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023. Is ChatGPT a good NLG evaluator? a preliminary study. In *Proceedings of the 4th New Frontiers in Summarization Workshop*, pages 1–11, Singapore. Association for Computational Linguistics.

A Exercise Identification Details

We defined an eligible exercise as a tutor turn presenting a language practice opportunity satisfying three conditions:

1. **Content Presence:** The turn contains the actual exercise body (e.g., a sentence to complete), rather than merely announcing an upcoming task.
2. **Contextual Completeness:** The exercise is self-contained within the chat history, excluding tasks requiring external resources (textbooks, URLs, images, audio).
3. **Productive Intent:** The task requires the student to produce English (structured drills, translation, open-ended prompts), excluding scaffolding for prior exercises, social exchanges, and administrative setup.

The annotation pipeline proceeded in four stages: (1) one author manually annotated a random 10% sample ($n = 252$), reviewing each target turn with a context window of 15 surrounding turns; (2) GPT-4.1 (temperature 0) was evaluated on this verification set against the precision threshold defined in Section 3.1; (3) the automated pipeline was applied to the remaining sequences; (4) student-initiated exercises were removed to restrict the analysis to tutor-led pedagogical strategies. The prompt is documented in Appendix C.1.

B State-Level Agreement on the Verification Set

To complement the turn-label agreement reported in Section 4.2, we applied the trajectory-to-challenge mapping rule (Section 3.3) separately to the human and GPT-5.2 turn-label sequences for the 157 manually annotated exercises, obtaining a derived difficulty alignment state from each annotation source. Table 6 reports the resulting confusion matrix.

Overall agreement is substantial (accuracy = 0.885, Cohen's $\kappa = 0.786$ under the Landis and Koch scale (Landis and Koch, 1977)). The Optimally Challenged class has high precision (0.944) and modest recall (0.773): when the model labels an exercise Optimally Challenged it almost always agrees with the human annotator (17 of 18 cases), but it misses 5 of 22 human-labeled Optimally Challenged exercises, assimilating them into the neighboring Under- or Over-Challenged categories. The model never labels a human-Under exercise as Optimally Challenged, supporting the interpretation that LLM-derived Optimal-state counts are conservative rather than inflated. The dominant disagreement pattern is between Under-Challenged and Over-Challenged (12 of 18 total errors).

	LLM (GPT-5.2)			
Human	Under	Optimal	Over	**Total**
Under-Challenged	90	0	9	99
Optimally Challenged	3	17	2	22
Over-Challenged	3	1	32	36
Total	96	18	43	157

Table 6: Confusion matrix between human-derived and GPT-5.2-derived difficulty alignment states on the 157-exercise verification set. Rows: human; columns: LLM. Overall accuracy = 0.885, Cohen's $\kappa = 0.786$.

C Prompts

This appendix lists the specific prompts provided to the Large Language Model for the automated annotation pipelines described in Sections 3.1 and 3.2.

C.1 Exercise Identification Prompt

System Prompt: Exercise Identification

```
You are an expert annotator for educational dialogues in English language learning
    contexts.
```

```
Your task is to determine if the "Current Turn" is an **eligible exercise turn**
    posed by the {role}.

## Definition of Eligible Exercise

An eligible exercise is a turn where the {role} presents a language practice
    opportunity that:

1. **Contains the actual exercise content** (not just an announcement or instruction
     or scaffolding about an upcoming or previous stated exercise)
2. **Is self-contained within the chat** (does not require external resources like
    textbooks, videos, links, or materials not present in the conversation)
3. **Requires the student to produce or practice English**

## Types of Eligible Exercises

- **Structured exercises**: fill-in-the-blank, grammar drills, vocabulary exercises,
     sentence completion, translation tasks, writing prompts (e.g., "Write a
    paragraph about...")
- **Eliciting questions for language practice**: open-ended questions designed to
    get the student to practice expressing themselves in English (e.g., "Can you
    describe a typical day for you?", "Tell me about your favorite hobby", "What
    would you do if...?")

## What is NOT an Eligible Exercise

### 1. Announcements or setup without exercise body
Turns that merely introduce an exercise without containing the exercise itself.
    However, if a turn combines the announcement AND the exercise body, it IS
    eligible.

**Example sequence:**
- Turn X -- Teacher: "Let's try to add the word 'yesterday' to the following
    sentence!" -> **NOT eligible** (announcement only)
- Turn X+1 -- Teacher: "She went back home." -> **ELIGIBLE** (this is the exercise
    body)
- Turn X+2 -- Teacher: "Insert the word 'yesterday' in the sentence I just sent you
    ." -> **NOT eligible** (instruction or scaffolding referencing previous exercise
    )

**Contrast with:**
- Teacher: "Let's try to add the word 'yesterday' to the following sentence: 'She
    went back home.'" -> **ELIGIBLE** (contains the exercise body in one turn)

**Note:** If you cannot find the exercise body in the current turn, previous context
    , OR post context, then the turn is NOT eligible -- this implies it requires
    external materials not present in the chat.

### 2. Scaffolding or hints for a previous exercise
Follow-up prompts that support an already-given exercise but do not constitute a new
     exercise themselves (e.g., "Please answer in a full sentence.", "Any other
    ideas?", "Can you try again?", "Think about what tense to use.")

### 3. Social/administrative exchanges
Casual greetings, check-ins, or procedural questions that do not involve language
    practice (e.g., "How are you?", "Are you ready to start?", "Can we do some
    exercises?", "Did you finish your homework?")

### 4. Exercises requiring external materials
If completing the exercise requires access to resources not fully present in the
    conversation:
- Textbooks, worksheets, images, charts, videos, audio files
- External links or apps (e.g., Skype messages, Instagram, websites)
- Reading passages or materials not included in the chat

**Key indicators that an exercise requires external materials:**
- Mentions of "chart," "image," "picture," "textbook," "video," "audio," "page," or
    "link"
```

```
- References to specific question numbers (e.g., "What do you think about question
    43?") without providing the question content -- this implies they are looking at
     an external source
- Instructions like "Look at the reading passage" or "Based on the video we watched"
     when that content is not in the chat

### 5. Simple yes/no or factual questions
Questions that can be answered with a single word or simple fact and do not promote
    meaningful language practice (e.g., "Is this correct?", "Do you understand?", "
    What's 2+2?")

## Decision Process

1. **Read the Current Turn carefully.** Does it contain actual exercise content that
     requires the student to produce English?

2. **Check the Post Context.** If the exercise body appears AFTER the current turn,
    then the current turn is just an announcement -> NOT eligible.

3. **Check for external material dependencies.** Using the Previous and Post Context
    , determine whether the exercise references materials or content not present in
    the conversation -> NOT eligible.

4. **Distinguish from scaffolding.** Is this turn providing hints or follow-up for
    an exercise that was already given in a previous turn? -> NOT eligible.

---

Previous Context:
{prev_context}

Current Turn ({role}):
{current_turn}

Post Context:
{post_context}

---

First, write a brief reasoning (maximum 50 words) explaining your decision.
Then provide the final label.

Output format:
<analysis>
Your reasoning here (<= 50 words).
</analysis>
<judgement>
YES or NO
</judgement>
```

C.2 Turn-Level Behavioral Annotation Prompt

System Prompt: Post-Exercise Discussion Annotation

```
You are an expert linguistic annotator analyzing English as a Foreign Language (EFL)
     tutoring dialogues.
Your task is to analyze the "Post-Exercise Discussion" and label the pedagogical
    function of each turn relative to a specific "Exercise Body."

### INPUT DATA
1. Previous Context: The conversation history leading up to the exercise.
2. Exercise Body: The specific problem, sentence, or grammar task the student must
    solve.
3. Post-Turn Text: A sequence of up to 10 dialogue turns immediately following the
    exercise introduction.

### LABEL TAXONOMY
Choose exactly one label for each turn from this list:
```

```
TEACHER LABELS
- Hint/Instruction
- Direct Answer  [STOP CONDITION]

STUDENT LABELS
- Procedure Question or Confusion
- Content Question or Confusion
- Response Attempt - Correct  [STOP CONDITION]
- Response Attempt - Incorrect
- Response Attempt - Partial

OTHER
- NA

### DEFINITION OF LABELS
- Hint/Instruction: Scaffolding, hints, elicited questions (e.g., "Tell me more"),
    or instructions.
    Also, if the Exercise Body is vague (e.g., "Make a sentence") and the teacher
        provides the specific exercise text in the chat, label that turn as `Hint/
        Instruction`.
- Direct Answer: The teacher explicitly provides the correct answer.
- Procedure Question or Confusion: Student asks how to do the task or expresses
    confusion about the activity steps (not language knowledge).
- Content Question or Confusion: Student asks about language knowledge/content, ways
     to improve the answers, or expresses lack of knowledge specific to this
    exercise.
- Response Attempt - Correct: Student provides a grammatically and semantically
    complete answer that satisfies the exercise.
- Response Attempt - Incorrect: Student attempts to answer, but the content is wrong
     or contains typos.
- Response Attempt - Partial: Student provides a correct but incomplete answer (e.g
    ., filling only 1 of 2 blanks).
- NA: The turn cannot be labeled with any of the above labels. Turn does not fit
    above categories OR is off-topic / meta-talk.

### CRITICAL ANNOTATION RULES
1. Strict Scope: Labels must strictly relate to the Current Exercise Body.
2. Stop Condition: Process turns sequentially. STOP labeling immediately after
    assigning Direct Answer or Response Attempt - Correct.
3. Always follow the teacher's judgement on correctness of the student's answer.

### DIALOGUE INPUT
Previous Context:
{prev_context}

Exercise Body:
{exercise_body}

Post-Turns:
{turns_text}

### OUTPUT FORMAT
Return pure JSON only. No markdown and no extra keys.
{{
  "annotations": [
    {{
      "turn_id": 1,
      "speaker": "Student or Teacher",
      "text": "turn text",
      "reasoning": "brief reason (<=30 words)",
      "label": "one label from taxonomy"
    }}
  ]
}}
```

D Feature Dictionary and L1 Selection Frequencies

Table 7 reports the full pool of 24 candidate features and their L1 selection frequencies across 5-fold session-level cross-validation, using the inclusion criterion described in Section 3.4. The 13 features retained for the final GLMM are highlighted in bold. Structural control variables (the student random intercept and the tutor fixed effect) were not subject to L1 penalization and were unconditionally retained.

Feature Name	Description	Selection Ratio
Learner Background		
CEFR Level Ordinal	Ordinal encoding of CEFR proficiency (1=B1, 2=B2, 3=C1, 4=C2).	**1.00**
CEFR Level (Categorical)	Categorical representation of the student's CEFR proficiency.	0.30
Trajectory Length		
Prefix Length	Number of prior exercises observed in the current session.	0.00
Number of Previous Exercises	Redundant feature mathematically identical to Prefix Length.	0.00
Historical Performance		
Consecutive Successes	Count of consecutive correct responses immediately preceding $t+1$.	**1.00**
Consecutive Failures	Count of consecutive incorrect responses immediately preceding $t+1$.	**0.80**
Last 3 Success Rate	The accuracy rate over the most recent 3 exercises.	**0.60**
Last 2 Success Rate	The accuracy rate over the most recent 2 exercises.	0.40
Cumulative Success Rate	The overall accuracy rate across all prior exercises in the session.	0.00
Previous Exercise Correct	Binary indicator of whether the immediately preceding exercise was correct.	0.00
Historical Difficulty Alignment (i+1)		
Prop. Over-Challenged	Cumulative proportion of Over-Challenged exercises in the trajectory.	**1.00**
Prop. Under-Challenged	Cumulative proportion of Under-Challenged exercises in the trajectory.	**0.60**
Prop. Optimally Challenged	Cumulative proportion of Optimally Challenged exercises (dropped to avoid collinearity).	0.20
Consecutive Optimal States	Count of uninterrupted Optimally Challenged states immediately preceding $t+1$.	**1.00**
Consecutive Over-Challenged	Count of uninterrupted Over-Challenged states immediately preceding $t+1$.	0.00
Previous Difficulty State	Categorical variable for the difficulty state of the immediately preceding exercise.	0.00
Scaffolding & Interaction		
Avg. Hints per Exercise	Average number of tutor hints provided per exercise so far.	**1.00**
Total Teacher Hints	Cumulative count of all tutor hints provided in the trajectory.	**1.00**
Total Procedural Qs	Cumulative count of procedural questions asked by the student.	**0.80**
Total Content Qs	Cumulative count of content questions asked by the student.	**0.60**
Previous Ended Direct Answer	Binary indicator of whether the previous exercise ended in direct answer from teacher.	0.40
Previous Required Hint	Binary indicator of whether the previous exercise involved any tutor hints.	0.00
Next Exercise Properties (Complexity Controls)		
Next Exercise # Chars	Character length of the upcoming exercise prompt.	**0.60**
Next Exercise # Words	Word count of the upcoming exercise prompt.	**0.60**

Table 7: Dictionary of all initial trajectory features and their L1 selection frequencies across 5-fold cross-validation. Features selected in ≥ 0.60 of folds (highlighted in bold) were retained for the final GLMM.

E Robustness Checks for the Next-Exercise Success GLMM

To ensure that the predictive relationships identified in Section 3.4 are not artifacts of a specific statistical structure, we evaluated four variations of the Generalized Linear Mixed-Effects Model (GLMM).

The models tested include:

1. **Student Random + Tutor Fixed (Main Model):** Includes both a student-level random intercept and a categorical tutor fixed effect.
2. **Tutor Fixed Only:** Includes only the tutor fixed effect, omitting student-level variance control.
3. **Student Random Only:** Includes only the student-level random intercept, omitting the tutor fixed effect.
4. **Base Model:** A standard logistic regression omitting both student and tutor controls.

Table 8 demonstrates that all four model specifications yielded near-identical out-of-fold predictive performance. Table 9 reports the standardized coefficient estimates (β) and p-values across the three alternative specifications. Consistent with the main findings, the count of consecutive optimally challenged states (`consec_aligned_t`) remains a significant positive predictor across all model variations, confirming the robustness of this core pedagogical feature regardless of the structural controls applied.

Model Specification	Mean AUC	Mean Acc.
Student Random + Tutor Fixed (Main)	0.610	0.701
Tutor Fixed Only	0.610	0.701
Student Random Only	0.599	0.703
Base Model (No Controls)	0.599	0.703

Table 8: Cross-validated predictive performance across model structures ($N = 5$ folds). Hierarchical controls do not fundamentally alter discriminative power.

	Tutor Fixed Only		Student Random Only		Base Model	
Fixed Effects	β	p	β	p	β	p
Consec. Aligned States	.234	**.029**	.232	**.029**	.233	**.029**
Avg. Hints per Exercise	.203	**.047**	.156	.112	.156	.113
CEFR Level Ordinal	−.183	**.019**	−.125	.094	−.125	.091
Prop. Under-Challenged	.150	.337	.173	.261	.176	.254
Prop. Over-Challenged	−.135	.412	−.098	.545	−.096	.555
Next Ex. Length: Words	−.113	.547	−.122	.495	−.121	.508
Consec. Successes	.111	.249	.133	.164	.134	.161
Total Teacher Hints	−.104	.282	−.129	.178	−.131	.170
Next Ex. Length: Chars	−.073	.702	−.152	.393	−.155	.394
Consec. Failures	−.056	.551	−.053	.573	−.053	.572
Total Procedural Qs	−.038	.582	−.053	.443	−.053	.444
Last 3 Success Rate	.029	.879	.032	.865	.031	.870
Total Content Qs	−.004	.957	.006	.933	.006	.927
Intercept	.625	**< .001**	.922	**< .001**	.924	**< .001**
Tutor Fixed Effect	.432	**.035**	–	–	–	–

Table 9: Standardized coefficients and p-values across three alternative model specifications. All continuous predictors are z-scored per fold. Significant effects ($p < .05$) in bold. The effect of consecutive aligned states remains stable across all variations.Models include student random intercepts.