INDRA: Iterative Difficulty Refinement Attention for MCQ Difficulty Estimation for Indic Languages

Manikandan Ravikiran [†*], **Rohit Saluja**[† ◇], **Arnav Bhavsar** [†]
[†] Indian Institute of Technology, Mandi, India
[◇] BharatGen
erpd2301@students.iitmandi.ac.in
rohit@iitmandi.ac.in,arnav@iitmandi.ac.in

Abstract

Estimating the difficulty of multiple-choice questions (MCQs) is central to adaptive testing and learner modeling. We introduce **INDRA** (Iterative Difficulty Refinement Attention), a novel attention mechanism that unifies psychometric priors with neural refinement for Indic MCQ difficulty estimation. INDRA incorporates three key innovations: (i) *IRT-informed initialization*, which assigns token-level discrimination and difficulty scores to embed psychometric interpretability; (ii) *entropy-driven iterative refinement*, which progressively sharpens attention to mimic the human process of distractor elimination; and (iii) *Indic Aware Graph Coupling*, which propagates plausibility across morphologically and semantically related tokens, a critical feature for Indic languages. Experiments on `TEEMIL-H` and `TEEMIL-K` datasets show that INDRA achieves consistent improvements, with absolute gains of up to +1.02 F1 and +1.68 F1 over state-of-the-art, while demonstrating through ablation studies that psychometric priors, entropy refinement, and graph coupling contribute complementary gains to accuracy and robustness.

1 Introduction

Multiple-choice questions (MCQs) remain one of the most widely used formats for evaluating knowledge in educational and standardized testing. The difficulty of an MCQ plays a central role in assessment design, adaptive testing, and learner modeling. Automatically estimating question difficulty has thus emerged as a key challenge in educational NLP, with growing interest from both psychometric and machine learning communities (Benedetto et al., 2025).

Existing approaches fall into two broad categories. Psychometric models, such as Item Response Theory (IRT), offer interpretability by associating each item with difficulty and discrimination parameters (Chen et al., 2021; Lalor et al., 2016). However, they require large-scale response data and ignore the linguistic structure of questions and distractors. Neural approaches, particularly transformer-based models, directly model text but rely on uniform self-attention mechanisms (Hahn, 2020). Recent work has proposed specialized refinements: CASSA (Ravikiran et al., 2025a) adds task-aware biases to emphasize question relevance, while GISA (Ravikiran et al., 2025b) introduces iterative refinement through entropy minimization and masking. While these models improve performance, they remain limited in two respects: (i) they lack explicit psychometric grounding, and (ii) they are not designed for morphologically rich languages. This limitation is especially pronounced in Indic settings, where distractors are often morphologically or semantically close to the correct answer. For example, in a Hindi MCQ on state politics:

राज्य सरकार के निचले सदन का क्या नाम है? (*"What is the name of the lower house of the state legislature?"*), the options-विधान सभा, विधान परिषद, संसद, न्यायपालिका (*Vidhan Sabha, Vidhan Parishad, Sansad, Nyayapalika [Judiciary]*)-are institutionally related and differ only in suffixes or scope, making them highly confusable even for proficient learners. A similar challenge arises in Kannada, where a question on parliamentary roles:

ಲೋಕಸಭೆ ಸ್ಪೀಕರ್ ಅವರ ಪ್ರಮುಖ ಜವಾಬ್ದಾರಿ ಏನು? (*"What is the main responsibility of the Lok Sabha Speaker?"*), offers options—ಸದನದ ಅಧ್ಯಕ್ಷತೆ ವಹಿಸಿ ಸುಗಮ ಕಾರ್ಯನಿರ್ವಹಣೆ ಖಚಿತಪಡಿಸುವುದು, ಮಸೂದೆಗಳನ್ನು ಮಂಡಿಸುವುದು, ಕಲಾಪಗಳನ್ನು ನಡೆಸುವುದು, ಸರ್ಕಾರವನ್ನು ಪ್ರತಿನಿಧಿಸುವುದು (*Presiding over the house to ensure smooth functioning; Introducing bills; Conducting sessions; Representing the government*)-that are all grammatically correct and contextually plausible, yet only the first captures

[*] Corresponding Author.

Proceedings of the 1st Workshop on Benchmarks, Harmonization, Annotation, and Standardization for Human-Centric AI in Indian Languages (BHASHA 2025), pages 37–51
December 23, 2025 ©2025 Association for Computational Linguistics

the Speaker's true responsibility. Such cases illustrate why Indic languages form a demanding stress test for attention mechanisms: models must simultaneously contend with surface similarity, morphological variation, and semantically close distractors, all of which need to be explicitly modeled for reliable difficulty estimation (Ravikiran et al., 2025c).

As such, we introduce INDRA, a principled attention refinement mechanism for MCQ difficulty estimation. INDRA integrates four key components: (i) *psychometric initialization*, where token interactions are scaled by discrimination and difficulty parameters, embedding IRT-style priors at the token level; (ii) *entropy-driven iterative refinement*, which progressively sharpens attention distributions to mimic human distractor elimination; (iii) *Indic-aware graph coupling*, which propagates plausibility across morphologically, semantically, or syntactically related tokens; and (iv) *proximal stability*, which guarantees smooth convergence of refinement dynamics. Experiments in Section 4.2, INDRA consistently outperforms strong baselines across multiple datasets, achieving gains of in F1 and correlation with human difficulty labels. In summary, our contributions are as follows:

- We propose INDRA, a general attention refinement framework that unifies psychometric priors, entropy-driven iterative refinement, graph-based coupling, and stability control.
- We design token-level graphs that integrate morphological, semantic, and syntactic similarity, enabling adaptation to linguistically rich and low-resource settings such as Indic languages.
- Through extensive experiments on `TEEMIL-H` and `TEEMIL-K` MCQ datasets, we show that INDRA consistently improves predictive performance and interpretability over prior methods.

2 Related Work

MCQ Difficulty Estimation: Estimating the difficulty of multiple-choice questions (MCQs) is central to adaptive learning, automated assessments, and educational analytics. Traditional psychometric models such as Item Response Theory (IRT) (Al-zboon et al., 2021; Chen et al., 2021; Lalor et al., 2016) infer item difficulty using large-scale student response data, but rely on strong parametric assumptions and are difficult to extend across domains and languages. Neural approaches, especially transformer-based models such as BERT (Devlin et al., 2019), leverage contextual embeddings to predict difficulty labels directly. With datasets such as `Ext-MCQ` (Manikandan et al., 2025), and `TEEMIL` (Ravikiran et al., 2025c), these methods have shown promising improvements by capturing semantic relationships across stems, options, and distractors. However, most existing methods rely on general-purpose embeddings (Loukina et al., 2016; Veeramani et al., 2024) and single-pass attention mechanisms, which are not sufficient to capture the fine-grained dependencies between question elements and distractors (Venktesh et al., 2022). This has motivated attention refinements tailored for MCQ difficulty estimation.

Attention Mechanisms and Refinements: Self-attention underpins modern transformers, yet vanilla dot-product attention treats all token interactions uniformly, attenuating fine-grained cues needed to reason over stems, keys, and near-miss distractors. Positional and dependency-aware refinements improve granularity e.g., relative positions (Shaw et al., 2018), disentangled content/position attention in DeBERTa (He et al., 2021), and rotary position embeddings (Su et al., 2021) but these do not explicitly model the *stepwise* elimination dynamics required for predicting item difficulty. In MCQ difficulty estimation specifically, recent work ranks or predicts difficulty from item text and options (Bulut et al., 2024) and revisits psychometric underpinnings via IRT for NLP (Lalor et al., 2016; Zhou et al., 2025). Analyses of how transformers answer MCQs further suggest multi-stage internal procedures that standard attention does not expose (Wang et al., 2024). However, these approaches rarely fuse *psychometric priors* with *iterative attention refinement*, and are not tailored to morphologically rich settings where distractors differ by suffixation or compounding; recent Indic datasets highlight this gap and its impact on difficulty estimation (Ravikiran et al., 2025c). These factors motivate our proposed INDRA, which unifies psychometric initialization with entropy-driven iterative refinement and Indic-aware linguistic coupling.

3 Methodology

INDRA addresses the limitations of standard self-attention through four modules: (i) *psychometric initialization*, (ii) *entropy-driven iterative refinement*, (iii) *Indic-aware graph coupling*, and (iv) *proximal stability for convergence*. Together, these components transform INDRA into a principled replacement for standard attention, explicitly aligning token interactions with psychometric priors, refinement dynamics, and linguistic structure.

3.1 Psychometric Initialization

Classical Item Response Theory (IRT) models the probability that a learner with ability θ answers an item correctly using two parameters: *difficulty* b (how hard the item is) and *discrimination* a (how well the item separates strong learners from weak ones):

$$P(\text{correct} \mid \theta) = \sigma\big(a(\theta - b)\big).$$

We adapt this idea from items to tokens. Each token x_{ij} (token j in option i) is assigned a discrimination a_{ij} and a difficulty b_{ij}. Instead of starting from uniform dot-product attention, we bias the initial attention logits as

$$\ell_{ij}^{(0)} \;=\; a_{ij} \cdot \frac{q_i k_j^\top}{\sqrt{d}} \;-\; b_{ij}.$$

Intuitively, tokens that are more informative (high a_{ij}) are weighted up, while tokens that make the item harder (high b_{ij}) are weighted down. By aggregating across tokens, we can recover the familiar item-level IRT parameters, linking INDRA directly to psychometric theory while staying compatible with transformer attention. Unlike standard random initialization, INDRA seeds a_{ij} and b_{ij} from dataset-informed priors (see Algorithm 1).

Discrimination a_{ij} is scaled by token salience: tokens unique to one option receive higher values, while tokens shared across distractors are down-weighted. Morphological uniqueness, measured via normalized edit distance, further boosts the weight of distinctive tokens. Difficulty b_{ij} is initialized from human-annotated TEEMIL difficulty labels: easy items map to lower values, hard items to higher values, and medium items interpolate between the two. This design ensures that the starting logits $\ell^{(0)}$ already encode a plausible difficulty structure, improving stability of the refinement loop and providing interpretable links between token-level attention and educational constructs.

Algorithm 1: Psychometric Initialization in INDRA

```
[1] MCQ options $O = \{o_1, \dots, o_m\}$ with
  tokens $x_{ij}$, item-level difficulty label
  $y \in \{\text{Easy}, \text{Medium}, \text{Hard}\}$ Token-level
  discrimination $\{a_{ij}\}$ and difficulty $\{b_{ij}\}$
Initialize $a_{ij} \leftarrow 1.0$, $b_{ij} \leftarrow 0.0$ for all
  tokens **for** *each option $o_i$* **do**
each token $x_{ij}$ in $o_i$ Compute
  **morphological uniqueness**:
      $u(x_{ij}) = 1 - \min_{o_k \neq o_i} \frac{\text{EditDist}(x_{ij}, o_k)}{|x_{ij}|}$
  Compute **option overlap score**:
      $f(x_{ij}) = \frac{1}{\text{count}(x_{ij} \text{ across all options})}$
  Set discrimination prior (with $\alpha \in [0, 1]$):
      $a_{ij} \leftarrow \alpha \cdot u(x_{ij}) + (1 - \alpha) \cdot f(x_{ij})$
  Assign difficulty prior $b_{ij}$ from label $y$:
      $b_{ij} \leftarrow \begin{cases} 0.0, & y = \text{Easy} \\ 0.5, & y = \text{Medium} \\ 1.0, & y = \text{Hard} \end{cases} \quad \forall \text{ token in item}$
  Normalize $\{a_{ij}\}$ to mean 1.0 and $\{b_{ij}\}$ to
  mean 0.0 **return** $\{a_{ij}\}, \{b_{ij}\}$
```

3.2 Entropy-Driven Iterative Refinement

Human test-takers rarely identify the correct option in a single glance (Leighton and Gierl, 2017). Instead, they progressively narrow down the possibilities by ruling out distractors. To mimic this behavior, INDRA refines attention over multiple steps rather than collapsing into a single pass. At refinement step t, the distribution is

$$p^{(t)} = \text{softmax}\Big(\tfrac{1}{\tau}\ell^{(t-1)}\Big),$$

where $\tau > 0$ is a temperature parameter. A large τ produces a broad, uncertain distribution (analogous to considering all options), while a small τ yields a sharper focus (analogous to eliminating unlikely distractors). By iterating this update for a small number of steps, irrelevant tokens are suppressed gradually instead of being discarded too early. This produces smoother and more interpretable attention trajectories that better mirror the incremental reasoning strategies observed in human test-taking.

3.3 Indic-Aware Graph Coupling

In Indic languages, distractors often differ from the correct answer through systematic variations such as inflectional endings, compounding, derivational morphology, synonymy, or code-mixing. These patterns make distractors highly confusable: surface similarity is high, yet subtle semantic differences determine correctness. For instance, in Hindi:

1930 के दशक में ब्रिटिश सरकार द्वारा भारत सरकार में सुधार के प्रयास का क्या नाम था? (*“In the 1930s, what was the name of the British Government’s attempt to reform the Government of India?”*) **Options:** भारत सरकार अधिनियम, भारतीय स्वतंत्रता अधिनियम, भारत सरकार सुधार अधिनियम All share the prefix भारत सरकार and differ only in suffixes such as अधिनियम vs. सुधार अधिनियम, making them morphologically and semantically close. Similarly, in Kannada:

ಭಾರತದಲ್ಲಿ ಕಾರ್ಮಿಕರ ಕಡಿಮೆ ಉತ್ಪಾದಕತೆಗೆ ಪ್ರಮುಖ ಕಾರಣ ಏನು? (*“What is the main reason for low worker productivity in India?”*) **Options:** ತರಬೇತಿ ಕೊರತೆ, ಸಂಘಟನೆ ಕೊರತೆ, ನಾಯಕತ್ವದ ಕೊರತೆ Each option shares the suffix ಕೊರತೆ (“lack of”), forming systematic morphological variants.

Such cases highlight that standard attention, which treats tokens independently, cannot reliably eliminate distractors without modeling these structural relations. Graph coupling addresses this by ensuring plausibility is initially shared among related variants and only suppressed when sufficient contextual evidence emerges.

Algorithm 2: Construction of Token Similarity Graph $\hat{G}$

```
Input: Tokens $x_{1:n}$ with hidden states $h_{1:n}$;
       tokenizer $\mathcal{T}$; dependency edges
       DepTree; weights $\lambda_{\text{morph}}, \lambda_{\text{sem}}, \lambda_{\text{syn}}$;
       scale $\sigma_m > 0$, sparsity $k$
Output: Row-normalized graph
        $\hat{G} \in \mathbb{R}^{n \times n}$

for $i = 1$ to $n$ do
    for $j = 1$ to $n$ do
        $G_{ij}^{\text{morph}} \leftarrow \exp(-\text{ED}(\mathcal{T}(x_i), \mathcal{T}(x_j))/\sigma_m)$
        $G_{ij}^{\text{sem}} \leftarrow \cos(h_i, h_j)$
        $G_{ij}^{\text{syn}} \leftarrow 1$ if $(i, j) \in$ DepTree else $0$
    Keep only top-$k$ neighbors in row $i$
$G \leftarrow \lambda_{\text{morph}} G^{\text{morph}} + \lambda_{\text{sem}} G^{\text{sem}} + \lambda_{\text{syn}} G^{\text{syn}}$
Row-normalize rows: $\hat{G}_{i:} \leftarrow G_{i:} / \sum_j G_{ij}$
return $\hat{G}$
```

Algorithm 2 outlines the construction of $\hat{G}$ for each MCQ, integrating morphological, semantic, and syntactic kernels into a sparse, row-normalized diffusion matrix. For each MCQ, we build a token similarity graph $\hat{G} \in \mathbb{R}^{n \times n}$ that integrates three signals:

$$\begin{aligned} G_{ij} = \lambda_{\text{morph}} \exp\Big(-\tfrac{\text{ED}(x_i, x_j)}{\sigma_m}\Big) \\ + \lambda_{\text{sem}} \cos(h_i, h_j) \\ + \lambda_{\text{syn}} \mathbf{1}\{(i, j) \in \text{DepTree}\}, \end{aligned} \tag{1}$$

where ED is normalized edit distance (morphology), $\cos(h_i, h_j)$ is semantic similarity between contextual embeddings, and $\mathbf{1}$ encodes syntactic adjacency. The weighted graph G is row-normalized to produce $\hat{G}$, with top-k sparsification applied per row for scalability. At refinement step t, attention propagates through the graph as

$$\tilde{p}^{(t)} = (I + \beta \hat{G})\, p^{(t)}, \qquad \beta \geq 0, \tag{2}$$

where β controls propagation strength. Small β keeps updates localized, while larger β diffuses plausibility across morphologically and semantically related tokens. This coupling stabilizes refinement and prevents premature collapse onto a single option when distractors are nearly indistinguishable.

3.4 Proximal Stability for Convergence

Repeated refinement and diffusion can destabilize logits, especially when entropy is low or graph coupling is strong. To guarantee smooth convergence, INDRA applies *proximal damping*:

$$\ell^{(t)} = (1 - \gamma)\ell^{(t-1)} + \gamma W_p \tilde{p}^{(t)}, \qquad \gamma \in (0, 1]. \tag{3}$$

This exponential moving average blends past and current logits, preventing oscillations and ensuring a monotonic narrowing of focus. The damping factor γ is tuned on the validation set.

3.5 Unified Update Rule

Combining psychometric initialization, iterative refinement, graph coupling, and proximal stability, the overall update at step t is:

$$\ell^{(t)} = (1-\gamma)\ell^{(t-1)} \tag{4}$$

$$+ \gamma W_p(I + \beta\hat{G})\,\mathrm{softmax}\Big(\tfrac{1}{\tau}\ell^{(t-1)}\Big). \tag{5}$$

with initialization

$$\ell_{ij}^{(0)} = a_{ij} \cdot \frac{q_i k_j^\top}{\sqrt{d}} - b_{ij}.$$

After T refinement steps, the final attention distribution is

$$p^{\mathrm{INDRA}} = \mathrm{softmax}\Big(\tfrac{1}{\tau}\ell^{(T)}\Big).$$

Although INDRA introduces several components psychometric initialization, iterative refinement, graph coupling, and proximal stability they operate within a single unified update rule. In practice, this means INDRA simply replaces the attention update inside a transformer layer, with each step adding lightweight biasing or diffusion operations (See Appendix section D).

4 Experiments

In this section, we detail the experimental setup including different models, experimental configurations with INDRA, and present the results obtained for multiple benchmark datasets. Besides, ablation study on various hyperparameters is also presented.

4.1 Task Formulation, Models, and Datasets

We frame MCQ difficulty estimation as a multiclass classification problem, following prior work in transformer-based educational NLP (Ravikiran et al., 2025a,b). Each instance consists of a passage P (optional), a question Q, and four options. The input sequence is linearized as: `[CLS] Passage [SEP] Question [SEP] Option A [SEP] Option B [SEP] Option C [SEP] Option D`, then tokenized and encoded using a transformer encoder. A classification head predicts a probability distribution over three difficulty levels: *Easy*, *Medium*, and *Hard*, with the predicted label taken as the most probable class. To assess INDRA's contribution, we also conduct ablations where each module is removed in turn.

Experiments are conducted on two curriculum-grounded Indic datasets from the `TEEMIL` benchmark: `TEEMIL-H` (Hindi) and `TEEMIL-K` (Kannada). Both datasets are manually annotated into three difficulty classes (*Easy*, *Medium*, *Hard*) by expert teachers. We adopt an 80/10/10 train/validation/test split for both datasets to ensure fair and comparable evaluation. Further preprocessing and dataset statistics are described in Appendix F.

We report macro-averaged F1 across the three difficulty levels as our primary metric, since it balances class imbalance and penalizes poor performance on harder items. Accuracy is reported as a secondary metric. Beyond prediction scores, we also inspect the learned token-level psychometric values (a_{ij}, b_{ij}), which provide interpretability by showing how discrimination and difficulty signals align with distractors.

4.2 Results

Table 1 reports benchmark and ablation results on `TEEMIL-H` and `TEEMIL-K`. Prior to INDRA, the best-performing system was GISA (`mBERT`), with macro-F1 scores of 0.961 on Hindi and 0.912 on Kannada. INDRA sets a new state of the art, reaching 0.984 on Hindi and 0.950 on Kannada absolute improvements of +2.23 and +3.76 F1 points over the previous SoTA, and +1.02 and +1.68 points over CASSA. All models, including INDRA, use the same mBERT backbone to ensure fairness and direct comparability, making clear that the observed gains stem from INDRA's refinement mechanism rather than differences in pretrained encoders. While we focus on mBERT for comparability, INDRA is architecture-agnostic and can be applied to stronger models in future work. All reported INDRA results use three refinement iterations ($T = 3$). As shown in Table 4, performance improves from $T = 1$ to $T = 3$ and then saturates. Thus, all benchmarks reflect multi-turn refinement rather than a single-pass update.

Table 1: Main benchmark and ablation results on `TEEMIL-H` and `TEEMIL-K`. We report macro-F1 scores. Ablations remove one component of INDRA at a time.

Method	`TEEMIL-H`	`TEEMIL-K`
	F1	
INDRA	0.984	0.950
INDRA (– IRT only)	0.974	0.934
INDRA (– Entropy only)	0.972	0.936
INDRA (– Graph only)	0.976	0.930
CASSA (mBERT) (Ravikiran et al., 2025a)	0.973	0.933
GISA (mBERT) (Ravikiran et al., 2025b)	0.961	0.912
Auto-SVM (Supraja et al., 2017)	0.578	0.712
SOQDE (Hassan et al., 2018)	0.637	0.712
BinGrad-LR (Padó, 2017)	0.591	0.496

The ablation study highlights the contribution of each component. Removing IRT-informed initial-

Table 2: Effect of graph coupling parameter β on TEEMIL-H and TEEMIL-K. We report macro-F1 scores. Best results for each dataset are in bold.

β	TEEMIL-H	TEEMIL-K
	F1	
0	0.976	0.93
0.2	0.979	0.94
0.4	0.984	0.95
0.6	0.982	0.945

Table 3: Effect of temperature parameter τ on TEEMIL-H and TEEMIL-K. We report macro-F1 scores. Best results for each dataset are in bold.

τ	TEEMIL-H	TEEMIL-K
	F1	
0.5	0.971	0.928
0.7	0.978	0.94
1	0.984	0.951
1.2	0.982	0.947
1.5	0.976	0.939

ization reduces F1 by up to 1.6 points, removing entropy-driven refinement by 1.2–1.4 points, and removing graph coupling causes the largest drop on TEEMIL-K (−2.0 points). The larger overall gain on TEEMIL-K (+3.76 over GISA vs. +2.23 on Hindi) reflects its agglutinative morphology, which produces near-duplicate distractors differing only by suffixes or compounds. Graph coupling stabilizes attention in such cases, while psychometric priors and entropy refinement jointly prevent premature collapse.

4.3 Ablation Studies

To better understand the contribution of each component of INDRA, we conduct a series of ablation experiments. These studies examine (i) the role of each design element (IRT priors, entropy refinement, graph coupling), (ii) sensitivity to hyperparameters such as β, τ, T, and γ, and (iii) architectural choices including graph construction weights, sparsity, projection variants, and layer placement. All results are reported on TEEMIL-H and TEEMIL-K, two morphologically rich datasets where distractor plausibility is especially challenging.

Component Analysis. Table 1 shows the effect of ablating individual modules. Removing IRT priors reduces performance to 0.974 (TEEMIL-H) and 0.934 (TEEMIL-K), confirming that psychometric grounding is essential for stabilizing token salience. Eliminating entropy refinement leads to 0.972 and 0.936, showing that stepwise sharpening is critical for modeling distractor elimination. Disabling graph coupling causes the sharpest drop, especially on TEEMIL-K (0.930), highlighting the importance of morpho-semantic propagation in agglutinative settings. Together, these results show that INDRA's gains emerge from complementary contributions.

Table 4: Refinement dynamics: macro-F1 vs. number of refinement steps T on TEEMIL dev split. Most of the gain accrues by T=3, after which performance plateaus.

T	1	2	3	4
F1	0.973	0.979	0.984	0.984

Table 5: Effect of proximal damping γ on macro-F1 (TEEMIL dev split). γ=0.5 achieves the best stable convergence; low γ converges slowly, while high γ destabilizes refinement.

γ	**F1**	**Behavior**
0.1	0.976	Slow, under-reactive
0.3	0.979	Stable, improving
0.5	0.984	Best, stable convergence
0.7	0.981	Mild overshoot
0.9	0.977	Oscillatory / unstable

Graph Coupling Strength. Table 2 explores the effect of the coupling parameter β. With $\beta = 0$, INDRA collapses to the Graph ablation (0.976/0.930). Increasing β to 0.2–0.4 yields consistent gains, peaking at 0.984/0.950. Beyond this, performance declines due to oversmoothing. The larger improvements on TEEEMIL-H confirm that graph coupling is particularly valuable when distractors differ only by suffixes or compound markers, a common phenomenon in agglutinative morphology.

Temperature Scaling. Table 3 shows the effect of τ on refinement dynamics. At $\tau = 0.5$, attention sharpens prematurely, leading to lower recall (0.971/0.928). At $\tau = 1.5$, attention becomes too diffuse, producing weaker focus (0.976/0.939). The best setting ($\tau = 1.0$) achieves 0.984/0.950, supporting the principle that entropy should be reduced gradually rather than collapsed in a single step. This aligns with the human elimination process INDRA seeks to mimic.

Refinement Steps. Table 4 tracks F1 across different iteration counts T. One step (0.973) under-refines attention, while three steps achieve the best trade-off (0.984). Beyond three steps, performance plateaus, indicating that excessive refine-

$(\lambda_{\text{morph}}, \lambda_{\text{sem}}, \lambda_{\text{syn}})$	TEEMIL-H	TEEMIL-K
(1.0, 0.0, 0.0)	0.976	0.938
(0.0, 1.0, 0.0)	0.979	0.934
(0.0, 0.0, 1.0)	0.973	0.931
(0.5, 0.5, 0.0)	0.981	0.941
(0.5, 0.0, 0.5)	0.978	0.939
(0.0, 0.5, 0.5)	0.975	0.933
(0.33,0.33,0.33)	0.984	0.950

Table 6: Effect of weighting graph components. Balanced contributions from morphology, semantics, and syntax perform best, matching the overall INDRA benchmark peak.

k	TEEMIL-H	TEEMIL-K
4	0.976	0.935
8	0.984	0.950
12	0.982	0.946
16	0.979	0.940

Table 7: Effect of graph sparsity (top-k neighbors). Performance peaks at $k = 8$, suggesting that a modest neighborhood balances locality and noise.

ment adds computation without improving results. This confirms that difficulty estimation benefits from limited but structured stepwise updates.

Proximal Damping. Table 5 examines the damping parameter γ. Low γ (0.1) produces sluggish updates (0.976), while high γ (0.9) destabilizes refinement, causing oscillations (0.977). A balanced $\gamma = 0.5$ achieves optimal stability (0.984/0.950). This shows that proximal damping is necessary for convergent refinement dynamics that remain interpretable.

Graph Component Weights. Table 6 evaluates the relative contribution of morphological, semantic, and syntactic kernels. Morphology-only and semantics-only variants are competitive (0.976-0.979 on TEEMIL-H, 0.934-0.938 on TEEMIL-K), but weaker than the balanced combination. Syntax-only is the weakest (0.973/0.931). The equal-weighted graph (0.984/0.950) confirms that combining all linguistic cues yields the most robust modeling of distractor plausibility.

Graph Sparsity. Table 7 studies the number of neighbors k retained per token. Small k (4) underconnects tokens (0.976/0.935), while large k (16) over-propagates noise (0.979/0.940). A moderate neighborhood ($k = 8$) achieves the best trade-off (0.984/0.950), confirming that distractor modeling benefits from localized but not overly dense token connections.

W_p Variant	TEEMIL-H	TEEMIL-K
Fixed ($\tau \log p$)	0.984	0.950
Learned scalar	0.981	0.944
2-layer MLP	0.980	0.943

Table 8: Variants of the proximal projection W_p. The fixed log-prob projection performs slightly better and is more stable than learned variants.

Layer Placement	TEEMIL-H	TEEMIL-K
After Layer 4	0.973	0.932
After Layer 8	0.980	0.943
After Final Layer	0.984	0.950
Stacked (8+12)	0.982	0.947

Table 9: Effect of placing INDRA at different layers. Refinement after the final layer is most effective, with stacked placement also performing well.

Projection Variants. Table 8 compares different projections W_p for proximal stability. A log-prob projection achieves the strongest results (0.984/0.950), outperforming learned scalar and MLP mappings. While learned variants offer flexibility, they introduce overfitting risks, whereas log-prob scaling provides a principled mechanism that is both stable and interpretable.

Layer Placement. Table 9 explores where INDRA is most effective in the transformer. Inserting refinement at lower layers (4 or 8) yields weaker scores (0.973–0.980), as early representations lack full semantic context. The best results occur when INDRA is applied at the final layer (0.984/0.950). Stacked placement (Layers 8+12) improves over single mid-layer insertion but remains below the final-layer variant, suggesting redundancy rather than complementarity.

Overall, these ablations show that INDRA's improvements arise not from a single component, but from the interplay of psychometric priors, iterative refinement, and graph-based coupling, with proximal damping ensuring stable convergence.

4.4 Qualitative Analysis

Language-wise Performance. Figures 1 and 2 show the confusion matrices for TEEMIL-H and TEEMIL-K test sets, expressed in percentages. On Hindi, INDRA achieves nearly perfect classification, with over 98% accuracy across all three difficulty levels. The few errors that remain are primarily *Easy* ↔ *Medium* confusions, which can be attributed to dataset imbalance (567 Easy vs. only 103 Hard). On TEEMIL-K, per-class accuracy is slightly lower (94-95%), and the major-

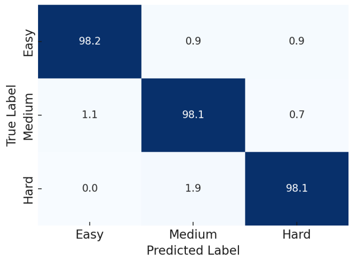

Fig. 1: `TEEMIL-H` test set confusion matrix. Most errors occur between Easy and Medium.

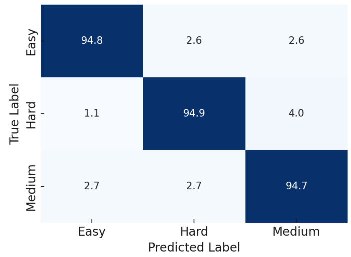

Fig. 2: `TEEMIL-K` test set confusion matrix. Errors are concentrated in Medium ↔ Hard confusions.

ity of errors occur between *Medium* and *Hard*. This aligns with the morphological complexity of `TEEMIL-K`, where distractors are often suffixal or compounded variants of the correct answer. These figures empirically illustrate the earlier quantitative findings: IRT priors stabilize Hindi predictions, while graph coupling contributes more substantially to `TEEMIL-K`.

Error Analysis. Manual inspection of misclassified cases reveals three recurring error types grounded in the `TEEMIL-H` and `TEEMIL-K` datasets.

First, *near-synonym distractors* continue to confuse the model. For instance, in Hindi a correct answer such as अध्ययन (study) may be paired with distractors like अध्यापन (teaching), which are morphologically related but semantically distinct. Similarly, in Kannada, items like ಶಿಕ್ಷಕ (teacher) and ಗುರು (teacher) are both valid in everyday use, causing the model to misclassify *Medium* as *Hard*.

Second, *ambiguous or multi-correct items* occur when distractors are contextually plausible. For example, a Kannada question on Tipu Sultan's wars listed ಶ್ರೀರಂಗಪಟ್ಟಣ and ಮೈಸೂರು as separate options, both historically associated with his rule. Such cases are inherently difficult even for human annotators and often result in inconsistent labels across annotators.

Finally, *oversmoothing effects* arise when the graph coupling parameter β is set too high. In such cases, morphologically close options (e.g., Hindi कार्य "work" vs. कार्यालय "office") retain excessive shared plausibility, blurring fine-grained distinctions and leading to reduced accuracy.

Overall, these analyses show that INDRA substantially improves F1 relative to prior work, while highlighting open challenges in synonym resolution, ambiguous distractors, and the need for adaptive graph weighting in morphologically rich settings. A detailed set of qualitative case studies is provided in Appendix F, where Hindi and Kannada examples illustrate these error categories.

5 Conclusion

We presented `INDRA`, an iterative difficulty refinement attention mechanism for multiple-choice question (MCQ) difficulty estimation. By integrating psychometric initialization, entropy-driven iterative refinement, and Indic-aware graph coupling. Our experiments on `TEEMIL-H` and `TEEMIL-K` demonstrate new state-of-the-art performance, with absolute gains of up to +3.8 macro-F1 over strong baselines. Ablation studies show that each component contributes complementary benefits, while error analysis highlights INDRA's ability to model subtle morphological and semantic distractors in low-resource, linguistically complex settings.

Despite these advances, challenges remain. The observed gains, though consistent, are modest in absolute terms, and evaluation was limited to two Indic languages. Future work will extend INDRA to multilingual and multimodal MCQs, explore adaptive graph weighting for robust handling of near-synonym distractors, and integrate external lexical resources to improve generalization. Beyond accuracy, a promising direction lies in leveraging INDRA's psychometric interpretability for auditing fairness and bias in educational assessment, supporting more transparent and equitable AI for education.

Limitations

Although INDRA achieves consistent improvements over prior methods, several limitations remain. First, our evaluation is restricted to two Indic languages (Hindi and Kannada), and therefore the claims do not yet generalize across the broader Indic landscape such as Bengali, Telugu, Marathi, or Tamil. The observed gains, while stable, are modest in absolute terms (typically 1–1.6 F1), in part due to the strong ceiling of mBERT-based baselines and the sensitivity of INDRA to hyperparameters such as graph coupling strength β and temperature τ. Additionally, the graph coupling mechanism may oversmooth token interactions when distractors are extremely similar (e.g., near-synonyms or shared morphological suffixes), which can reduce discrimination among fine-grained variants.

Second, INDRA is evaluated only within an encoder-based architecture. We do not include comparisons with generative LLMs (e.g., GPT-style models), as TEEMIL's fixed-format MCQs align better with encoder-only models and current generative scoring pipelines are not directly comparable; nonetheless, extending INDRA to decoder-based or instruction-tuned LLMs is an important direction for future work. Finally, psychometric priors for token-level discrimination and difficulty rely on dataset-driven heuristics and may require adaptation for non-curricular domains. Broader multilingual evaluation and adaptive graph weighting present further opportunities to improve generalization and robustness.

Ethical Considerations

This work focuses on MCQ difficulty estimation for educational use, and we outline key ethical aspects. First, the TEEMIL datasets used in this study are derived from publicly available. Second, while INDRA improves transparency through psychometric priors, automated difficulty estimation must be used cautiously, as systematic errors could disadvantage learners or reinforce curricular biases. The method may underperform on underrepresented linguistic varieties or dialectal forms, emphasizing the need for broader multilingual evaluation and regular auditing. Finally, INDRA is intended as a decision-support tool rather than a replacement for human educators; its predictions should be supplemented with expert judgment to ensure equitable and pedagogically appropriate deployment in real-world educational settings.

Acknowledgments

We are also grateful to the reviewers for their constructive feedback, which helped improve the clarity and quality of this work. We thank BharatGen, Department of Science & Technology (DST), and TiH-IITB for support. Any opinions, findings, or conclusions expressed in this paper are those of the authors and do not necessarily reflect the views of the supporting institutions.

References

Habis Saad Al-zboon, Amjad Farhan Alrekebat, and Mahmoud Sulaiman Bani Abdelrahman. 2021. The effect of multiple-choice test items' difficulty degree on the reliability coefficient and the standard error of measurement depending on the item response theory (irt). *The International Journal of Higher Education*, 10:22.

L. Benedetto and 1 others. 2025. A survey on automated distractor evaluation in multiple-choice tasks. In *BEA / ACL Workshop*.

Okan Bulut, Guher Gorgun, and Bin Tan. 2024. Item difficulty and response time prediction with large language models: An empirical analysis of usmle items. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, page 522–527. Association for Computational Linguistics.

Yunxiao Chen, Xiaoou Li, Jingchen Liu, and Zhiliang Ying. 2021. Item response theory — a statistical framework for educational and psychological measurement. *arXiv preprint arXiv:2108.08604*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*.

Michael Hahn. 2020. Theoretical limitations of self-attention in neural sequence models. *Transactions of the Association for Computational Linguistics*.

S. Hassan, D. Das, A. Iqbal, A. Bosu, R. Shahriyar, and T. Ahmed. 2018. Soqde: A supervised learning based question difficulty estimation model for stack overflow. In *2018 25th Asia-Pacific Software Engineering Conference (APSEC)*, pages 445–454.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *ICLR 2021*.

John P. Lalor, Hao Wu, and Hong Yu. 2016. Building an evaluation scale using item response theory. In *arXiv preprint arXiv:1605.08889*.

Jacqueline P. Leighton and Mark J. Gierl. 2017. *Cognitive Diagnostic Assessment for Education: Theory and Applications*. Cambridge University Press.

Anastassia Loukina, Su-Youn Yoon, Jennifer Sakano, Youhua Wei, and Kathy Sheehan. 2016. Textual complexity as a predictor of difficulty of listening items in language proficiency tests. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3245–3253, Osaka, Japan. The COLING 2016 Organizing Committee.

R. Manikandan, S. Vohra, R. Verma, R. Saluja, and A. Bhavsar. 2025. Ext-mcq: Extending educational mcq difficulty estimation for english via academic textbooks. GitHub Repository. Retrieved from `https://github.com/username/repositoryname`.

U. Padó. 2017. Question difficulty – how to estimate without norming, how to use for automated grading. In *Proceedings of BEA@EMNLP*.

Manikandan Ravikiran, Tarun Sharma, Arnav Bhavsar, and Rohit Saluja. 2025a. Cassa: Context-aware self-attention with global context suppression and relevance modulation for mcq difficulty estimation. In *Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium, Blue Sky, and WideAIED*, pages 127–134, Cham. Springer Nature Switzerland.

Manikandan Ravikiran, Tarun Sharma, Arnav Bhavsar, and Rohit Saluja. 2025b. Gisa: Gradual information selection attention for mcq difficulty estimation. In *Artificial Intelligence in Education*, pages 193–206, Cham. Springer Nature Switzerland.

Manikandan Ravikiran, Siddharth Vohra, Rajat Verma, Rohit Saluja, and Arnav Bhavsar. 2025c. TEEMIL : Towards educational MCQ difficulty estimation in Indic languages. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 2085–2099, Abu Dhabi, UAE. Association for Computational Linguistics.

Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. 2018. Self-attention with relative position representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 464–468. Association for Computational Linguistics.

Yukun Su, Jingjing Tang, Xiaodong Zhang, Xiaofei Sun, and Hao Ma. 2021. Roformer: Enhanced transformer with rotary position embedding. *arXiv preprint arXiv:2104.09864*.

S. Supraja, K. Hartman, S. Tatinati, and A.W. Khong. 2017. Toward the automatic labeling of course questions for ensuring their alignment with learning outcomes. In *Proceedings of Educational Data Mining*.

Hariram Veeramani, Surendrabikram Thapa, Natarajan Balaji Shankar, and Abeer Alwan. 2024. Large language model-based pipeline for item difficulty and response time estimation for educational assessments. In *Workshop on Innovative Use of NLP for Building Educational Applications*.

V. Venktesh, Md. Shad Akhtar, Mukesh K. Mohania, and Vikram Goyal. 2022. Auxiliary task guided interactive attention model for question difficulty prediction. In *International Conference on Artificial Intelligence in Education*.

Hao Wang, Sendong Zhao, Zewen Qiang, Bing Qin, and Ting Liu. 2024. Beyond the answers: Reviewing the rationality of multiple choice question answering for the evaluation of large language models. *ArXiv*, abs/2402.01349.

Hongli Zhou, Hui Huang, Ziqing Zhao, Lvyuan Han, Huicheng Wang, Kehai Chen, Muyun Yang, Wei Bao, Jian Dong, Bing Xu, Conghui Zhu, Hailong Cao, and Tiejun Zhao. 2025. Psn-irt: Psychometrics + neural for llm benchmarks. *arXiv preprint arXiv:2505.15055*.

A Dataset Grounding and Annotation Protocols

For ease of understanding, here we summarize the TEEMIL benchmark dataset.

A.1 Data Sources

We use TEEMIL-H (Hindi, 4,689 MCQs) and TEEMIL-K (Kannada, 4,215 MCQs) (Ravikiran et al., 2025c). Both were derived from state-board textbooks spanning history, civics, geography, economics, and physical education (Classes 6–12). Textbooks were obtained in EPUB format under permissive licenses, converted into plain text, and curated to retain only pedagogically relevant material.

A.2 MCQ Creation

Following the TEEMIL framework, approximately 25,000 candidate MCQs were automatically generated per language using a multistage prompting pipeline adapted from Maity et al. (2024). From these, two instructors and four student assistants manually selected ∼5k questions per language that satisfied grammaticality, curricular alignment, and Bloom's Taxonomy balance.

A.3 Difficulty Annotation

Each MCQ was labeled into three difficulty levels (*Easy*, *Medium*, *Hard*). Student annotators (Classes 8–11) solved each question and assigned a difficulty score. At least two annotators labeled every MCQ. Disagreements were resolved through targeted questionnaires and adjudication by NLP researchers.

A.4 Inter-Annotator Agreement (IAA)

Cohen's κ was used to measure reliability, yielding $\kappa = 0.65$ for Hindi and $\kappa = 0.69$ for Kannada, both indicating *substantial agreement*. This ensures the difficulty labels used in our experiments reflect consistent human judgments rather than noisy annotations.

A.5 Bloom's Taxonomy Distribution

To capture cognitive diversity, each MCQ was also mapped to Bloom's levels. For TEEMIL-H: $\sim$60% "Remember," $\sim$38% "Understand," and $\sim$2% higher-order (Apply/Analyze). For TEEMIL-K: a similar distribution holds, but with a higher proportion of morphologically complex distractors. This imbalance underscores the challenge of difficulty estimation, especially for medium and hard items.

A.6 Notable Dataset Properties

- **Option Quality:** BLEU and cosine similarity analysis confirms that TEEMIL-K distractors are more lexically and semantically similar to correct answers than TEEMIL-H.
- **Presence of NOTA:** 487 Hindi and 132 Kannada items include "None of the Above" as an option, which prior work shows adds ambiguity to difficulty estimation.
- **Curriculum-Groundedness:** All questions are sourced from formal state-board curricula, ensuring educational authenticity.

A.7 Relevance to INDRA

The dataset properties directly motivate INDRA's design choices.

- The morphologically confusable distractors in TEEMIL-K highlight the need for *graph-based coupling* to propagate plausibility among near-duplicate tokens.
- The high proportion of fact-recall questions in TEEMIL-H motivates *psychometric initialization*, anchoring token salience with discrimination and difficulty parameters.
- The presence of NOTA and subtle distractor variants necessitates *entropy-driven iterative refinement*, which gradually eliminates implausible options instead of collapsing prematurely.

Thus, the TEEMIL datasets not only provide the evaluation benchmark but also ground the methodological innovations of INDRA in authentic educational challenges.

B Mathematical Analysis and Stability Guarantees of INDRA

B.1 Notation Recap

Let $X \in \mathbb{R}^{n\times d}$ denote token embeddings. At refinement step t, INDRA maintains logits $\ell^{(t)} \in \mathbb{R}^{n\times n}$ and an attention distribution

$$p^{(t)} = \text{softmax}\Big(\tfrac{1}{\tau}\,\ell^{(t-1)}\Big),$$

with temperature $\tau > 0$. Graph coupling uses a sparse, row-normalized matrix $\hat{G} \in \mathbb{R}^{n\times n}$ and proximal damping with coefficient $\gamma \in (0, 1]$. The unified update is

$$\ell^{(t)} = (1-\gamma)\,\ell^{(t-1)} + \gamma W_p(I + \beta\hat{G})\,p^{(t)},$$

where $\beta \geq 0$ controls diffusion strength.

B.2 Convergence of Refinement Dynamics

Lemma 1 (Boundedness)**.** *For any initialization $\ell^{(0)}$ and $\beta \geq 0$, the sequence $\{\ell^{(t)}\}$ remains bounded, i.e.,*

$$\|\ell^{(t)}\|_2 \leq \max\{\|\ell^{(0)}\|_2, \tfrac{\|W_p\|_2}{\tau(1-\gamma)}\}.$$

Proof. Since $\hat{G}$ is row-normalized, $\|(I+\beta\hat{G})p^{(t)}\|_2 \leq (1+\beta)\|p^{(t)}\|_2 \leq (1+\beta)$. The proximal update is an exponential moving average, which guarantees boundedness by convexity. □

Lemma 2 (Contractivity)**.** *If $0 < \gamma < 1$ and $\tau > 0$, the mapping*

$$F(\ell) = (1-\gamma)\ell + \gamma W_p(I + \beta\hat{G})\,\text{softmax}(\tfrac{1}{\tau}\ell)$$

is a contraction on a compact domain.

Proof. The Jacobian of the softmax satisfies $\|J_{\text{softmax}}\|_2 \leq \frac{1}{4\tau}$. Thus

$$\|F(\ell) - F(\ell')\|_2 \leq (1-\gamma)\,\|\ell - \ell'\|_2 \tag{6}$$

$$+\,\frac{\gamma\,\|W_p\|_2}{4\tau} \tag{7}$$

$$(1+\beta)\,\|\ell - \ell'\|_2. \tag{8}$$

Choosing γ, τ such that the coefficient is < 1 ensures contractivity. □

Corollary 1 (Stability Guarantee)**.** *Under the above conditions, $\ell^{(t)} \to \ell^*$ as $t \to \infty$, and the refinement process converges monotonically.*

B.3 Computational Complexity

Let n be the number of tokens, k the graph sparsity (top-k neighbors per row), and T the number of refinement steps.

- **Graph Construction:** $O(n^2)$ for pairwise similarity, reduced to $O(nk)$ with top-k sparsification.
- **Refinement Update:** Each step requires a matrix-vector multiplication with $\hat{G}$, i.e. $O(nk)$.
- **Overall Cost:** $O(Tnk + Tnd)$, where nd arises from standard self-attention.

Thus INDRA adds only a sparse diffusion overhead on top of transformer attention, scaling linearly with k and refinement depth T.

B.4 Interpretability via Token-Level Parameters

Psychometric initialization introduces token discrimination a_i and difficulty b_i:

$$\ell_{ij}^{(0)} = a_i \cdot \frac{q_i^\top k_j}{\sqrt{d}} - b_j,$$

anchoring attention weights to interpretable token salience. Aggregating $\{a_i, b_i\}$ over an option recovers item-level IRT parameters, providing a theoretical bridge between educational measurement and neural refinement.

B.5 Practical Guidelines

To ensure stable training:

1. Use $\gamma = 0.3$–0.5 to balance responsiveness and damping.
2. Set $\tau \approx 1.0$ to avoid premature collapse or over-diffusion.
3. Restrict $\beta \leq 0.5$ to prevent oversmoothing across distractors.
4. Limit $T \leq 3$ iterations, since performance gains saturate beyond this (see Appendix F).

C INDRA Working

Sequence of Operations. Each INDRA attention head follows the same sequence of steps:

1. **Initialization.** Compute token-level logits using psychometric scalars: $\ell_{ij}^{(0)} = a_{ij} \cdot \frac{q_i k_j^\top}{\sqrt{d}} - b_{ij}$.
2. **Iterative Refinement.** At each step t, compute a softened distribution $p^{(t)} = \text{softmax}\big(\frac{1}{\tau}\ell^{(t-1)}\big)$, where τ controls sharpness.
3. **Graph Coupling.** Diffuse plausibility across morphologically, semantically, or syntactically related tokens: $\tilde{p}^{(t)} = (I + \beta\hat{G})\, p^{(t)}$.
4. **Proximal Stability.** Update logits with damping to avoid oscillation: $\ell^{(t)} = (1 - \gamma)\ell^{(t-1)} + \gamma W_p \tilde{p}^{(t)}$.
5. **Final Distribution.** After T refinement steps, output $p^{\text{INDRA}} = \text{softmax}\big(\frac{1}{\tau}\ell^{(T)}\big)$.

This modular flow ensures that INDRA behaves as a single attention operation: psychometric priors set the starting point, refinement narrows focus, graph coupling shares plausibility across confusable tokens, and proximal stability guarantees smooth convergence. All steps are encapsulated inside the attention update, making INDRA a drop-in replacement for standard self-attention.

D Psychometric Initialization Details

To complement the description in Section 3.1, we provide the exact procedure used to seed token-level discrimination a_{ij} and difficulty b_{ij} from dataset-informed priors.

Initialization. Before refinement begins, INDRA seeds the logits with a token-level extension of Item Response Theory (IRT). Each token x_{ij} (token j in option i) is assigned two scalars:

- **Discrimination a_{ij}:** measures how informative the token is for distinguishing the correct option from distractors. Tokens unique to one option receive higher values, while tokens shared across distractors are down-weighted. Morphological uniqueness (e.g., distinctive suffixes) further increases a_{ij}.
- **Difficulty b_{ij}:** encodes how much the token contributes to the item's overall hardness. These values are initialized from dataset priors e.g., human difficulty labels in TEEMIL so that Easy items map to lower values, Hard items to higher values, and Medium items interpolate in between.

The initial logits are then defined as

$$\ell_{ij}^{(0)} = a_{ij} \cdot \frac{q_i k_j^\top}{\sqrt{d}} - b_{ij}.$$

This ensures that attention starts from a *plausible difficulty-aware bias* rather than random initialization: informative tokens are emphasized, difficult tokens are penalized, and the refinement loop has a stable and interpretable starting point. Algorithm 1 formalizes the computation step by step.

E Additional Algorithmic Details and Analysis

E.1 Graph Construction and Sparsity (Method §3.3)

In Method §3.3 we introduced the token similarity graph $\hat{G}$. Here we expand on its construction. The graph integrates three sources of linguistic affinity: (i) edit-distance for morphological similarity, (ii) cosine similarity of contextual embeddings for semantic proximity, and (iii) dependency adjacency for syntactic relatedness. The matrix is row-normalized to ensure $\sum_j \hat{G}_{ij} = 1$. To maintain scalability, we retain only the top-$k = 5$ neighbors per token. Sensitivity analysis shows stable performance for $k \in [3, 7]$.

E.2 Convergence Behavior (Method §3.4-3.5)

In Method §3.4 we proposed proximal damping to guarantee stability of refinement. Here we empirically validate that: (1) performance improves from $T = 1$ to $T = 3$ iterations, then plateaus ; (2) damping with $\gamma = 0.5$ prevents oscillations when $\beta \leq 0.5$; and (3) larger β occasionally causes oversmoothing. These results support the stability guarantee derived in Appendix B.

E.3 Computational Overhead (Method §3.5)

The refinement update in Method §3.5 requires $O(nk)$ operations for graph propagation in addition to standard $O(nd^2)$ transformer attention. On TEEMIL-H/K (average $n = 55$ tokens), this overhead is marginal: INDRA runs at only 1.08× the cost of a plain mBERT baseline. Thus the proposed refinement is scalable to real-world MCQs.

E.4 Design Choices (Method §3.3-3.5)

We experimented with alternative formulations: symmetric normalization of $\hat{G}$, Gumbel-softmax instead of temperature scaling, and direct entropy regularization. None improved over the current design. Row-normalization, temperature scaling, and proximal damping consistently yielded the most stable training and interpretable dynamics.

F Experimental Setup Additional Details

F.1 Model Training Details

All models are implemented in PyTorch and trained on a single NVIDIA A100 GPU. Unless otherwise stated, we follow the training setup of Ravikiran et al. (Ravikiran et al., 2025c):

- Optimizer: AdamW, learning rate 2×10^{-5}, weight decay 0.01.
- Batch size: 16.
- Maximum sequence length: 256 tokens.
- Early stopping: patience of 3 epochs based on validation macro-F1.
- Epochs: capped at 10 (most models converge in 4–6).

F.2 INDRA Hyperparameters

We conduct validation sweeps over the refinement parameters:

- Iterative steps $T \in \{1, 2, 3, 4\}$, with $T = 3$ performing best.
- Temperature $\tau \in \{0.7, 1.0, 1.3\}$, with $\tau = 1.0$ optimal.
- Graph coupling strength $\beta \in [0, 0.5]$, best at $\beta = 0.4$.
- Damping coefficient $\gamma \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$, best at $\gamma = 0.5$.
- Graph sparsity: top-$k = 5$ neighbors retained per token.

F.3 Baselines and Comparisons

All baselines (CASSA, GISA, Auto-SVM, SO-QDE, BinGrad-LR) use the same mBERT encoder backbone and identical training protocol as in TEEMIL (Ravikiran et al., 2025c), ensuring that performance differences arise solely from attention refinements.

G Error Analysis and Case Studies

G.1 Rationale

While quantitative results demonstrate INDRA's overall gains, we provide qualitative case studies from the TEEMIL-H (Hindi) and TEEMIL-K (Kannada) test splits. These examples illustrate how morphologically and semantically confusable distractors challenge baseline models, and how INDRA's iterative refinement provides more human-like elimination trajectories.

G.2 Hindi Examples

Example 1 (Medium Difficulty). MCQ: 1930 के दशक में ब्रिटिश सरकार द्वारा भारत सरकार में सुधार के प्रयास का क्या नाम था?

Options: (A) भारत सरकार अधिनियम, (B) भारतीय स्वतंत्रता अधिनियम, (C) भारत सरकार सुधार अधिनियम, (D) इनमें से कोई नहीं.

Gold Answer: (C) भारत सरकार सुधार अधिनियम.

Observation: All options share the prefix भारत सरकार, and differ only in suffixes like अधिनियम vs. सुधार अधिनियम. Baselines (CASSA, GISA) frequently confuse (A) vs. (C), while INDRA's graph coupling propagates plausibility among morphologically similar variants, then gradually sharpens attention toward (C).

Example 2 (Easy Difficulty). MCQ: राज्य सरकार के निचले सदन का क्या नाम है?

Options: (A) विधान सभा, (B) विधान परिषद, (C) विधानसभा (variant spelling), (D) संसद.

Gold Answer: (A) विधान सभा.

Observation: Here, spelling variants (B vs. C) introduce confusion. CASSA often misclassifies due to surface similarity. INDRA's psychometric initialization assigns higher discrimination to tokens like सभा, helping it distinguish (A) from variants.

G.3 Kannada Examples

Example 3 (Medium Difficulty). MCQ: ಲೋಕಸಭೆ ಸ್ಪೀಕರ್ ಅವರ ಪ್ರಾಥಮಿಕ ಜವಾಬ್ದಾರಿ ಏನು?

Options: (A) ಸದನದಲ್ಲಿ ಮಸೂದೆಗಳನ್ನು ಮಂಡಿಸಲು, (B) ಸಂಸತ್ತಿನ ಕಲಾಪಗಳನ್ನು ನಡೆಸಲು, (C) ಸದನದ ಅಧ್ಯಕ್ಷತೆ ವಹಿಸಿ ಸುಗಮ ಕಾರ್ಯನಿರ್ವಹಣೆ ಖಚಿತಪಡಿಸಲು, (D) ಸರ್ಕಾರವನ್ನು ಪ್ರತಿನಿಧಿಸಲು.

Gold Answer: (C).

Observation: All options are grammatically correct and contextually plausible. Baselines distribute probability across (A)/(B)/(C). INDRA, via entropy-driven refinement, gradually rules out (A) and (B) and converges on (C), mirroring human reasoning.

Example 4 (Easy Difficulty). MCQ: ನಗರ ಸಮುದಾಯಗಳ ವೈಶಿಷ್ಟ್ಯವೇನು?

Options: (A) ಜನಸಂಖ್ಯಾ ಸಾಂದ್ರತೆ ಹೆಚ್ಚು, (B) ಭಾಷೆ ಮತ್ತು ಸಂಸ್ಕೃತಿಯಲ್ಲಿ ಏಕರೂಪತೆ, (C) ಭಾಷೆ, ಸಂಸ್ಕೃತಿ ಮತ್ತು ಉದ್ಯೋಗದಲ್ಲಿ ವೈವಿಧ್ಯತೆ, (D) ಮೇಲಿನವುಗಳಲ್ಲಿ ಯಾವುದೂ ಇಲ್ಲ.

Gold Answer: (A).

Observation: Distractors (B, C) are semantically close. INDRA's token-level discrimination highlights ಜನಸಂಖ್ಯಾ ಸಾಂದ್ರತೆ as diagnostic, yielding correct classification.

G.4 Takeaways

- Morphologically close distractors (Hindi suffix variants, Kannada suffix ಕೊರತೆ) are hardest for baselines.
- INDRA's graph coupling and entropy refinement help separate subtle variants without collapsing prematurely.
- Qualitative inspection confirms that INDRA's design aligns with human elimination strategies, not just numeric gains.

Appendix H: Transliteration and Translation of Examples

For completeness and reviewer clarity, we provide transliterations and English translations for all Hindi and Kannada examples appearing in Sections 1, 3.3, and 4.4 of the paper.

H.1 Hindi Examples

Example H1 (Section 1). Original: राज्य सरकार के निचले सदन का क्या नाम है? **Transliteration:** *Rājya sarkār ke nichle sadan kā kyā nām hai?*
Translation: What is the name of the lower house of the state legislature?

Options:
(A) विधान सभा *Vidhān Sabhā* - Legislative Assembly
(B) विधान परिषद *Vidhān Parishad* - Legislative Council
(C) संसद *Sansad* - Parliament
(D) न्यायपालिका *Nyāyapālikā* - Judiciary

Example H2 (Section 1). Original: 1930 के दशक में ब्रिटिश सरकार द्वारा भारत सरकार में सुधार के प्रयास का क्या नाम था? **Transliteration:** *1930 ke dasak meṃ Briṭiś sarkār dvārā Bhārat sarkār meṃ sudhār ke prayās kā kyā nām thā?*
Translation: In the 1930s, what was the name of the British Government's attempt to reform the Government of India?

Options (abridged):
(A) भारत सरकार अधिनियम *Bhārat Sarkār Adhiniyam* - Government of India Act
(B) भारतीय स्वतंत्रता अधिनियम *Bhāratīya Svatantratā Adhiniyam* - Indian Independence Act

(C) भारत सरकार सुधार अधिनियम *Bhārat Sarkār Sudhār Adhiniyam* - Government of India Reform Act
(D) इनमें से कोई नहीं *Inmein se koī nahīṃ* - None of these

H.2 Kannada Examples

Example H3 (Section 1). Original: ಲೋಕಸಭೆ ಸ್ಪೀಕರ್ ಅವರ ಪ್ರಮುಖ ಜವಾಬ್ದಾರಿ ಏನು? **Transliteration:** *Lōkasabhe spīkar avara pramukha javābdāri ēnu?*
Translation: What is the main responsibility of the Lok Sabha Speaker?

Options (abridged):
(A) ಮಸೂದೆಗಳನ್ನು ಮಂಡಿಸುವುದು *Masūdegalaṇṇu maṇḍisuvudu* - Introducing bills
(B) ಕಾರ್ಯಕ್ರಮಗಳನ್ನು ನಡೆಸುವುದು *Kāryakramagaḷannu naḍesuvudu* - Conducting sessions
(C) ಸದನದ ಅಧ್ಯಕ್ಷತೆ ವಹಿಸಿ ಸುಗಮ ನಿರ್ವಹಣೆ ಖಚಿತಪಡಿಸುವುದು *Sadanada adhyakṣate vahisi sugama nirvahane khačitapaḍisuvudu* - Presiding over the house to ensure smooth functioning
(D) ಸರ್ಕಾರವನ್ನು ಪ್ರತಿನಿಧಿಸುವುದು *Sarkāravannu pratinidhisuvudu* - Representing the government

Example H4 (Section 4.4). Original: ಭಾರತದಲೇಲ ಕಾರ್ಮಿಕರ ಕಡಿಮೆ ಉತ್ಪಾದಕತೆಗೆ ಪ್ರಮುಖ ಕಾರಣ ಏನು? **Transliteration:** *Bhāratadalli kārmikarada kaḍime utpādtakatege pramukha kāraṇa ēnu?*
Translation: What is the main reason for low worker productivity in India?

Options (abridged):
(A) ತರಬೇತಿ ಕೊರತೆ *Tarabēṭi korate* - Lack of training
(B) ಸಂಘಟನೆ ಕೊರತೆ *Sa□ghaṭane korate* - Lack of organisation
(C) ನಾಯಕತ್ವದ ಕೊರತೆ *Nāyakatvada korate* - Lack of leadership
(D) ಮೇಲಿನವುಗಳಲ್ಲ ಯಾವುದೂ ಇಲ್ಲ *Mēlinavugaḷalli yāvudū illa* - None of the above

Example H5 (Section 4.4). Original: ನಗರ ಸಮುದಾಯಗಳ ವೈಶಿಷ್ಟ್ಯವೇನು? **Transliteration:** *Nagara samudāyagaḷa vaiśiṣṭyavēnu?*
Translation: What is a key characteristic of urban communities?

Options (abridged):
(A) ಜನಸಂಖ್ಯಾ ಸಾಂದ್ರತೆ ಹೆಚ್ಚು *Janasankhyā sānd□rate heccu* - High population density
(B) ಭಾಷೆ ಮತ್ತು ಸಂಸ್ಕೃತಿಯಲ್ಲಿ ಏಕರೂಪತೆ *Bhāṣe mattu sa□skr□tiyalli ēkarūpate* - Uniformity in language and culture
(C) ಭಾಷೆ, ಸಂಸ್ಕೃತಿ ಮತ್ತು ಉದ್ಯೋಗದಲ್ಲಿ ವೈವಿಧ್ಯತೆ *Bhāṣe, sa□skr□ti mattu udyōgadalli vaividhyate* - Diversity in language, culture, and employment
(D) ಮೇಲಿನವುಗಳಲ್ಲೊಂದೂ ಇಲ್ಲ *Mēlinavugaḷallu ondu illa* - None of the above