Master of Technology in Artificial Intelligence Systems - Capstone Project

By Richard Chai
Email: e0700541@u.nus.edu
https://www.linkedin.com/in/richardchai/

Supervisor:
Dr Wang Aobo (Lecturer & Consultant, Artificial Intelligence Practice)
Email: isswan@nus.edu.sg

CEREBRO INDEX (CI) FRAMEWORK

Question Complexity Measurement Framework, Generation and Scoring System

AGENDA

- THE CORE PROBLEM & RESEARCH QUESTION
- WHY EXISTING METHODS FALL SHORT
- INTRODUCING THE CEREBRO INDEX (CI) FRAMEWORK
- SYSTEM ARCHITECTURE & IMPLEMENTATION
- THE VALIDATION STUDY - METHODOLOGY

AGENDA

- DISCUSSION, LIMITATIONS & FUTURE WORK
- CONCLUSION

THE CORE PROBLEM & RESEARCH QUESTION

Bridging the AI-Human Gap in Question Difficulty Perception

BRIDGING THE AI-HUMAN GAP IN QUESTION DIFFICULTY PERCEPTION

- **Market Impetus**

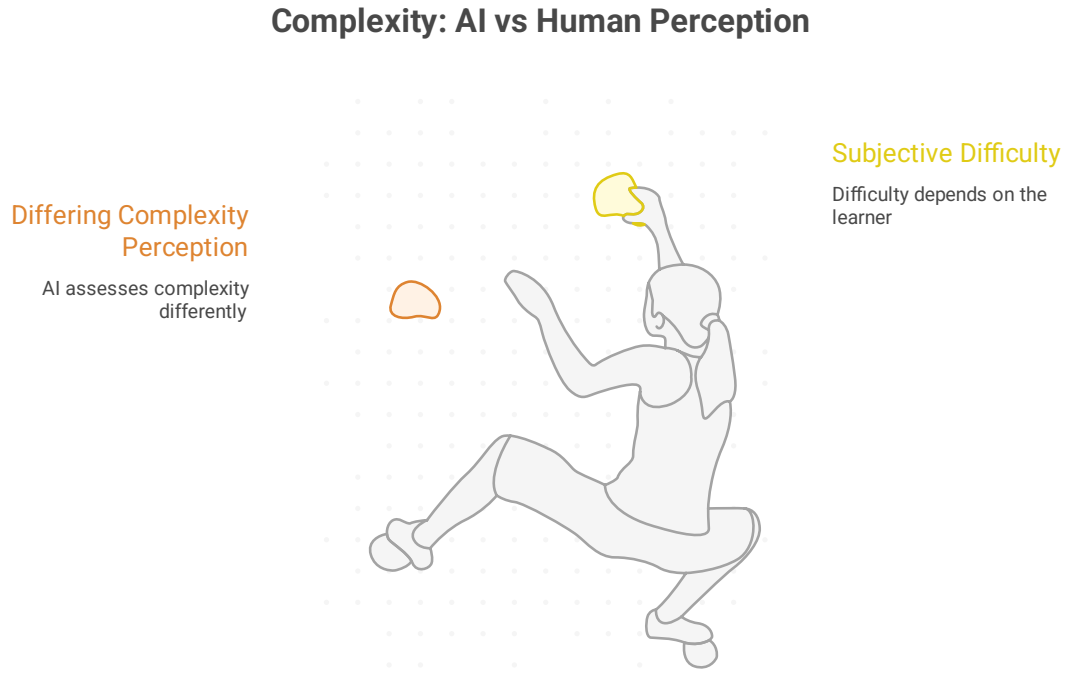
  AI-driven education (adaptive learning, LLM question generation) is growing rapidly.

- **Critical Challenge**

  AI systems interpret 'complexity' differently from humans and from each other too, which can lead to misaligned or inaccurate assessments.

- **Research Question**

  “How can we improve the AI system’s ability to generate questions with complexity levels that consistently align with human expectations of difficulty?”



WHY EXISTING METHODS FALL SHORT

The Limitations of Current Complexity/Difficulty Measurement

THE LIMITATIONS OF CURRENT COMPLEXITY/DIFFICULTY MEASUREMENT

- **Subjective**

  Educator intuition is inconsistent.

- **Rigid**

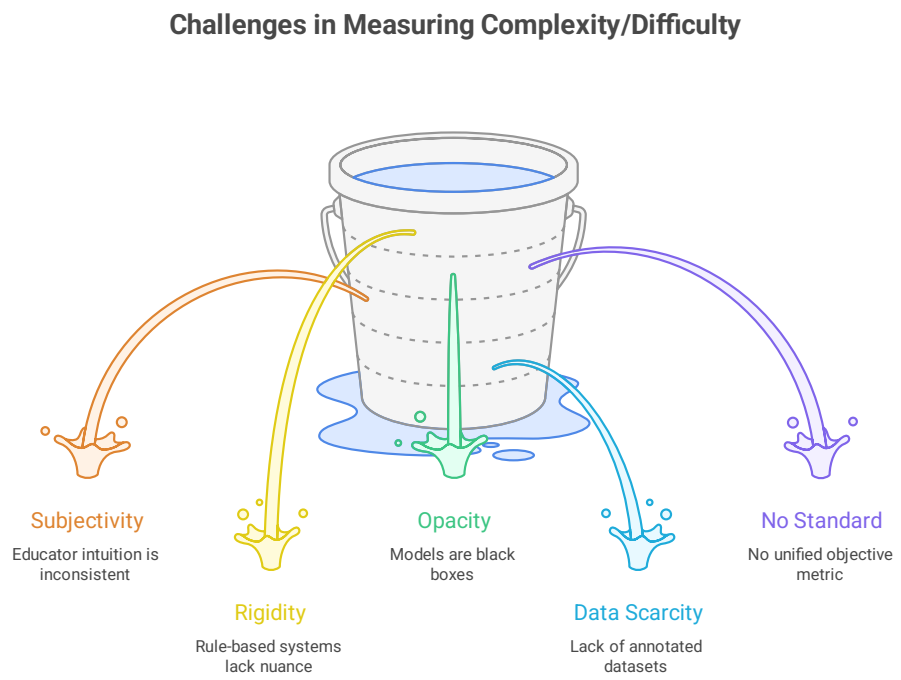
  Rule-based systems lack nuance and scalability.

- **Opacity**

  ML models can be black boxes and lack interpretability

- **No Standard**

  No unified, objective metric for complexity



INTRODUCING THE CEREBRO INDEX (CI)

A Multi-Dimensional, Quantifiable Question Complexity Scoring Framework

THE CEREBRO INDEX (CI) FRAMEWORK

- **Quantifying Question Complexity**

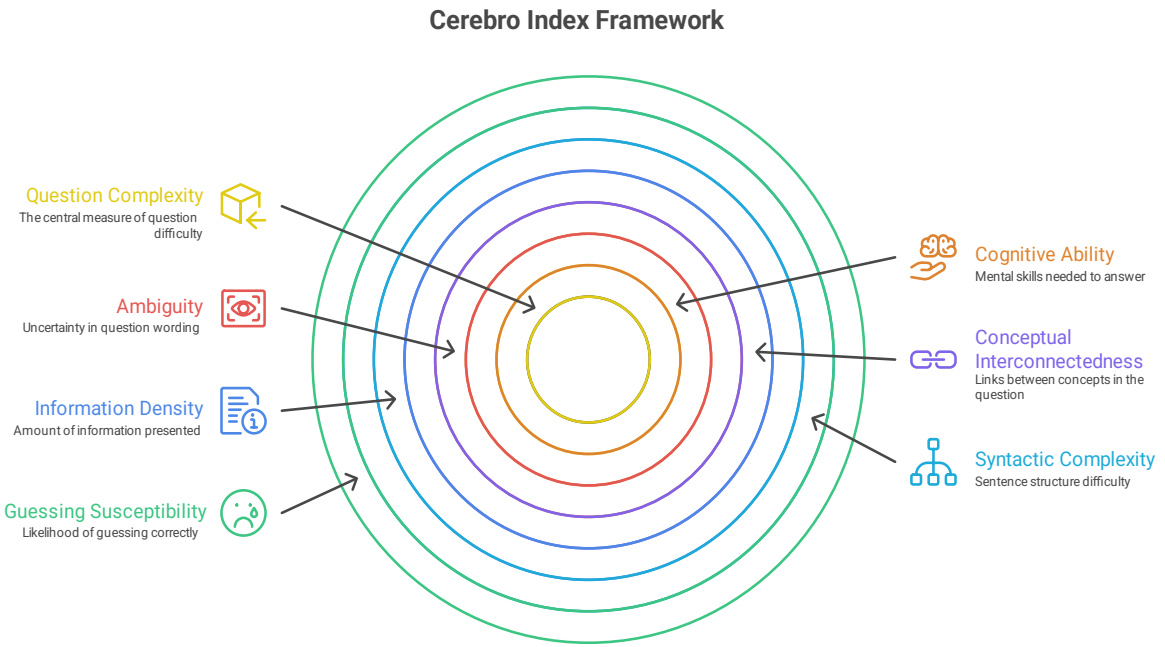
  The Cerebro Index (CI) assigns a score from 0.0 to 10.0, objectively measuring the inherent complexity of questions.

- **Six Core Attributes**

  CI integrates cognitive level, ambiguity, conceptual links, information density, syntactic structure, and susceptibility to guessing, with Bloom's Taxonomy driving 80% of the weight.

- **Complexity vs. Difficulty**

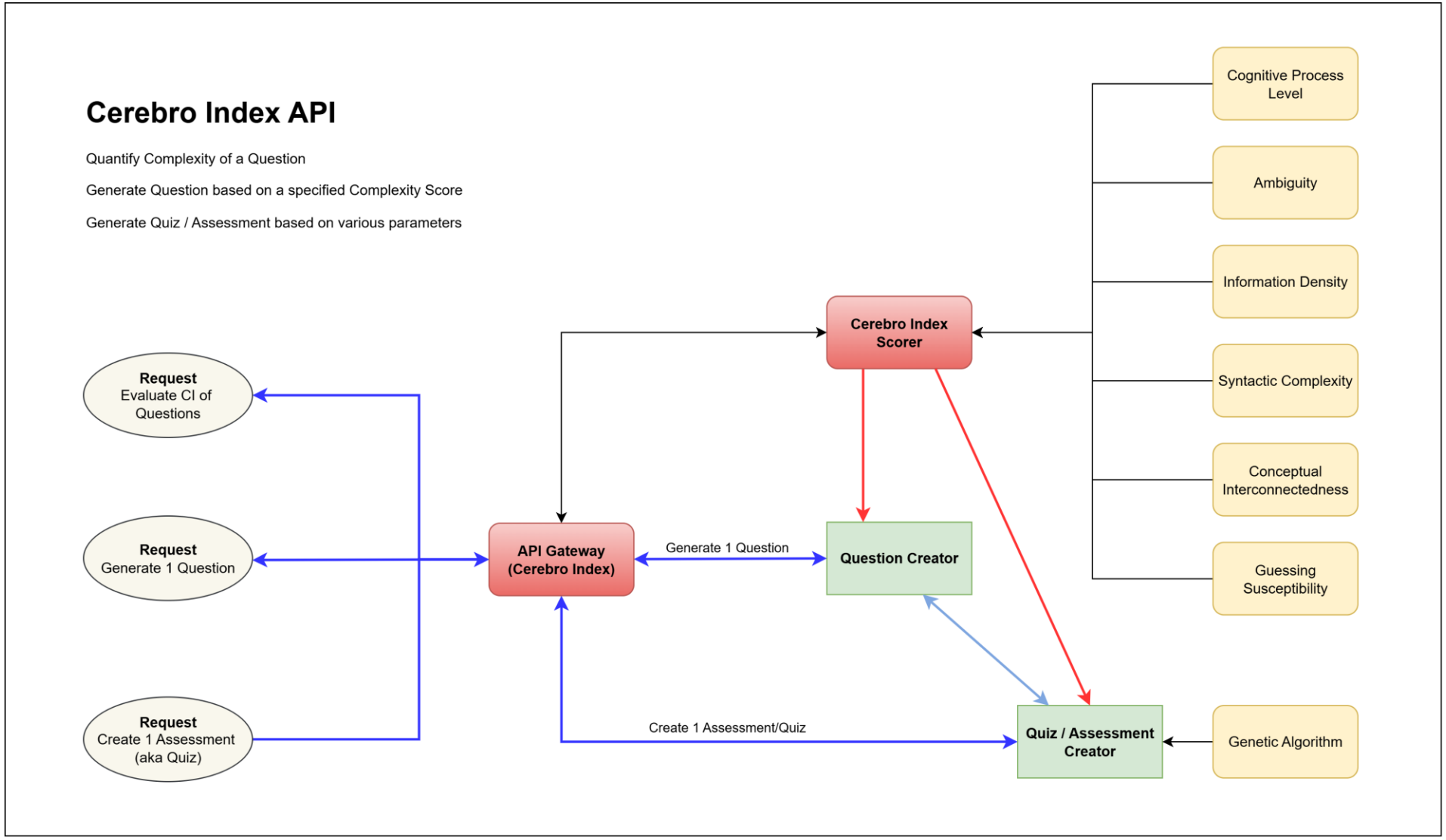
  Unlike learner-dependent difficulty, CI captures question complexity itself, making it essential for adaptive AI and education systems.



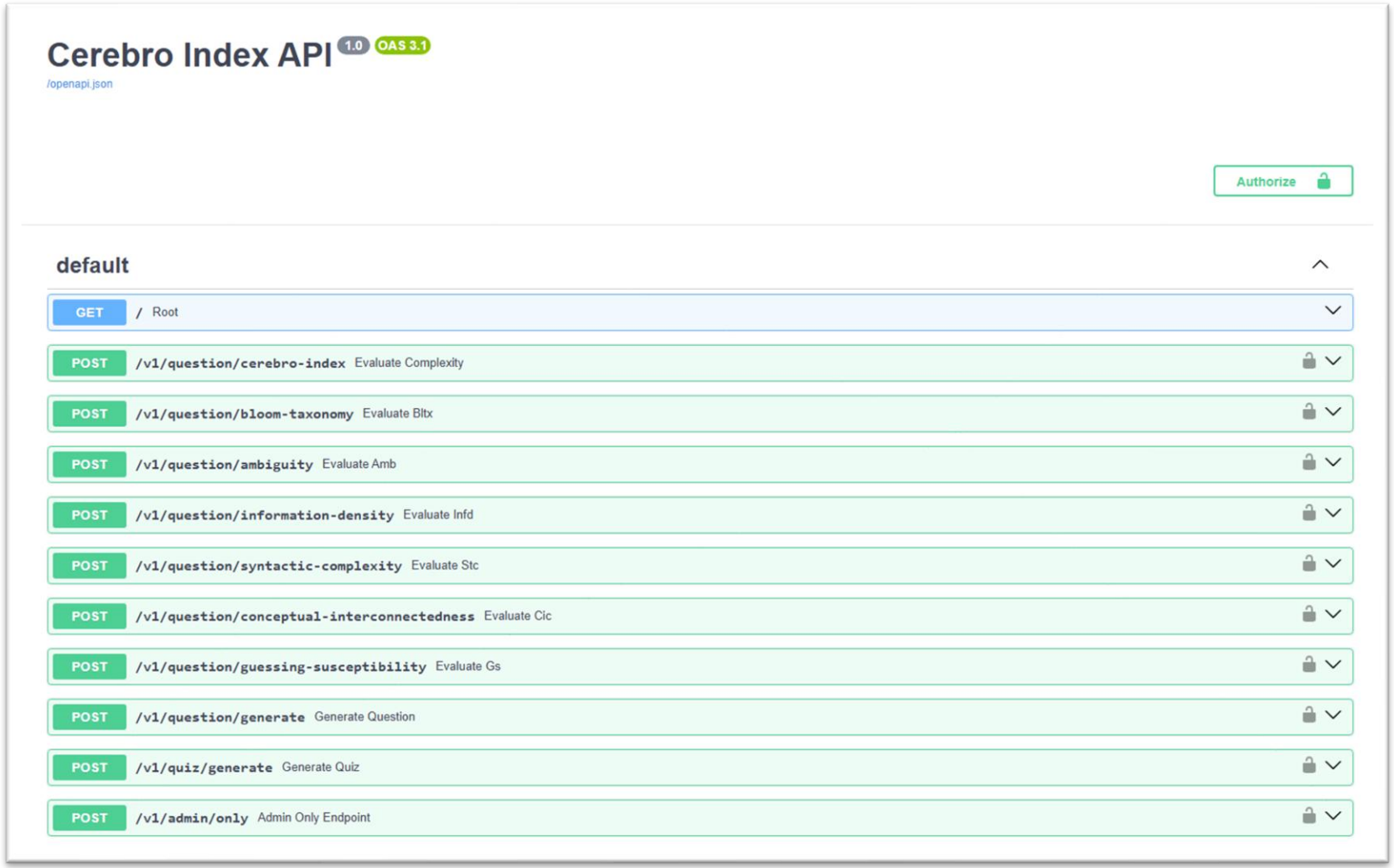
SYSTEM ARCHITECTURE & IMPLEMENTATION

From Theory to Practice: Building the CI System

SYSTEM ARCHITECTURE - COMPONENTS

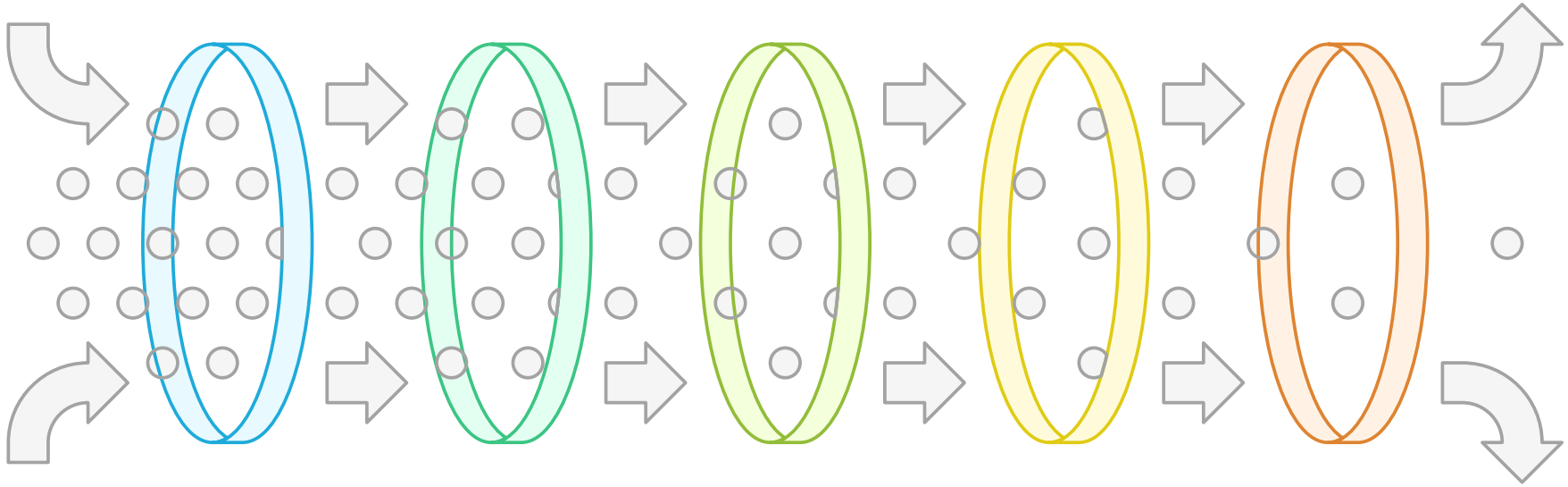


SYSTEM ARCHITECTURE – API ENDPOINTS



SYSTEM IMPLEMENTATION

Cerebro Index Development Funnel



Rapid Prototyping

Developing a minimal viable version

User Feedback Collection

Gathering insights for improvement

Iterative Refinement

Refining based on user feedback

Transition to Open Source

Moving to fine-tuned open-source models

Model Quantisation

Optimising for efficiency and scalability

THE VALIDATION STUDY - METHODOLOGY

Rigorous Empirical Validation: Does CI Align with Human Judgment?

VALIDATING THE CEREBRO INDEX

- **Objective**

  To validate the Cerebro Index (CI) as a proxy for human-perceived question complexity across multiple domains.

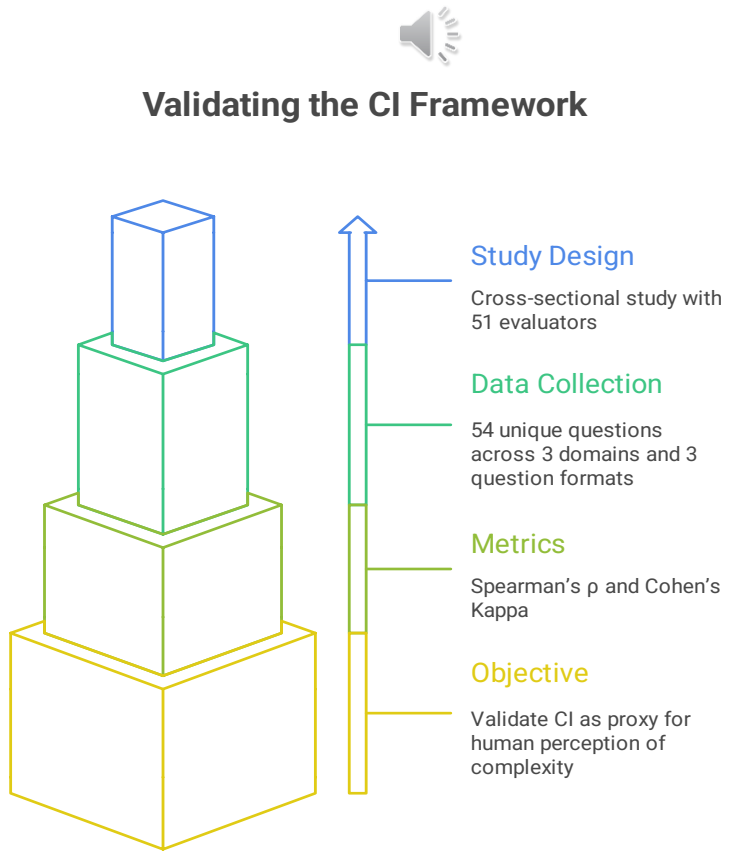
- **Design**

  Cross-sectional perceptual study using ordinal rankings and complexity labels by human raters, compared to CI outputs.

- **Research Questions**

  RQ1: Do CI scores positively correlate with human-perceived question complexity rankings?

  RQ2: Do CI complexity categories align with those assigned by human evaluators?



FINDING #1: CI SCORES STRONGLY CORRELATE WITH HUMAN RANKINGS

- **Overall**

  Significant, moderate-to-strong positive correlation ρ = 0.63, 95% CI [0.436, 0.769], p<0.05.

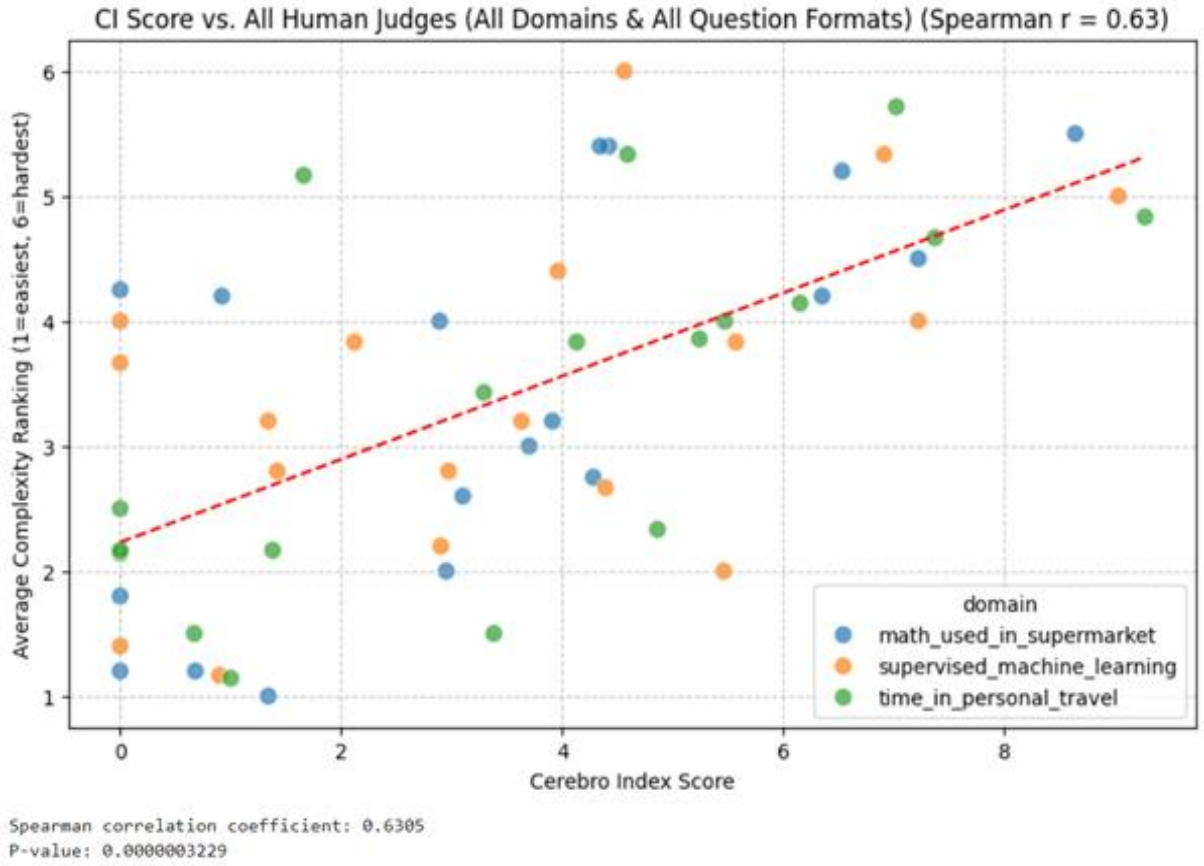
- **Domain Question Format Highlights.**

  True/False: Exceptional correlation in ML (ρ=0.986, p=0.0003) and Math (ρ=0.87, p=0.0244).

  MCQ: Very strong correlation in Time (ρ=0.943, p=0.0048)

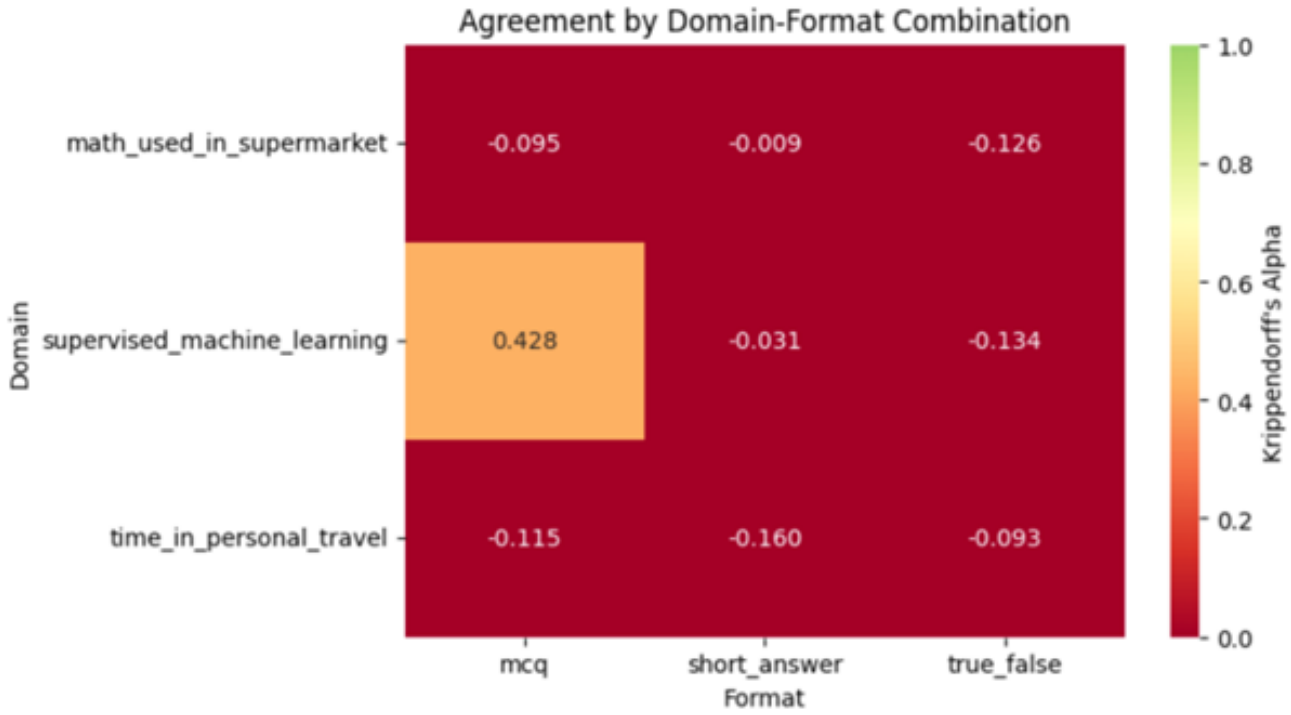
- **Conclusion**

  CI is a statistically valid measure of perceived complexity, especially for structured formats (MCQ, T/F).



FINDING #2: HUMAN EVALUATORS ALMOST ALWAYS DO NOT AGREE WITH EACH OTHER

```
Summary of Krippendorff's Alpha Values:
================================================
Overall                                 : -0.009 (Poor agreement)
Domain: math_used_in_supermarket        : -0.043 (Poor agreement)
Domain: supervised_machine_learning     : 0.106 (Poor agreement)
Domain: time_in_personal_travel         : -0.042 (Poor agreement)
Format: mcq                             : 0.016 (Poor agreement)
Format: short_answer                    : -0.044 (Poor agreement)
Format: true_false                      : -0.048 (Poor agreement)
math_used_in_supermarket - mcq          : -0.095 (Poor agreement)
math_used_in_supermarket - short_answer : -0.009 (Poor agreement)
math_used_in_supermarket - true_false   : -0.126 (Poor agreement)
supervised_machine_learning - mcq       : 0.428 (Moderate agreement)
supervised_machine_learning - short_answer: -0.031 (Poor agreement)
supervised_machine_learning - true_false: -0.134 (Poor agreement)
time_in_personal_travel - mcq           : -0.115 (Poor agreement)
time_in_personal_travel - short_answer  : -0.160 (Poor agreement)
time_in_personal_travel - true_false    : -0.093 (Poor agreement)
```



FINDING #2: BUT MAJORITY OF HUMAN EVALUATORS SHOW SUBSTANTIAL AGREEMENT WITH CI ADJACENT CATEGORIES

- Overall **κ** = 0.619:

  Substantial agreement beyond chance

- Off-by-one agreement = 0.778:

  CI mapped complexity category and human assigned (majority voting) are usually within one category of each other.

```
============================================================
AGREEMENT BETWEEN AI PREDICTIONS AND HUMAN MAJORITY VOTE
============================================================

Overall Agreement:
  Accuracy: 0.204
  Cohen's Kappa: 0.619
  Exact Agreement: 0.204
  Off-by-One Agreement: 0.778
  Number of Questions: 54

By Domain:
  math_used_in_supermarket:
    Accuracy: 0.222
    Cohen's Kappa: 0.615
    Exact Agreement: 0.222
    Off-by-One Agreement: 0.722
    Number of Questions: 18
  supervised_machine_learning:
    Accuracy: 0.278
    Cohen's Kappa: 0.640
    Exact Agreement: 0.278
    Off-by-One Agreement: 0.889
    Number of Questions: 18
  time_in_personal_travel:
    Accuracy: 0.111
    Cohen's Kappa: 0.612
    Exact Agreement: 0.111
    Off-by-One Agreement: 0.722
    Number of Questions: 18
```

FINDING #3:THE PRIMARY CRITERIA APPLIED BY HUMAN EVALUATORS APPEAR TO BE BROADLY CONGRUENT WITH THE FACTORS INCORPORATED INTO THE CI

- Key Assessment Factors considered by Human Evaluators

  Cognitive effort: Time to understand/solve questions and number of cognitive steps required.

  Knowledge requirements: Level of specialised expertise needed

  Language characteristics: Sentence length, vocabulary complexity, and clarity

- This is in-line with the components that Cerebro Index considers when determining the CI Score for a question.

DISCUSSION, LIMITATIONS & FUTURE WORK

Critical Reflection: Successes, Constraints, and the Path Forward

LIMITATIONS

- **Small Sample Size**

  Only 54 unique questions and 51 evaluators, resulting in 588 evaluated questions; limited statistical power

- **Methodological gaps**

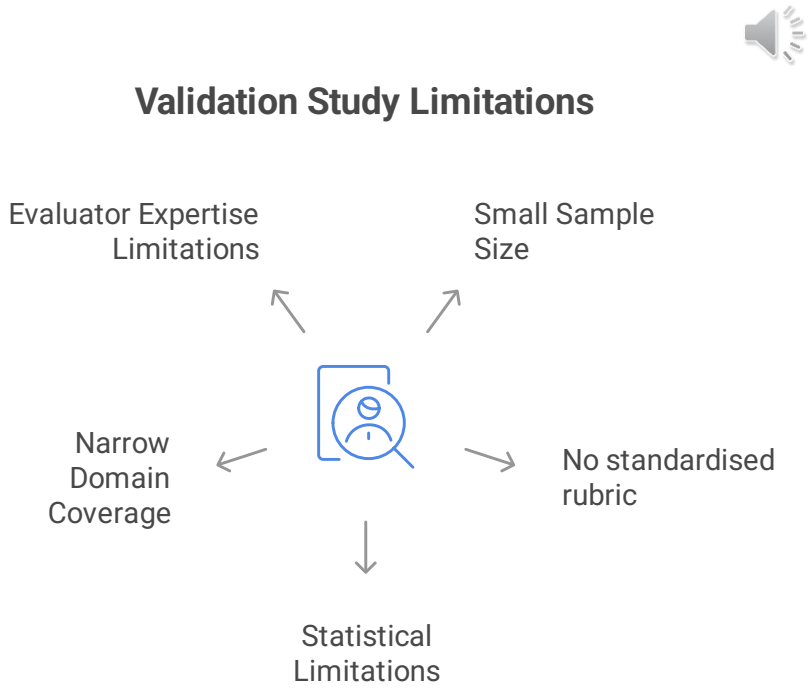
  No standardized rubric and varied evaluator expertise

- **Narrow Coverage**

  Only 3 domains, 3 question formats per domain and 6 questions per question format

- **Statistical Limits**

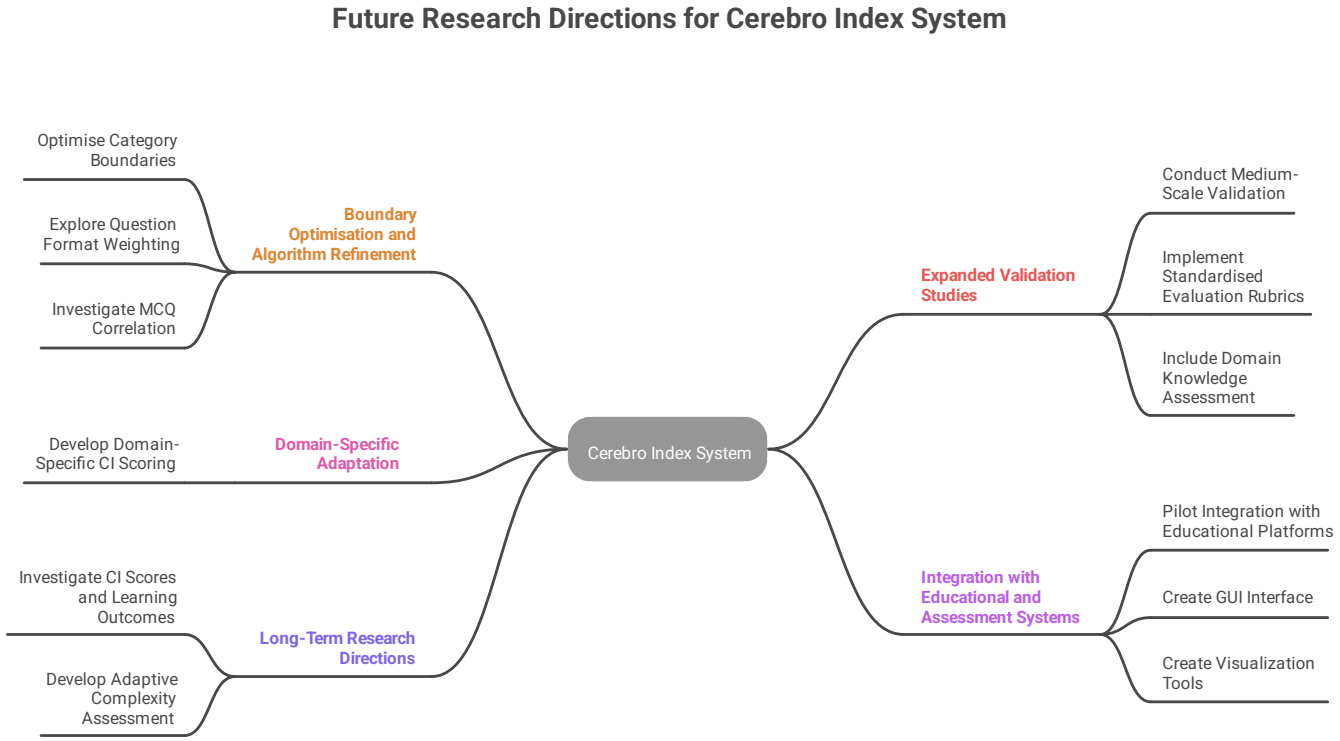
  Several weak/non-significant correlations



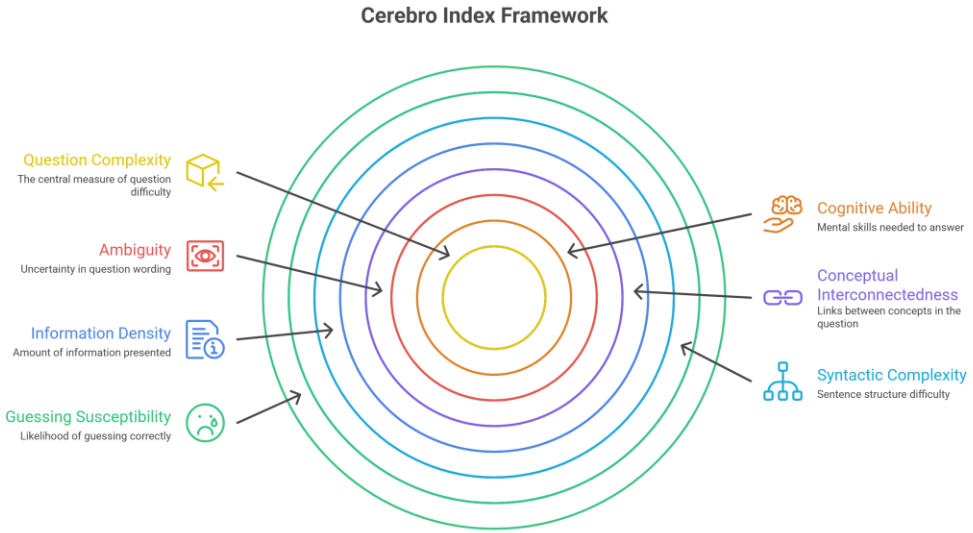
Despite these issues, the study confirmed the framework's potential and identified clear paths for refinement.

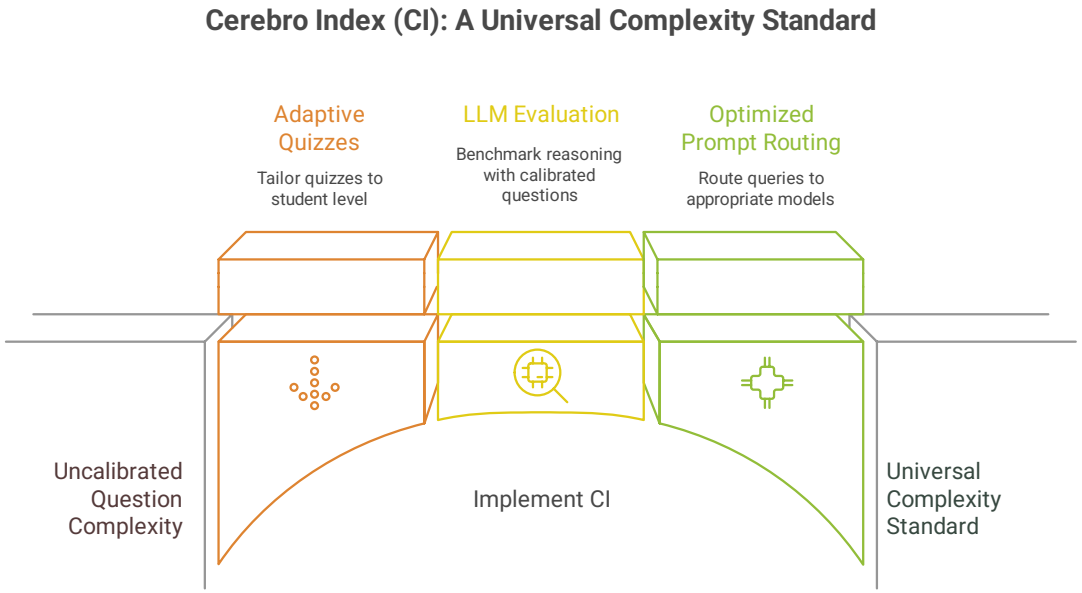
FUTURE WORKS

- Near Term:
  Optimize Complexity Category Boundary; develop format-specific weightings.
- Near Term :
  Conduct a larger-scale validation (200-500 unique questions) with a standardized rubric.
- Long-term:
  Integrate CI into existing learning platforms; explore its use for AI benchmarking.



BEYOND THE CAPSTONE: THE TRANSFORMATIVE POTENTIAL OF CEREBRO INDEX(CI) FRAMEWORK





The **Cerebro Index (CI)** Framework is a validated framework that quantifies question complexity using a continuous, multi-dimensional scoring system. It bridges subjective perception and objective measurement, offering a scalable and psychometrically grounded tool for advancing AI assessment and intelligent education.

CEREBRO INDEX (CI) FRAMEWORK

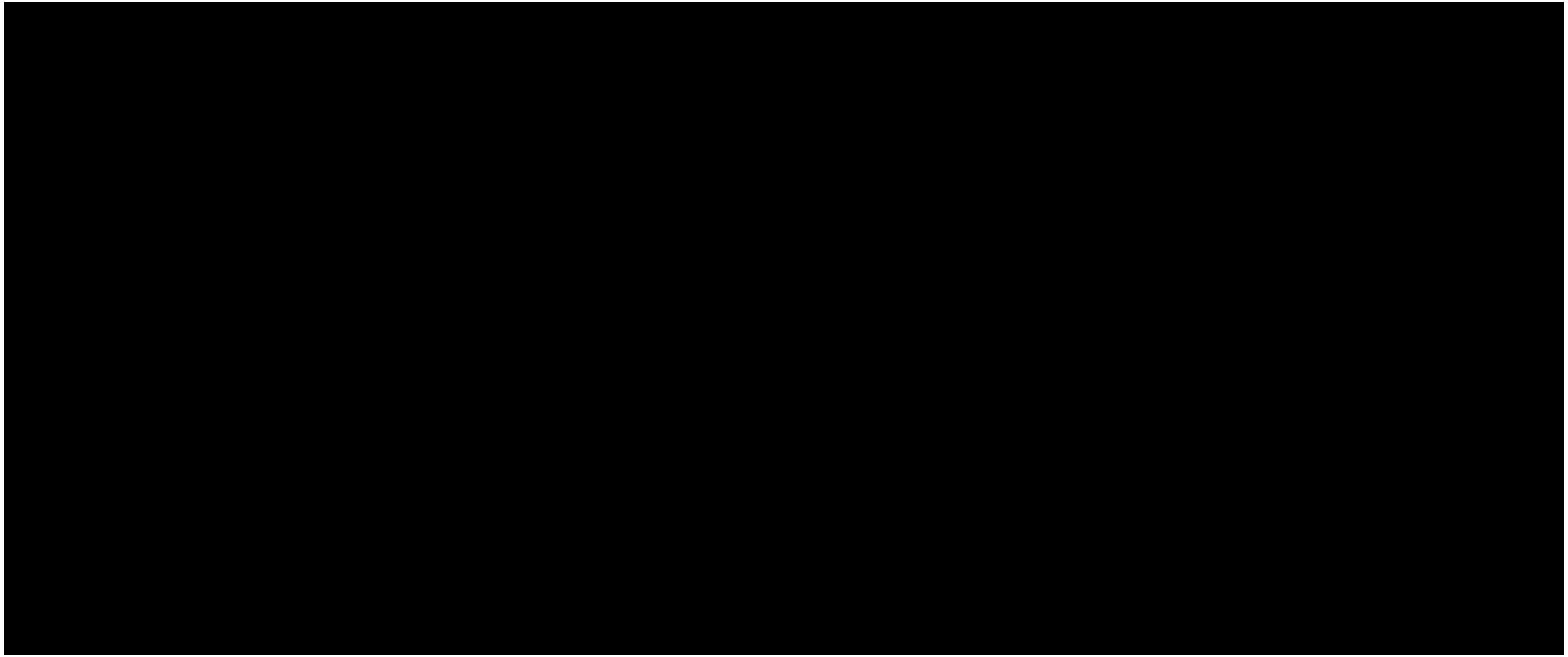
Validated Framework for Objective Question Complexity Measurement

DEMO

Richard Chai | e0700541@u.nus.edu | richard.chai@neutriumlabs.com

CEREBRO INDEX (CI) FRAMEWORK

Validated Framework for Objective Question Complexity Measurement



CEREBRO INDEX (CI) FRAMEWORK

A psychometrically grounded measure of question complexity for education and AI.

Thank you!