Leveraging LLMs for Automatic Item Generation and Grading in English Language Assessment

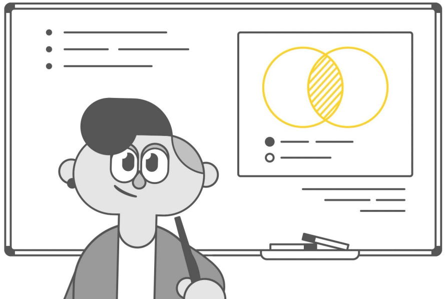
Andrew Runge
Duolingo English Test
September 13, 2023

Overview

- The Duolingo English Test (DET) as a digital-first assessment
- Developing automatic item generation (AIG) processes for complex item types using LLMs
- Assessing writing with GPT-4

Duolingo English Test

- Computer adaptive test (CAT) of English language proficiency
- Delivered online, remotely anywhere/anytime
- Proctored after test completion (2-3 people)
- Score turn-around < 48 hours
- Report an overall score and four subscores
- Much shorter (~1 hour) and cheaper than alternatives
- A **digital-first** assessment



A digital-first assessment

Born digital

- Designed with computational models, AI-algorithms, and tech capabilities in mind
- Theory-based & evidence-centered
- Valid & reliable

Integrated

- Fluid infrastructure
- Integrative theoretical frameworks
- Integrated (automatic) tools & databases

Superior test-taker experience (TTX)

- Available on demand, anytime, anywhere
- Online testing with remote proctoring

Digital-first involves a lot of automation

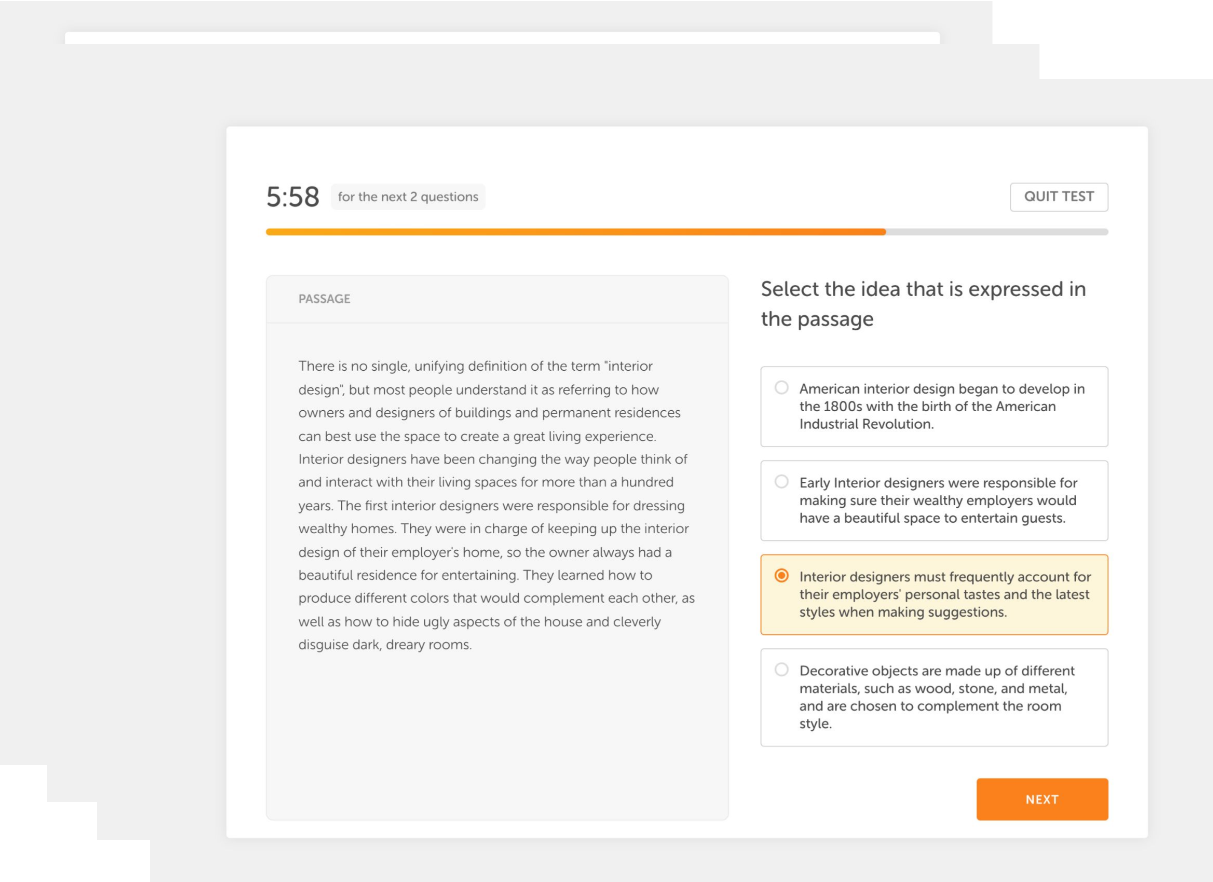
- **Automatic item generation**
- **Automatic grading**
- Computational psychometrics
- Automatic quality assurance for assessment (AQuAA) systems
- Automation in item piloting and review

Automatic Item Generation

Overview of Approach

- Use large language models for automated generation of items, creating texts and associated materials, such as:
  - Reading passages (**Interactive Reading**)
  - Conversations (**Interactive Listening**)
  - Questions, keys and distractors
  - Other information necessary for automated scoring
- Iteratively improve generated materials based on:
  - Review and edits of materials by experts
  - Test taker responses on a pre-pilot platform

The Interactive Reading Task



- A series of tasks where test takers interact with a text for several purposes targeting different reading skills
- Closely represents the task of university reading

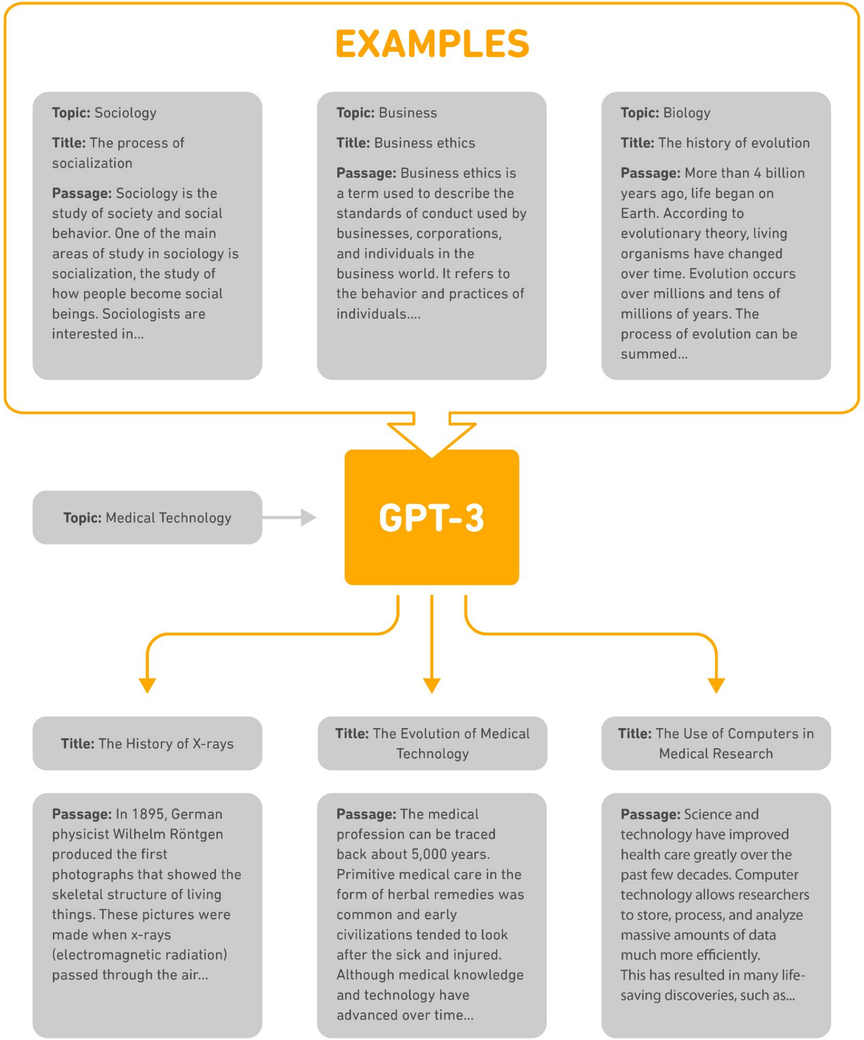
Passages & Titles

Goal

Generate self-contained passages in a range of domains - especially at an academic level

Method

Generate passages and titles conditioned by other academic content and a specified topic using prototypical examples



MC option generation

Goal:

Generate correct and incorrect options

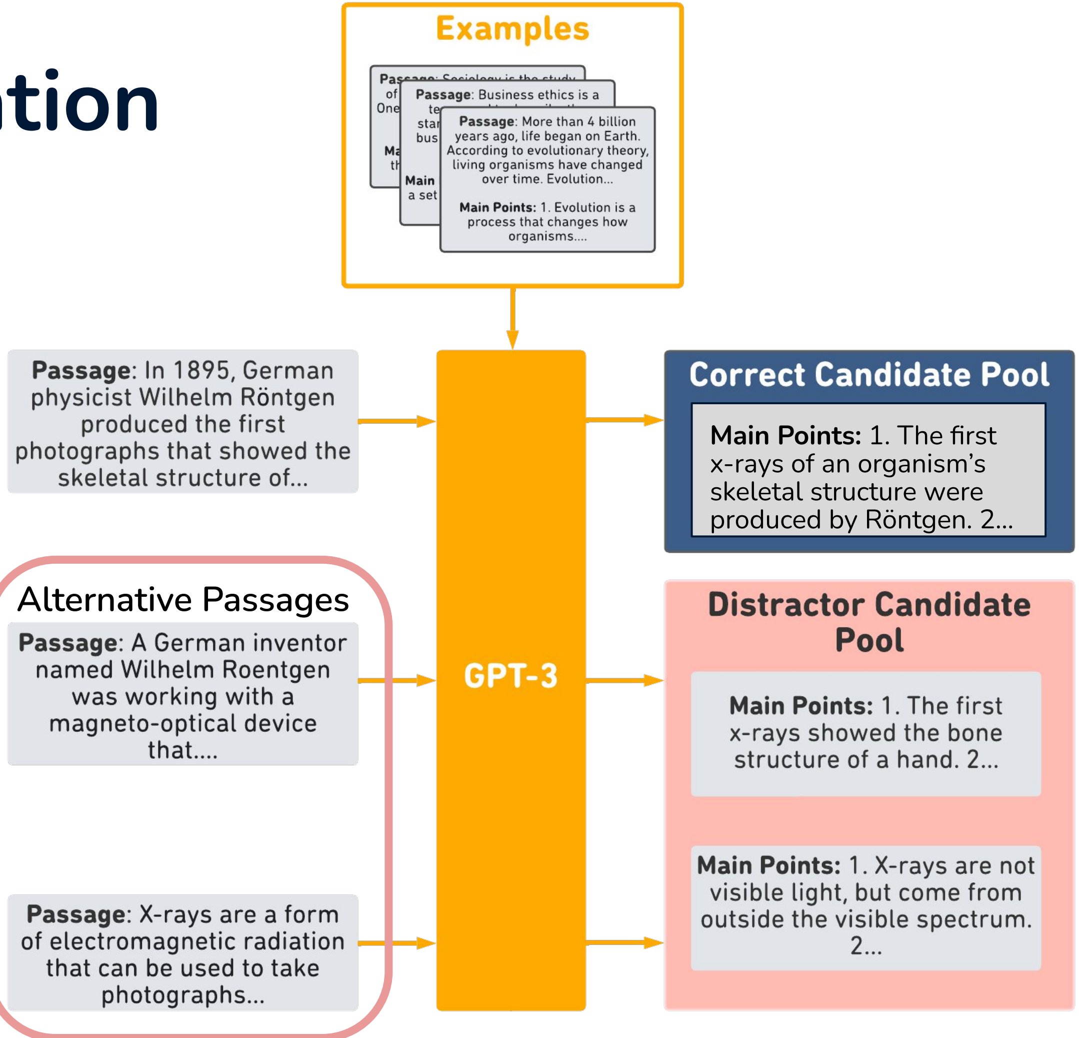
Challenge:

GPT struggles to generate intentionally wrong answers

Solution:

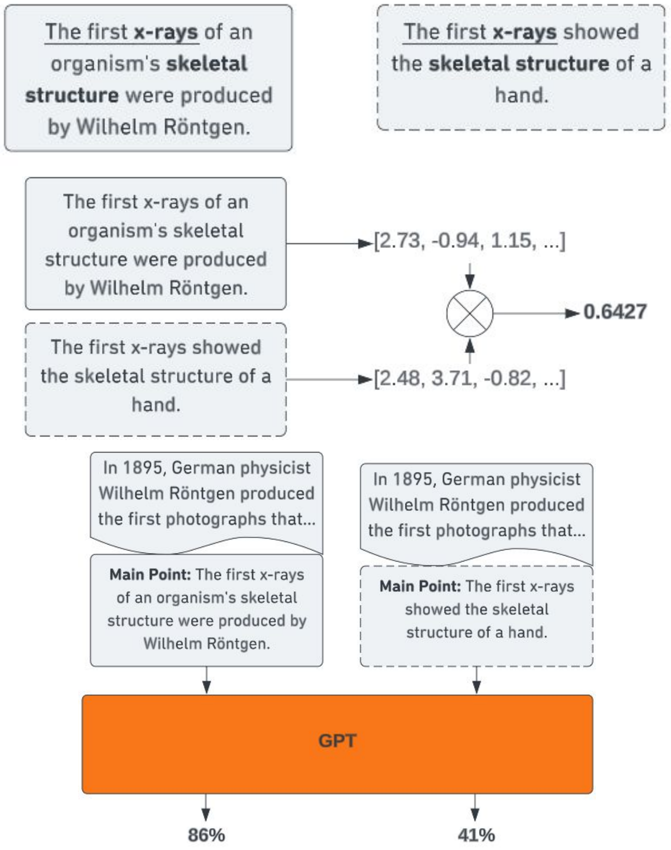
Generate alternative passages on the same topic/title

Use the alternative passages as inputs to generate answers that can



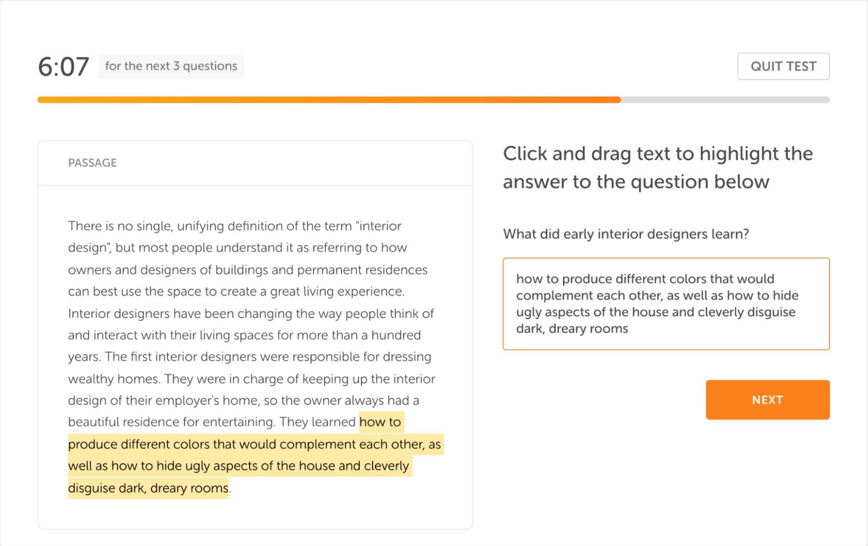
Selecting Keys and Distractors

- **Surface features**
  - Ex: content word overlap, prefix overlap, relative length of key vs. distractors
- **Semantic vector features**
  - Cosine similarity between text and candidate keys and distractors
- **LLM features**
  - Log likelihood from LLM of distractor as its answer for original passage

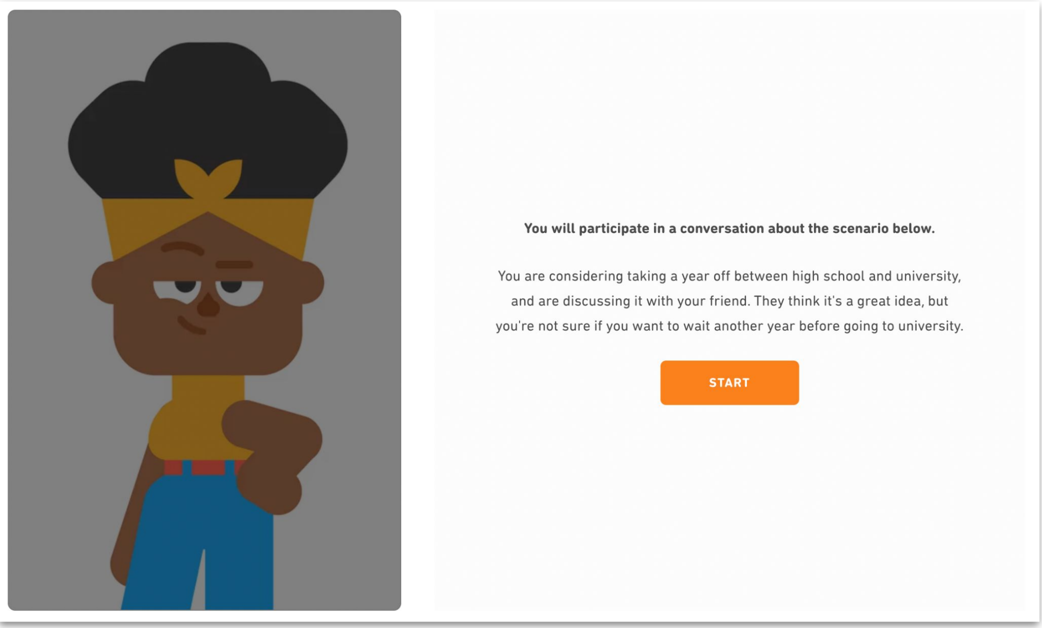


Highlight-the-Answer Questions

- Open-ended item type where test taker highlights the text that answers the question
- Generate question candidates conditioned on the passage and a target answer in the text
- Evaluate using mix of surface and LLM features and evaluation by a secondary QA model to confirm answers



Interactive listening



- Two common speech **situations in university contexts**
  - Student-student
  - Student-professor
- Two different power relationships (adds variation to the language of the turns)
- Representative **communicative purposes** (e.g., requests, advice, planning, and learning)
- Diverse **topics** and specific **details**

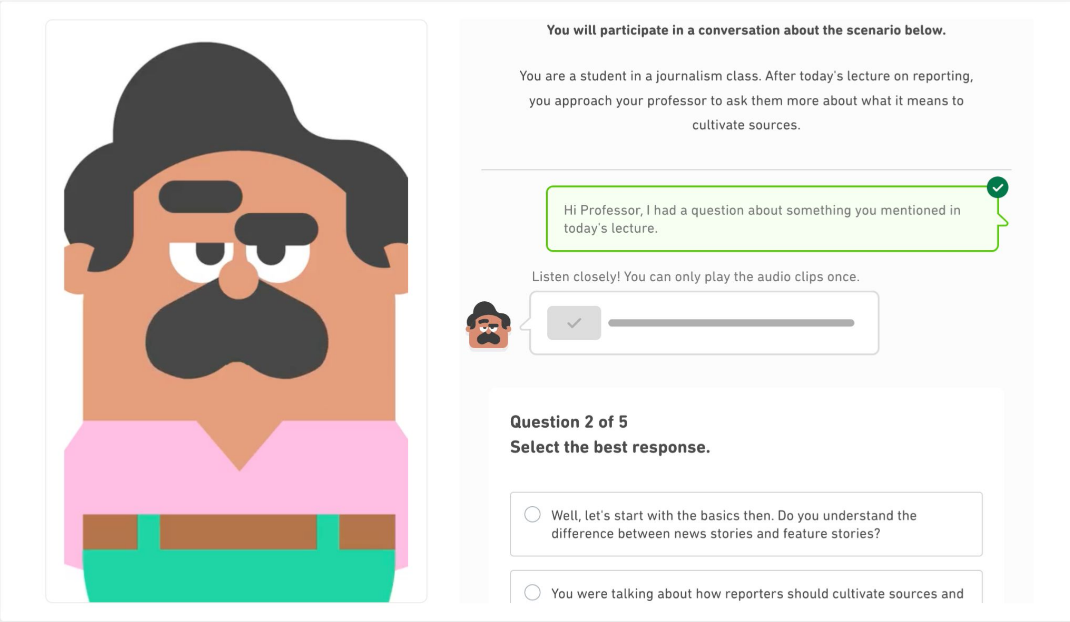
Conversations

Goal

Generate short conversations oriented around academic scenarios.

Key Challenge

How to get diverse conversations from limited human-created input?



Scenario Expansion with GPT

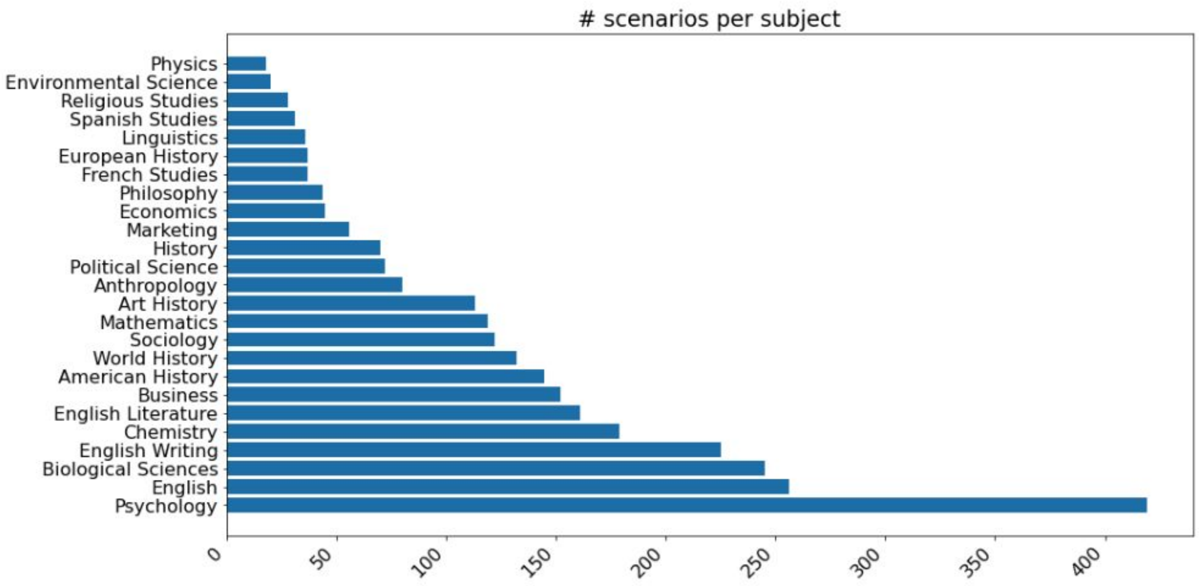
- **Base scenario:** A student is interested in a job post and wants to see if their advisor thinks it would be a good fit.

- **Expanded scenario:** You are a student who is about to graduate with a degree in English. You see a job posting for an editor at a publishing company. You're not sure if it would be a good fit for you, so you decide to ask your advisor who worked as an editor in the past for their advice.

Scenario Expansion

- Unguided scenario expansion leads to homogeneity
  - GPT has its own preferences for certain details

- Solution: Label expanded scenarios for key details, have GPT rewrite scenarios with shuffled details

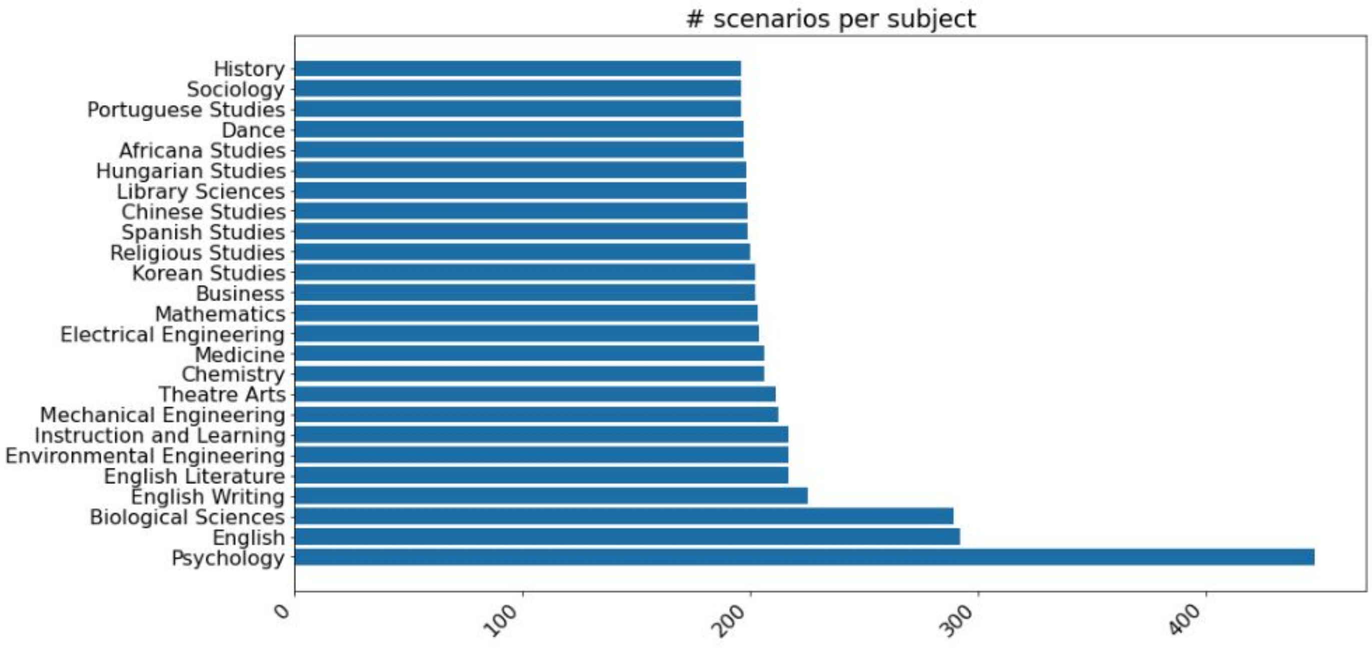


Scenario Expansion with GPT

- **Expanded scenario:** You are a student who is about to graduate with a degree in English. You see a job posting for an editor at a publishing company. You're not sure if it would be a good fit for you, so you decide to ask your advisor who worked as an editor in the past for their advice.

- **Rewrite with subject Engineering:** You are a student who is about to graduate with a degree in Mechanical Engineering. You see a job posting for an engineer at a large manufacturing company. You're not sure if it would be a good fit for you, so you decide to ask your advisor who worked as an engineer in the past for their advice.

Rewritten Scenario Expansions



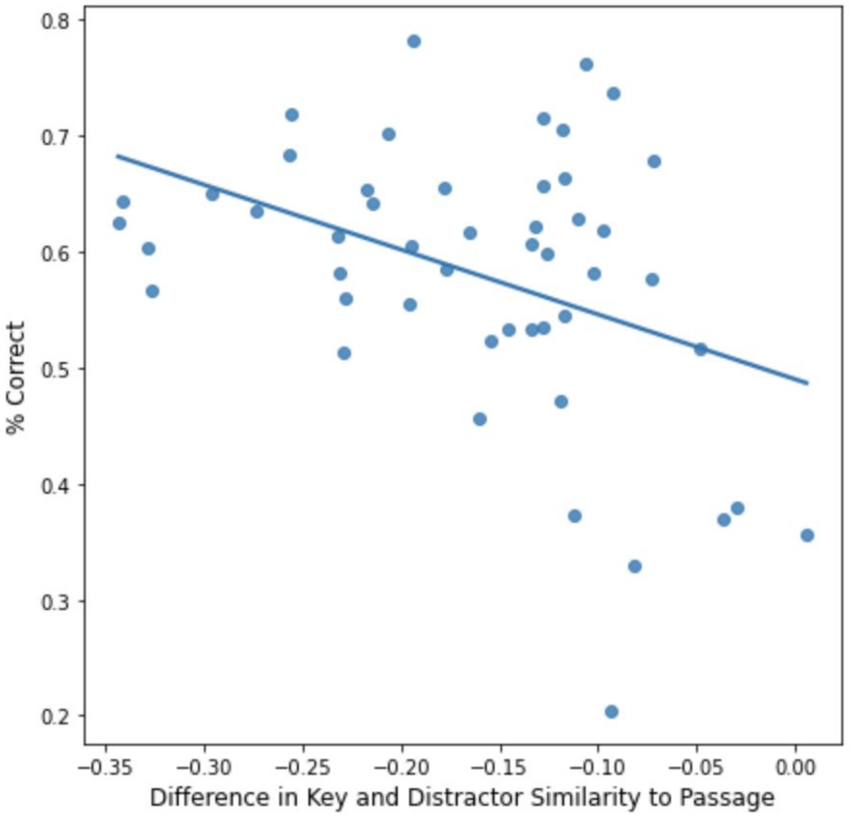
Dialogue MC Options

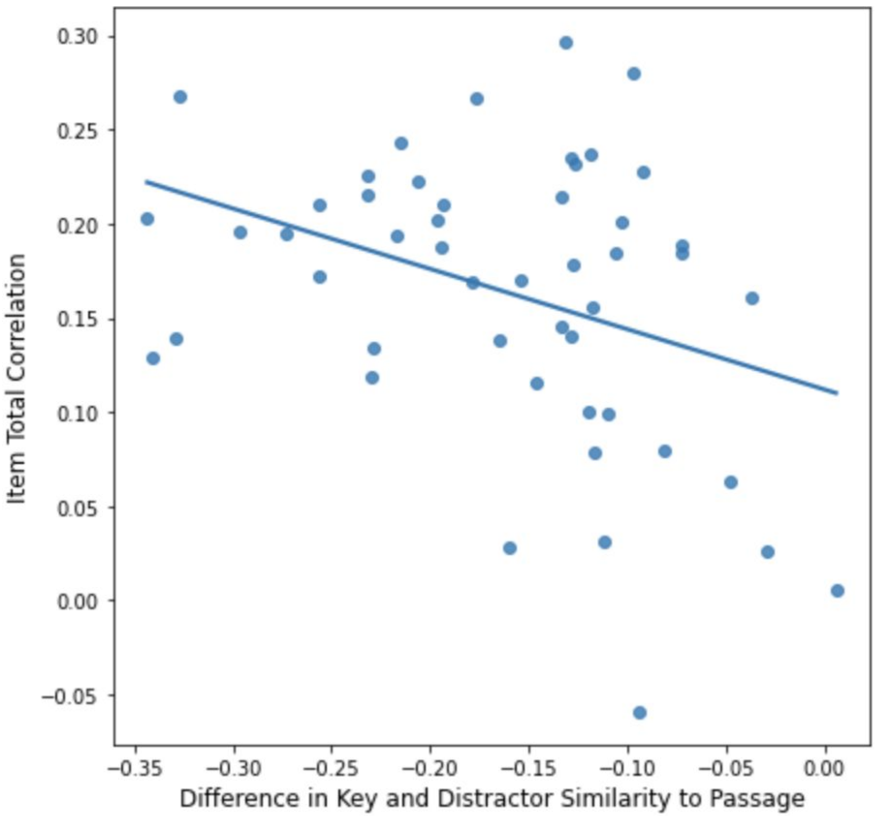
- Dialogues share many common lines, regardless of the specific topic
  - Greetings, expressions of gratitude, clarifying or planning questions

- Leverage all generated conversations as potential sources of distractors
  - Can use features of the line of dialogue (ex: position in the conversation) and its source conversation (ex: participants, scenario, topic) to identify, filter, and rank candidate distractors (+ previous features)

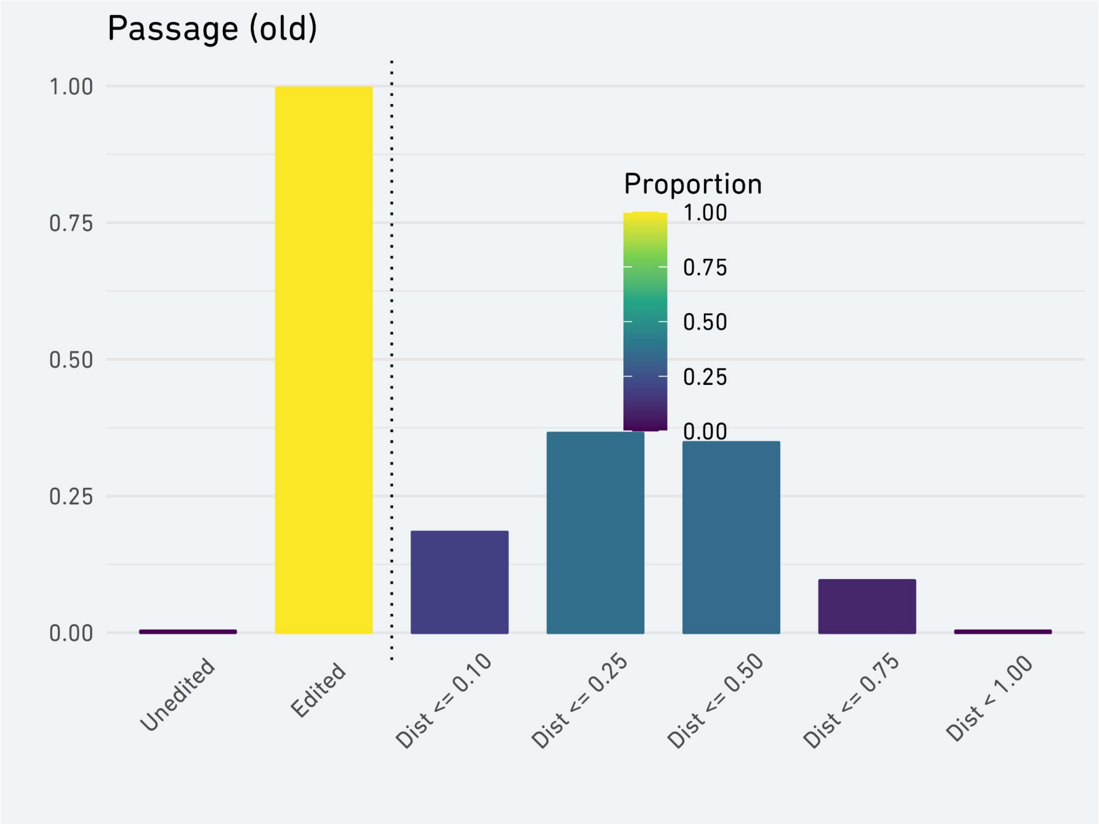
Iterative Improvement

Test-Taker Response Driven Changes

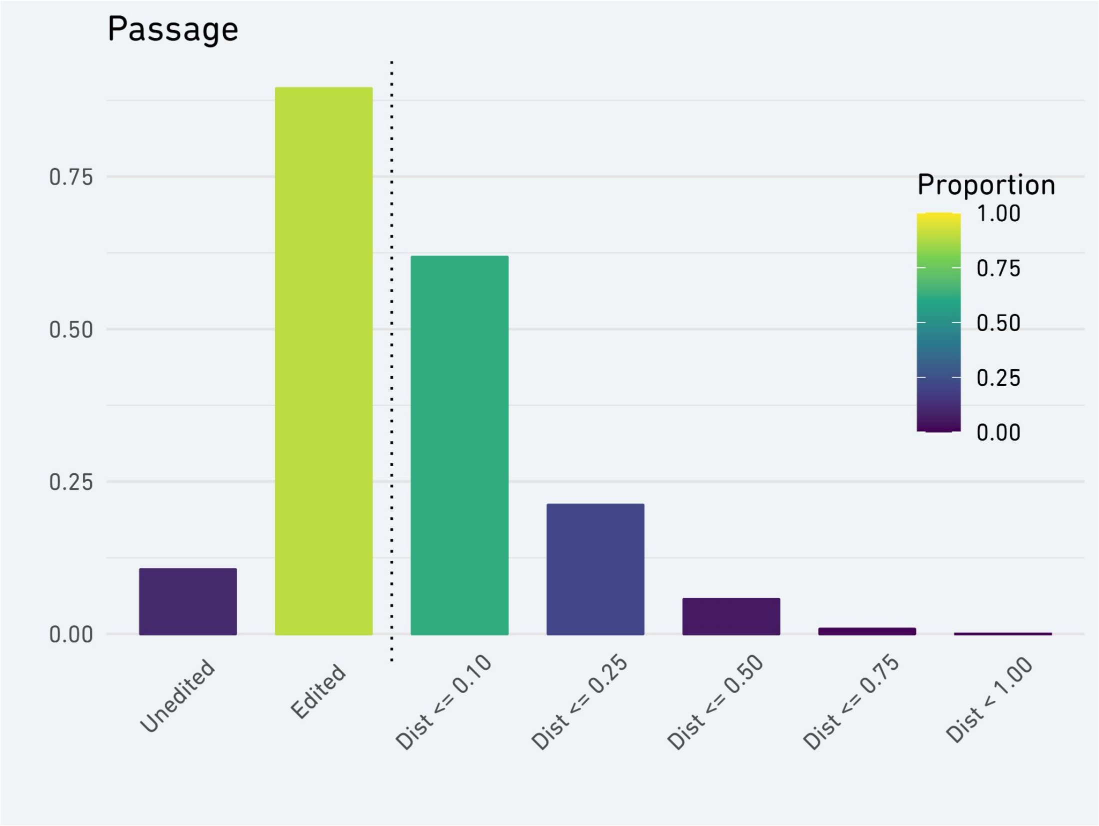




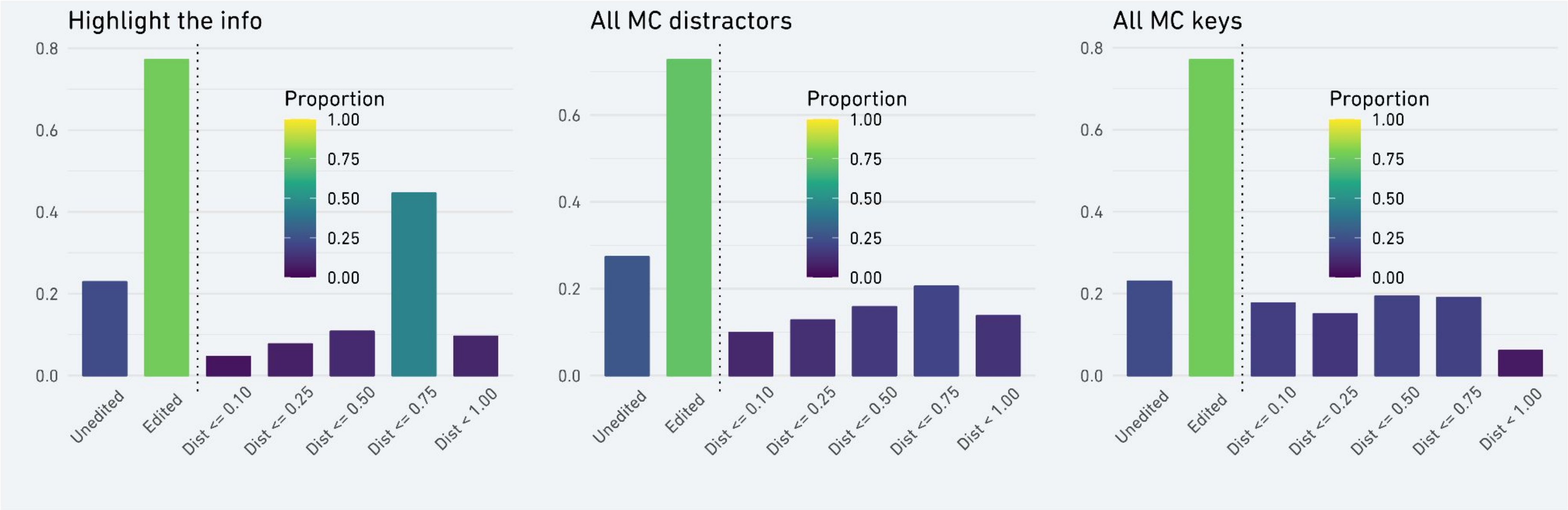
Measuring Improvements via Reviewer Edits



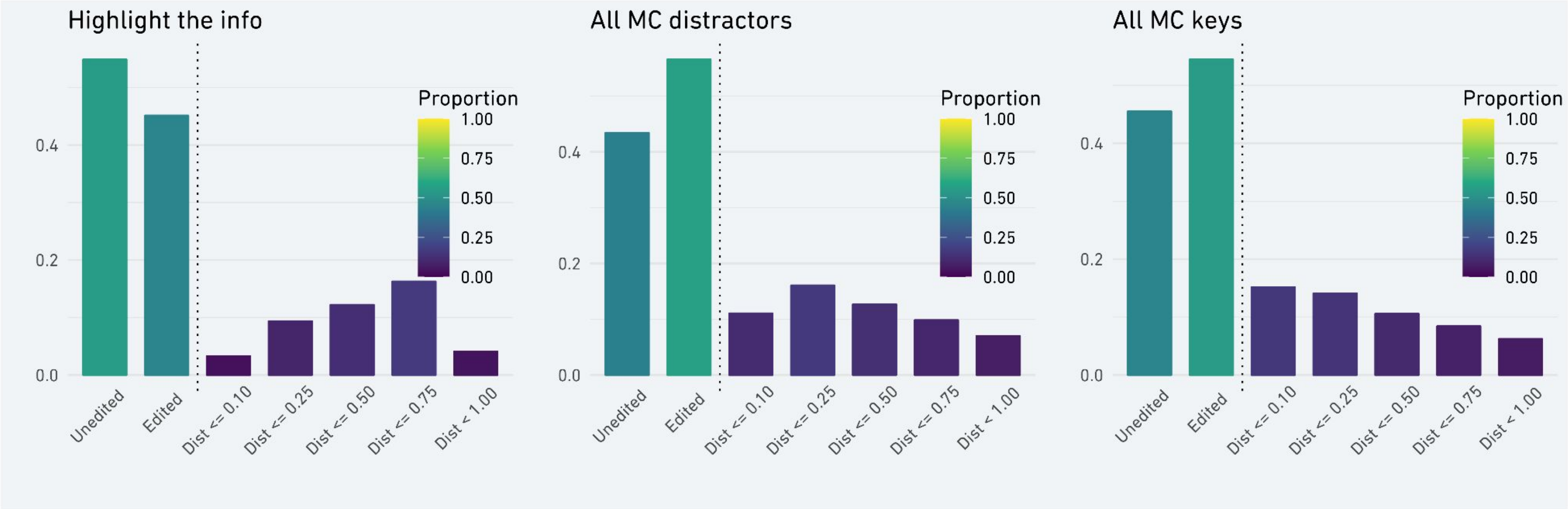
Char edit distances from last IR batch



Char edit distances from an early IR batch

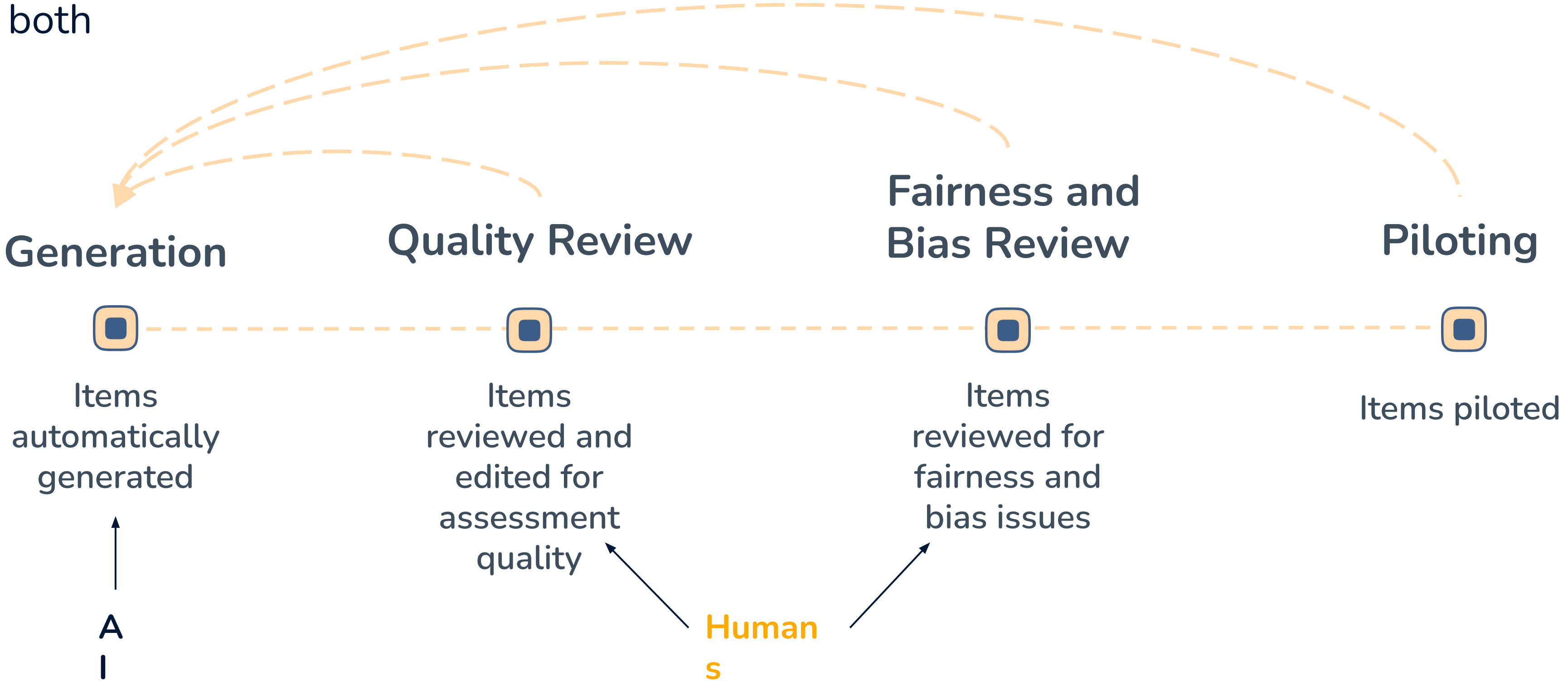


Char edit distances from last IR batch



Human-In-The-Loop AI

Goal: Use collaboration of humans and AI to perform a task that is difficult for both



Assessing Writing with LLMs

Rating English Learner Essays

- **Yancey et al. 2023 - Rating Short L2 Essays on the CEFR Scale with GPT-4**
- Naismith et al. 2023 - Automated Evaluation of Written Discourse Coherence with GPT-4

Evaluating Overall Writing Ability

- Many large-scale, high-stakes tests deploy Automated Writing Evaluation (AWE) systems to rate essays.
  - Typically rule-based and statistical feature-based models, more recently incorporating BERT features (Lagakis and Demetriadis, 2021)

- Much like with item generation, LLMs have the potential to revolutionize how we evaluating writing

- **How do LLMs perform at grading English learner (L2) compare to existing writing evaluation methods?**
  - Are they as accurate?
  - Are they fair?

Data and Models

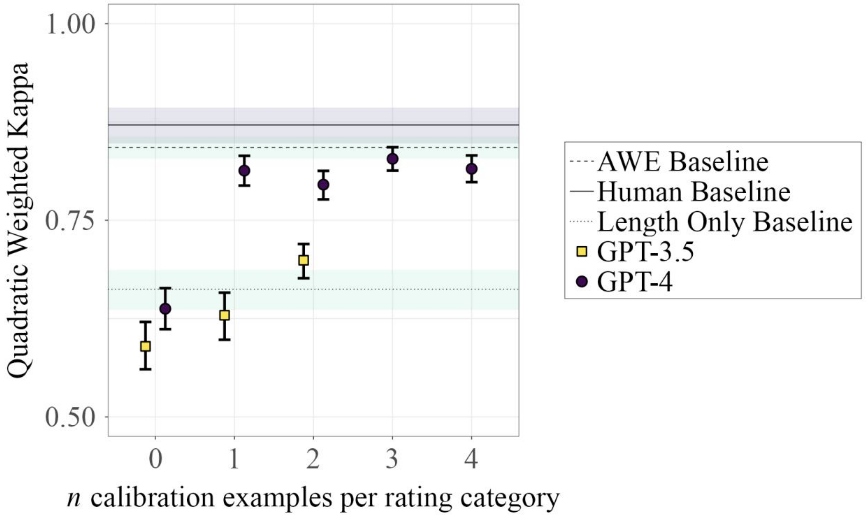
- **Data:**
  - 1,961 narrative or persuasive essays from 5-minute writing tasks
    - Balanced over 7 most common first languages (L1) and gender
  - 6 point rating scale
    - Based on the Common European Framework of Reference (CEFR)
    - 0 assigned to off-topic and minimal responses
  - 2 raters, 389 essays rated by both with 0.87 Quadratic Weighted Kappa

- **Models:**
  - Length-based baseline
  - Feature-based Automated Writing Evaluation (AWE) system
  - GPT-3.5 & GPT-4, where prompt included a short instruction set and 0-4 calibration examples per CEFR category

CEFR Rating Results

- With a single calibration example, agreement between human and GPT-4 nearly matches the AWE system
    - Further examples and more elaborate prompts gave no significant improvements

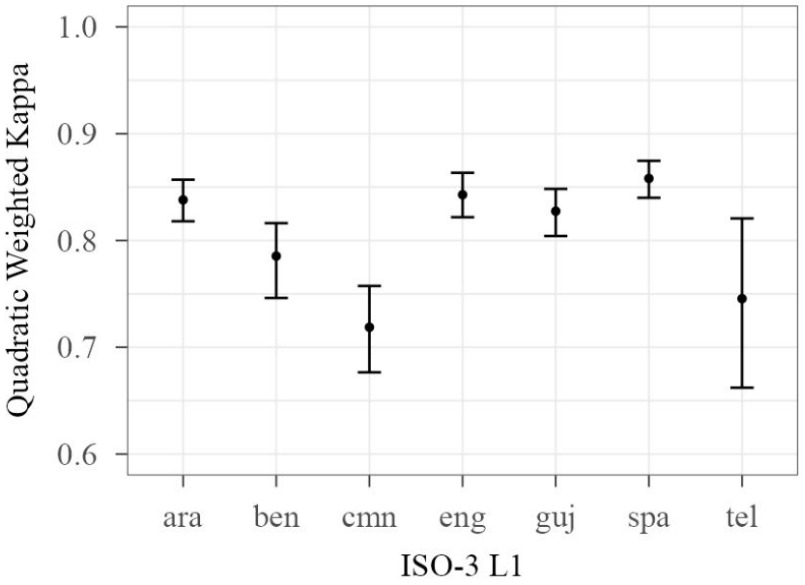
- GPT-3.5 barely improves over simple baseline



Fairness in Ratings

- No significant differences by gender

- But found significant differences in agreement across test taker's primary language (L1)
  - Could be due to lower reliability of human ratings on certain L1



Rating English Learner Essays

- Yancey et al. 2023 - Rating Short L2 Essays on the CEFR Scale with GPT-4

- **Naismith et al. 2023 - Automated Evaluation of Written Discourse Coherence with GPT-4**

Discourse Coherence

- Typically covers concepts including:
    - **Clarity** - easy of understanding of ideas, readability, vocabulary choice
    - **Flow** - progression of ideas, use of linking words, referencing
    - **Structure** - appropriate paragraphing, introducing and concluding ideas, connections between topics
    - **Effect on reader** - Naturalness of cohesion, appropriateness of cohesive features and how they help reader to understand the response

- Notoriously difficult to assess with AWE systems. **Can LLMs help with rating coherence?**
    - How do LLM ratings of coherence correlate with human ratings?
    - Do LLM analyses of discourse coherence align with human rationales?

Data & Evaluations

- **Dataset - DET-Coh**
  - 500 narrative and persuasive essays produced for a 5-minute writing task
    - Balanced over 7 most common first languages (L1) and gender
  - Rated by 2 raters on 6 point scale for coherence
    - Scale developed as a combination of CEFR descriptions, public test rubrics, and discourse research
  - 80 essays double-rated with 0.93 Quadratic Weighted Kappa

Models

- **Feature-based model** with NLP features over sentence pairs based on Coh-Metrix (Graesser et al. 2004)
  - Lexical, phrasal, and word stem overlap
  - Coreference overlap
  - LSA similarity
- Two versions of each feature - one computed over adjacent sentence pairs and one over all pairs
- Each example evaluated with hold-one-out training scheme

- **GPT-4:**
  - Prompt included:
    - Description of task
    - Evaluation rubric
    - Style guidelines
    - 1 example per CEFR level, consisting of prompt, response, and rating
  - Also tested giving rationale for each response, either before or after rating

Coherence Ratings

- All GPT-4 configurations outperform the baseline, but fall short of human performance
  - Tends to avoid the extremes of the scale more than humans
- May over-index on spelling errors - “spelling” mentioned in 43% of mismatches with human ratings

Metric	Human-baseline model	Human-GPT-4 (rating-rationale)	Human-GPT-4 (rationale-rating)	Human-GPT-4 (rating-only)
Exact agreement	0.36 (0.31-0.40)	**0.56** (0.52-0.60)	0.53 (0.49-0.58)	0.51 (0.46-0.56)
Adjacent agreement	0.82 (0.78-0.85)	0.96 (0.95-0.98)	**0.97** (0.95-0.98)	0.95 (0.93-0.97)
Cohen’s Kappa	0.13 (0.08-0.18)	**0.43** (0.38-0.48)	0.40 (0.36-0.46)	0.36 (0.31-0.42)
Quadratic Weighted Kappa	0.39 (0.33-0.45)	0.81 (0.79-0.84)	**0.82** (0.79-0.85)	0.78 (0.75-0.82)
Spearman’s rho	0.47 (0.39-0.53)	**0.82** (0.79-0.85)	**0.82** (0.79-0.85)	0.79 (0.76-0.83)

Sample Response: I think that we can learn different cultures from using books and watching movies and tv shows. We can be able to get to know how they live, eat, dance and party. There are many advantages and disadvantages learning about different cultures through books or movies. Firstly, the advantages of learning about the cultures are we can get to learn morals, codes and ethics of that certain culture and know their language, hardwork that they do in that characters and mostly the people mentality. Secondly, disadvantages are we acn also get encounter dru

Human Rationale:

The response has variable coherence. A relevant position is presented and a point in favor of using media to learn about different cultures is given. However, the topic is not developed, with the **second point about disadvantages ending abruptly** so that the intended message is not conveyed. There is some appropriate **use of basic discourse markers (e.g., “Firstly”, “Secondly”).** The language is sufficient to convey basic ideas on the topic, but lexis from the prompt is used repetitively and at times coherence is affected (e.g., “hardwork that they do in that characters”). Overall, the response produces a satisfactory effect on the reader, **negatively impacted by language inaccuracies and lack of development**

GPT-4 Rationale:

The response has variable coherence. The writer begins by stating that it is possible to learn about different cultures through books, movies, and TV shows, and provides some examples of what can be learned. The writer then moves on to discuss the advantages and disadvantages of this approach. **However, the response is unfinished and the disadvantages are not fully developed.** The response demonstrates some organization, with the **use of discourse markers such as “Firstly” and “Secondly”** to link ideas. However, there are some inaccuracies in grammar and vocabulary that impact coherence, for example “we acn also get encounter dru”. Overall, the response is somewhat effective in conveying the intended message, but the **unfinished nature of the response and inaccuracies in language limit the rating to a 3.**

Takeaways

- LLM-based AWE systems show strong early potential for evaluating essays holistically and on difficult-to-capture sub-constructs
- Rationales show limited impacts on ratings, but could be used as feedback to test takers for ways to improve essays
  - **Must not be interpreted as LLM reasoning**
- Further research needed to mitigate biases in LLMs to ensure fairness in operational systems

Thanks!
Questions?