QCRI
معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute
جامعة حمد بن خليفة
HAMAD BIN KHALIFA UNIVERSITY

جامعة حمد بن خليفة
HAMAD BIN KHALIFA UNIVERSITY
كلية العلوم الإنسانيّة والاجتماعيّة
College of Humanities & Social Sciences

جامعة قطر
QATAR UNIVERSITY

TOBB ETÜ
University of Economics & Technology

GenAI Content Detection Task 2: AI vs. Human – Academic Essay Authenticity Challenge

Shammur Absar Chowdhury, Hind Almerekhi, Mucahid Kutlu,
Kaan Efe Keleş, Fatema Ahmad, Tasnim Mohiuddin,
George Mikros, Firoj Alam

GenAI Content Detection Workshop

19 January, 2025

Motivation

- Usage of Generative AI (GenAI) is rapidly increasing.
- Excessive use and over-dependency hinders critical thinking, student learning and creativity.
- Can give rise to potential academic dishonest behaviour.

We urgently need solutions to detect AI-generated content, ensure fair evaluation, and foster student learning and creativity.

Tasks

"Given an essay, identify whether it is generated by a machine or authored by a human."

- This is a binary classification task
- The task is offered two in languages: English and Arabic

Training Datasets: Human-authored English Essays

- ***IELTS Writing Scored Essays***:
  - 1200 academic essays for varieties of prompts
- ***ETS Corpus of Non-Native Written English***:
  - 12,100 academic essays,
  - 8 different prompts,
  - Non-native speakers from 11 different countries, (part TOEFL English proficiency exam)

Metadata includes: Study level, Country, Proficiency level, Scores obtained

Training Datasets: Human-authored Arabic Essays

- ***Arabic Learner Corpus (ALC)***
  - 1,197 essays written by both native and non-native Arabic pre-university/university speakers from 67 nationalities
- ***Qatari Corpus of Argumentative Writing (QCAW) dataset***
  - a collection of 195 argumentative essays written by native Arabic undergraduate students
- ***The CERCLL corpus****
  - 270 essays written by non-native (L2) and heritage Arabic speakers

* Manual validation and annotation have been done on this dataset to use the data in this shared task.

Training Datasets: Human-authored Arabic Essays

Challenges with collecting Arabic essays

- ***Limited Digitization and Online Accessibility***
    - Many essays remain undigitized, and OCR tools for Arabic are unreliable.
- ***Inconsistent Metadata and Anonymization***
    - Metadata is often incorrect, and anonymization is inconsistent.
- ***Linguistic Diversity and Quality***
    - Essays vary in quality and are sometimes written in dialectal Arabic.
- ***Fragmented and Unreliable Archival Systems***
    - Broken links and missing backups make essays difficult to access.

Training Datasets: Machine Generated Essays

7 different open and closed Large Language Models (LLMs)

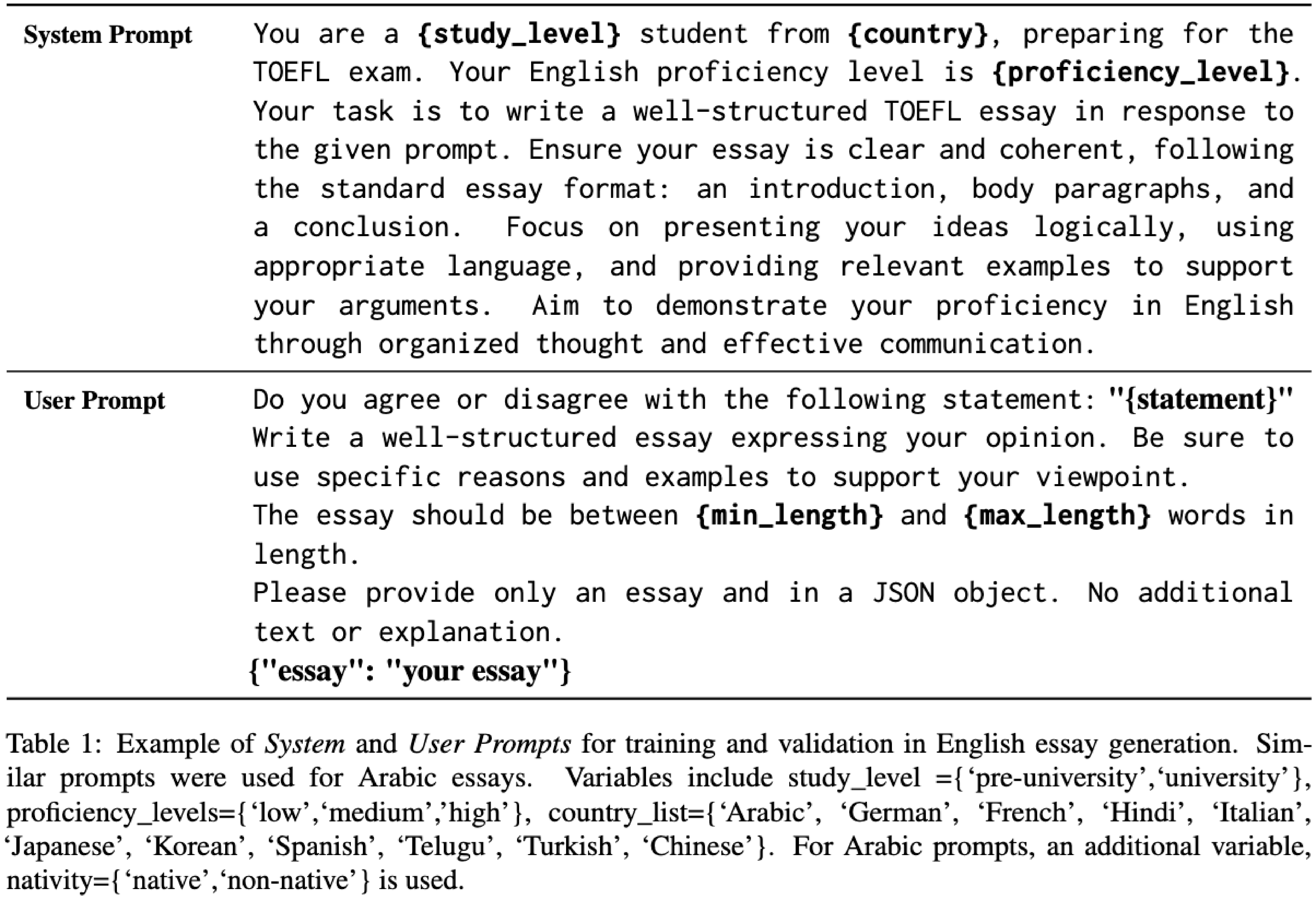
- GPT-3.5-Turbo,
- GPT-4o,
- GPT-4o-mini,
- Gemini-1.5,
- Llama-3.1 (8B),
- Phi-3.5-mini and Claude-3.5

Training Datasets: Machine Generated Essays

Train &

Development set

Prompt



System Prompt	You are a **{study_level}** student from **{country}**, preparing for the TOEFL exam. Your English proficiency level is **{proficiency_level}**. Your task is to write a well-structured TOEFL essay in response to the given prompt. Ensure your essay is clear and coherent, following the standard essay format: an introduction, body paragraphs, and a conclusion. Focus on presenting your ideas logically, using appropriate language, and providing relevant examples to support your arguments. Aim to demonstrate your proficiency in English through organized thought and effective communication.
User Prompt	Do you agree or disagree with the following statement: **"{statement}"** Write a well-structured essay expressing your opinion. Be sure to use specific reasons and examples to support your viewpoint. The essay should be between **{min_length}** and **{max_length}** words in length. Please provide only an essay and in a JSON object. No additional text or explanation. **{"essay": "your essay"}**

Table 1: Example of *System* and *User Prompts* for training and validation in English essay generation. Similar prompts were used for Arabic essays. Variables include study_level ={'pre-university','university'}, proficiency_levels={'low','medium','high'}, country_list={'Arabic', 'German', 'French', 'Hindi', 'Italian', 'Japanese', 'Korean', 'Spanish', 'Telugu', 'Turkish', 'Chinese'}. For Arabic prompts, an additional variable, nativity={'native','non-native'} is used.

Test Dataset:

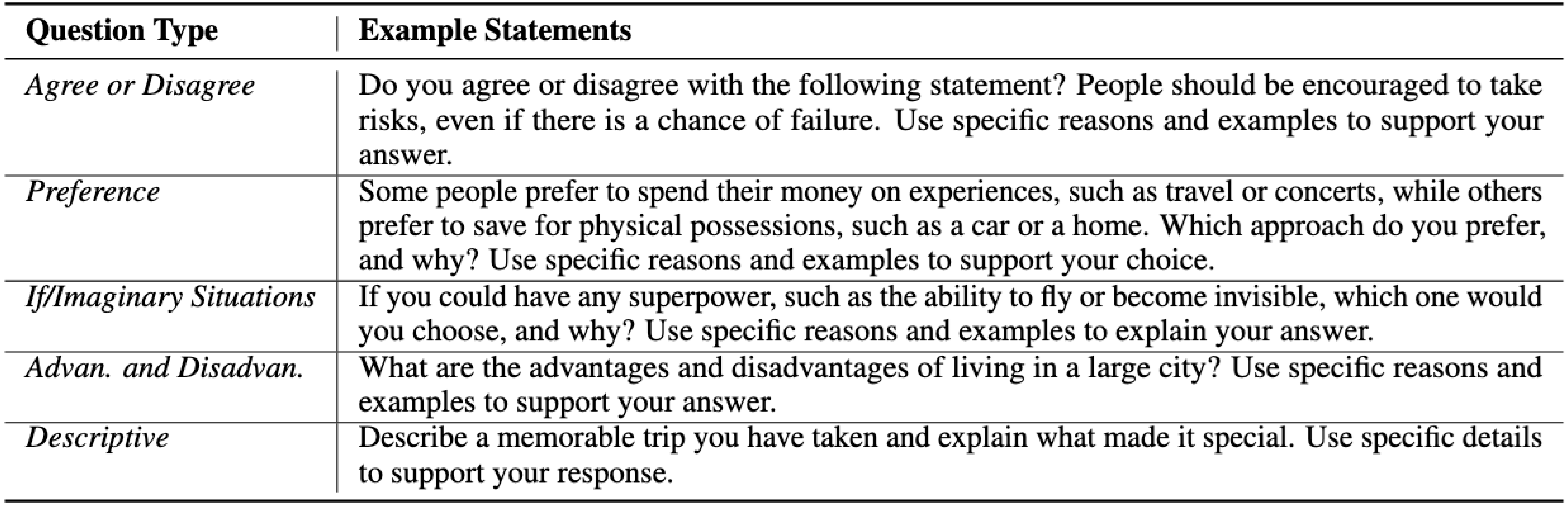
Introducing

GRACE: Generated and Real Academic Corpus for Evaluation

- **Bilingual (Arabic + English)**
- **Human-authored Essays**
- **AI-generated Essays**
  - Freehand Generation
  - Paraphrasing Human-written text
- Include different essay types
- Human-authored essays manually anonymized

Test Datasets: Human-authored – English + Arabic

- **Essay Writing by Recruited Participants**
  - Five different essay types, and under each type, we created several essay statements
  - A team of five trained annotators was recruited



Question Type	Example Statements
Agree or Disagree	Do you agree or disagree with the following statement? People should be encouraged to take risks, even if there is a chance of failure. Use specific reasons and examples to support your answer.
Preference	Some people prefer to spend their money on experiences, such as travel or concerts, while others prefer to save for physical possessions, such as a car or a home. Which approach do you prefer, and why? Use specific reasons and examples to support your choice.
If/Imaginary Situations	If you could have any superpower, such as the ability to fly or become invisible, which one would you choose, and why? Use specific reasons and examples to explain your answer.
Advan. and Disadvan.	What are the advantages and disadvantages of living in a large city? Use specific reasons and examples to support your answer.
Descriptive	Describe a memorable trip you have taken and explain what made it special. Use specific details to support your response.

Examples of different question types and corresponding essay statements (prompts)

Test Datasets: Human-authored – English

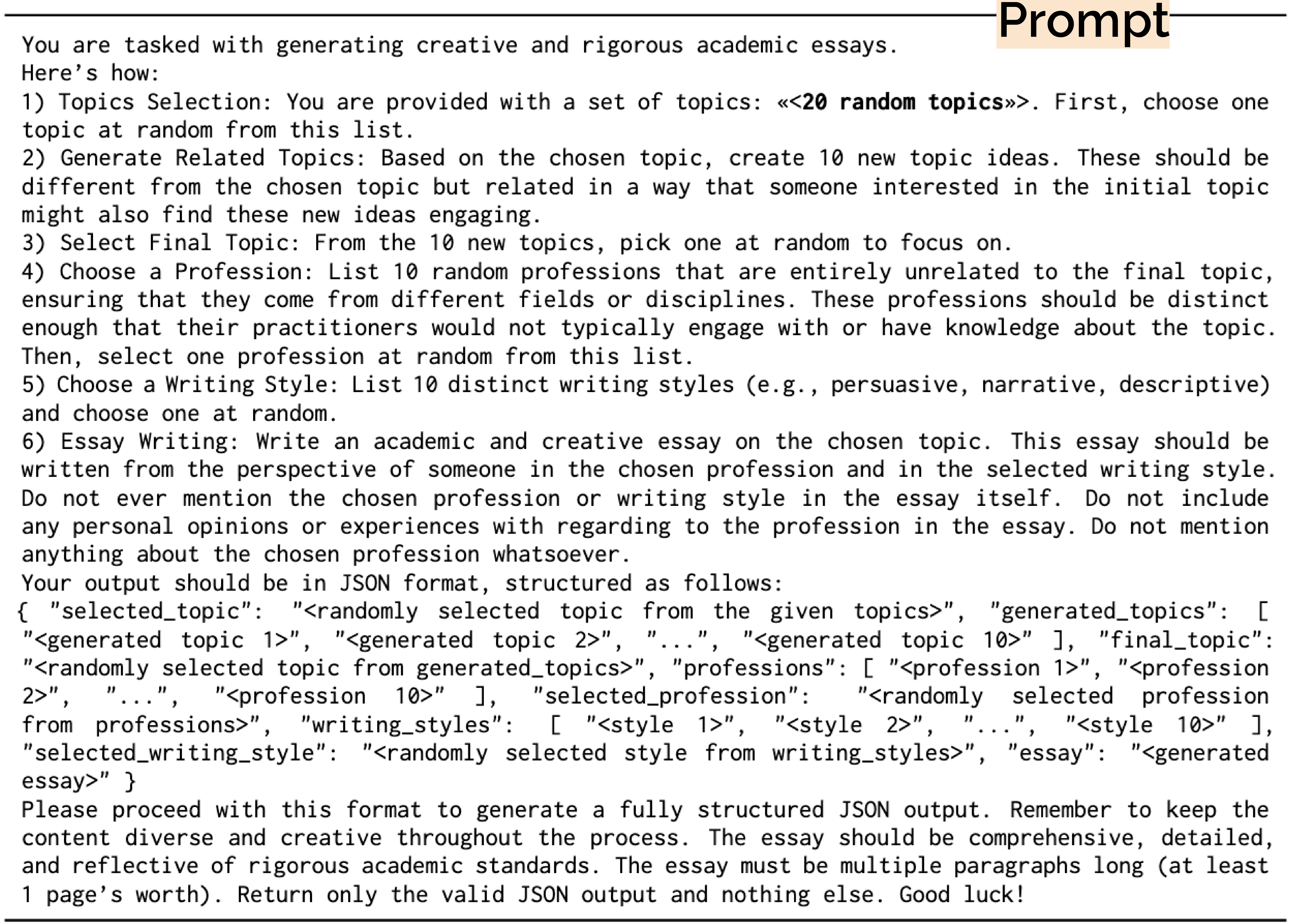
- **Collected Student-authored essays submitted as a part of university-level course**
- **Manual Removal/Anonymization of Personal Information**
  - Author Identification Removal
  - Private Entity Information
  - Sensitive Content
  - Consistency

Test Datasets: AI Generated - English + Arabic

Test set

Freehand prompt used to generate AI generated essays for the final test set

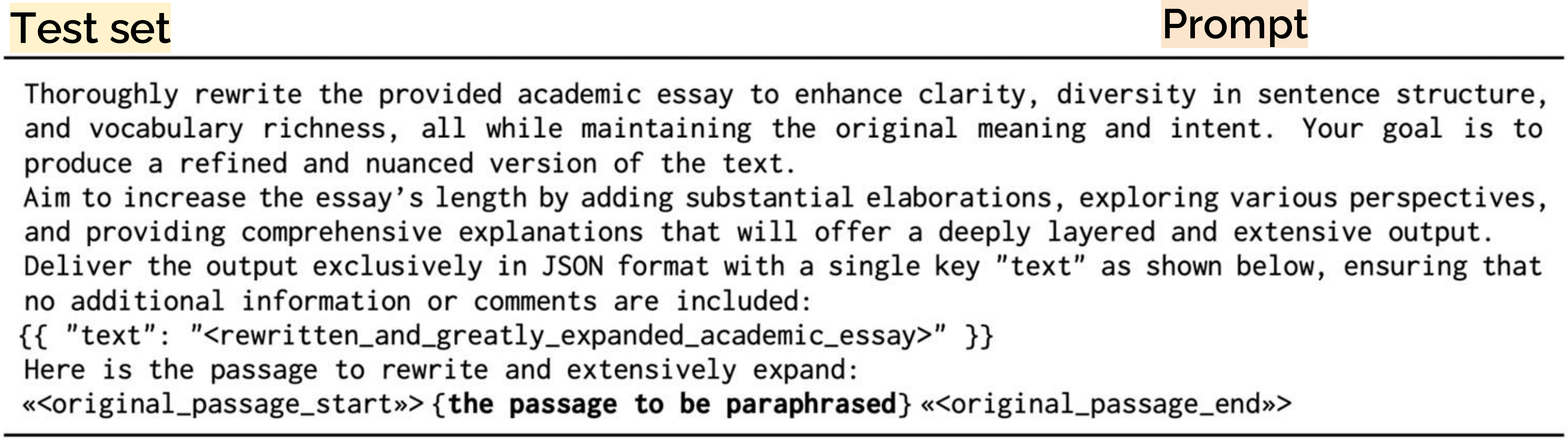
Prompt



You are tasked with generating creative and rigorous academic essays.
Here’s how:
1) Topics Selection: You are provided with a set of topics: «<**20 random topics**»>. First, choose one topic at random from this list.
2) Generate Related Topics: Based on the chosen topic, create 10 new topic ideas. These should be different from the chosen topic but related in a way that someone interested in the initial topic might also find these new ideas engaging.
3) Select Final Topic: From the 10 new topics, pick one at random to focus on.
4) Choose a Profession: List 10 random professions that are entirely unrelated to the final topic, ensuring that they come from different fields or disciplines. These professions should be distinct enough that their practitioners would not typically engage with or have knowledge about the topic. Then, select one profession at random from this list.
5) Choose a Writing Style: List 10 distinct writing styles (e.g., persuasive, narrative, descriptive) and choose one at random.
6) Essay Writing: Write an academic and creative essay on the chosen topic. This essay should be written from the perspective of someone in the chosen profession and in the selected writing style. Do not ever mention the chosen profession or writing style in the essay itself. Do not include any personal opinions or experiences with regarding to the profession in the essay. Do not mention anything about the chosen profession whatsoever.
Your output should be in JSON format, structured as follows:
{ "selected_topic": "<randomly selected topic from the given topics>", "generated_topics": ["<generated topic 1>", "<generated topic 2>", "...", "<generated topic 10>"], "final_topic": "<randomly selected topic from generated_topics>", "professions": ["<profession 1>", "<profession 2>", "...", "<profession 10>"], "selected_profession": "<randomly selected profession from professions>", "writing_styles": ["<style 1>", "<style 2>", "...", "<style 10>"], "selected_writing_style": "<randomly selected style from writing_styles>", "essay": "<generated essay>" }
Please proceed with this format to generate a fully structured JSON output. Remember to keep the content diverse and creative throughout the process. The essay should be comprehensive, detailed, and reflective of rigorous academic standards. The essay must be multiple paragraphs long (at least 1 page’s worth). Return only the valid JSON output and nothing else. Good luck!

Test Datasets: AI Generated - English + Arabic

Paraphrasing prompt used to generate AI generated essays for the final test set.

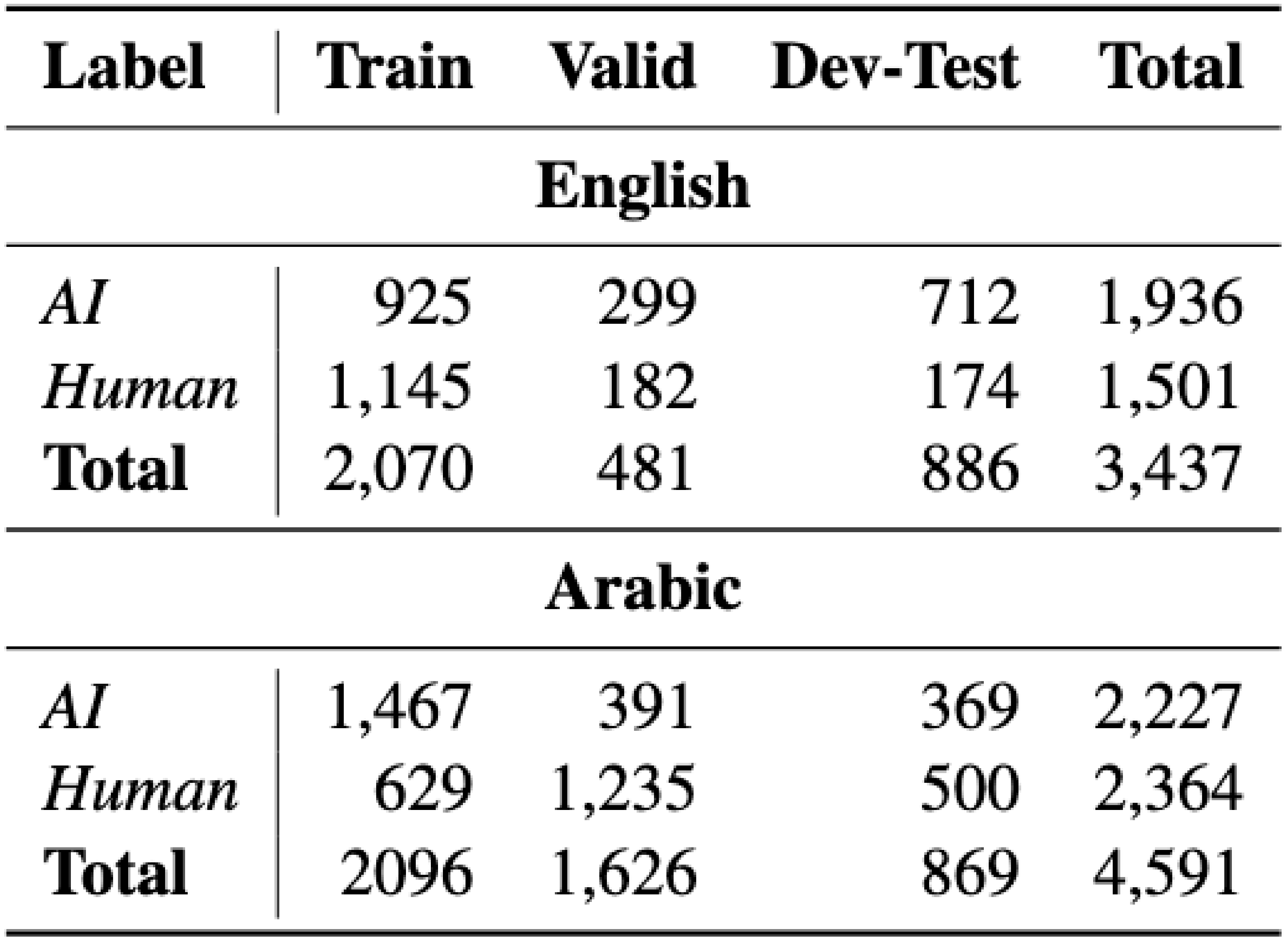


Test set   Prompt

Thoroughly rewrite the provided academic essay to enhance clarity, diversity in sentence structure, and vocabulary richness, all while maintaining the original meaning and intent. Your goal is to produce a refined and nuanced version of the text.
Aim to increase the essay's length by adding substantial elaborations, exploring various perspectives, and providing comprehensive explanations that will offer a deeply layered and extensive output.
Deliver the output exclusively in JSON format with a single key "text" as shown below, ensuring that no additional information or comments are included:
{{ "text": "<rewritten_and_greatly_expanded_academic_essay>" }}
Here is the passage to rewrite and extensively expand:
«<original_passage_start»> {**the passage to be paraphrased**} «<original_passage_end»>

Datasets: Splits

Development phase: dataset and label distribution

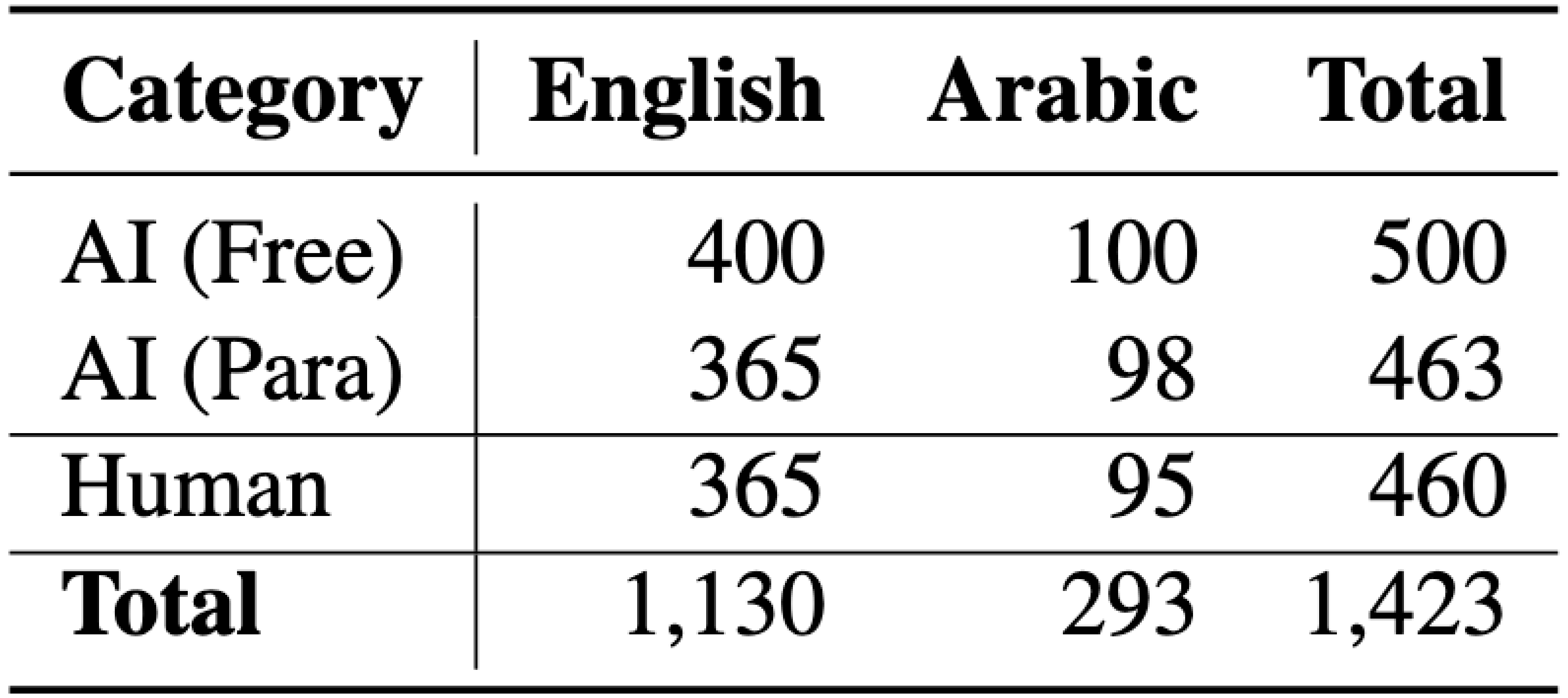


Label	Train	Valid	Dev-Test	Total
English				
AI	925	299	712	1,936
Human	1,145	182	174	1,501
Total	2,070	481	886	3,437
Arabic				
AI	1,467	391	369	2,227
Human	629	1,235	500	2,364
Total	2096	1,626	869	4,591

Distribution of essays by category and language across the test set

Free - freehand generation,

Para - paraphrasing-based generation.



Category	English	Arabic	Total
AI (Free)	400	100	500
AI (Para)	365	98	463
Human	365	95	460
Total	1,130	293	1,423

Evaluation Setup

Development phase:

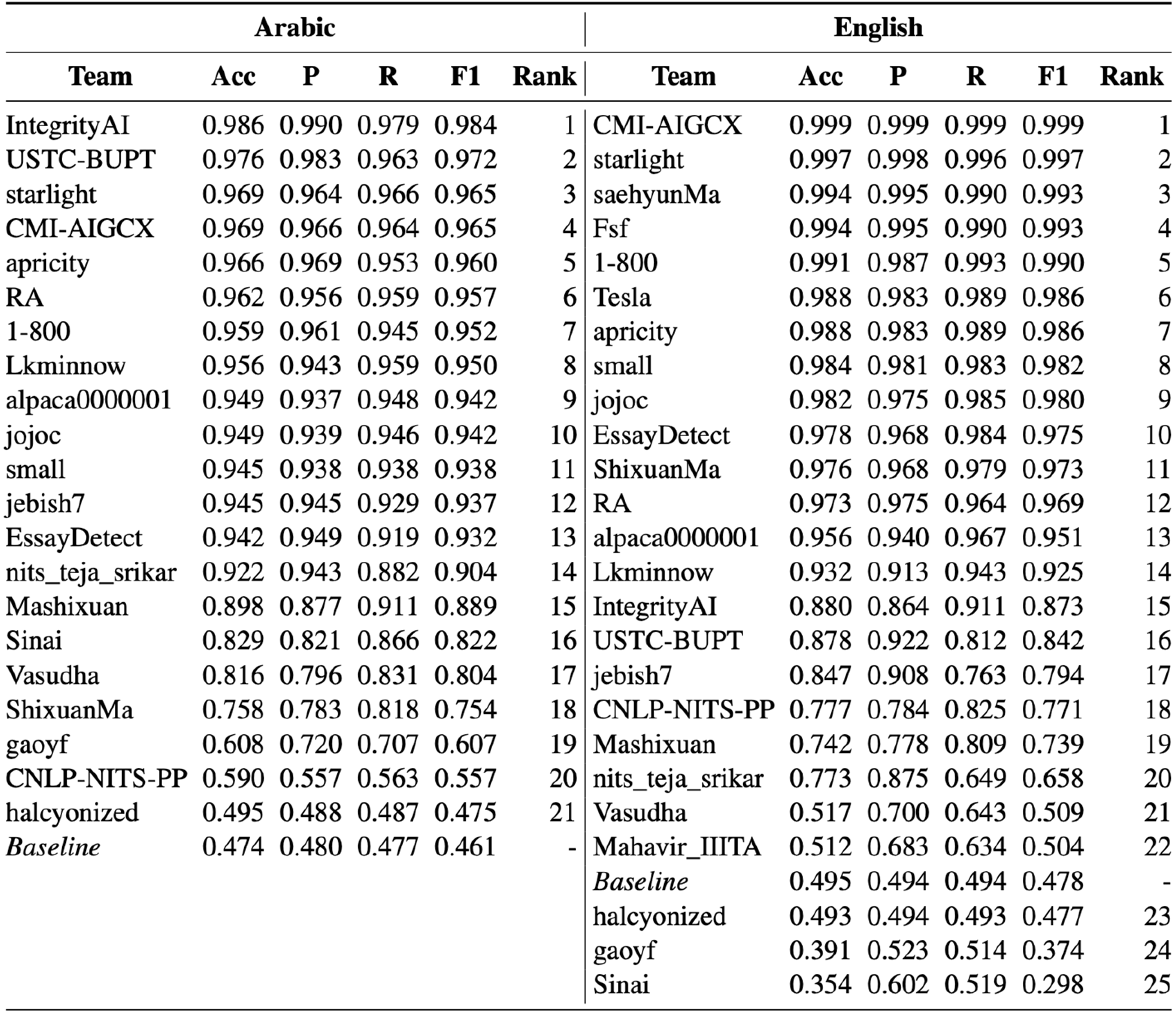
- Released the train and validation subsets
- Participants submitted runs on the dev-test set
- Using competition platform on Codalab

Evaluation phase:

- Released the official test subset – GRACE
- Participants were given four days to submit their final predictions.
- Only the latest submission from each team was considered for final team ranking.

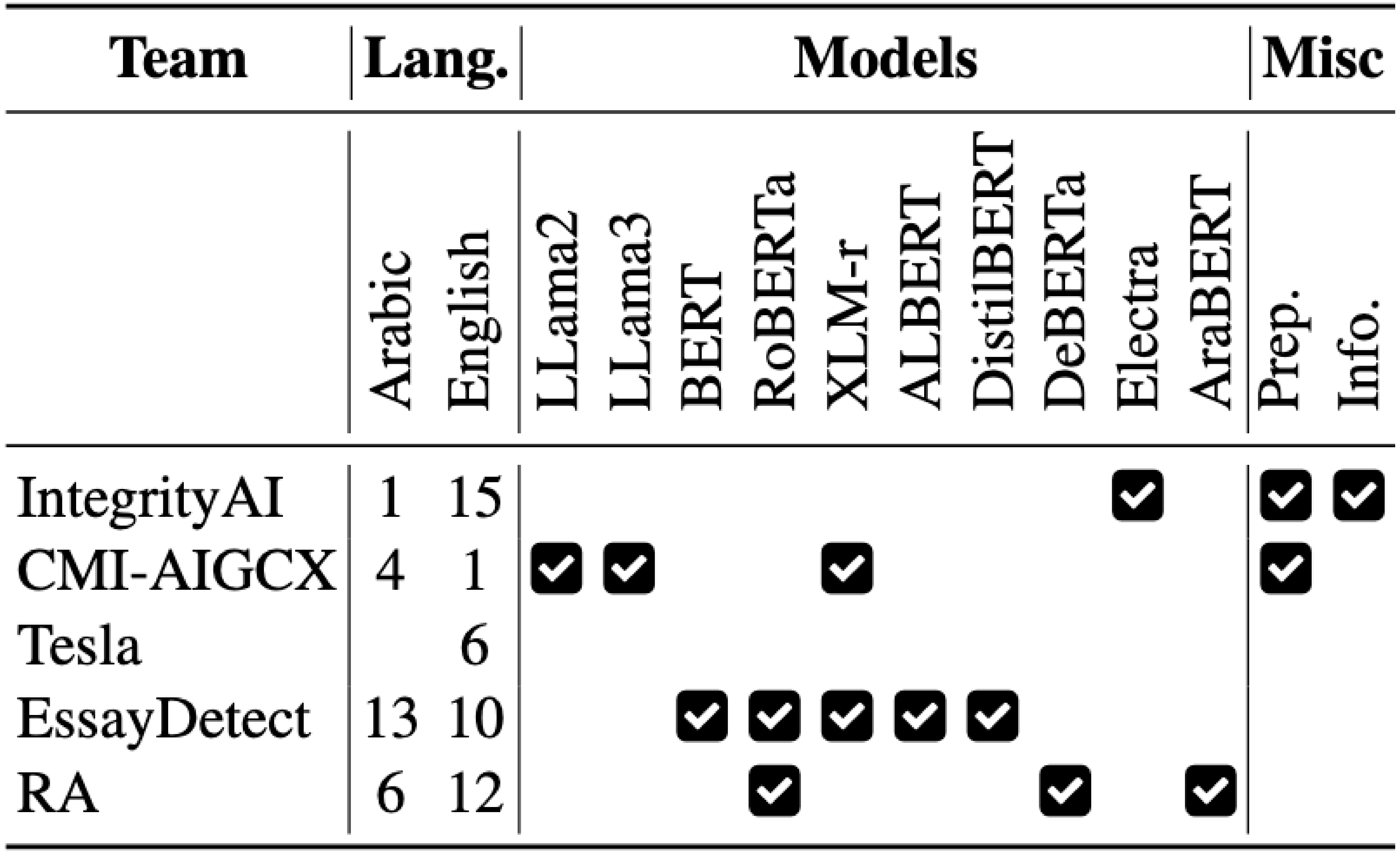
Evaluation Measure: macro-F1 (official ranking)

Results



Arabic						English					
Team	**Acc**	**P**	**R**	**F1**	**Rank**	**Team**	**Acc**	**P**	**R**	**F1**	**Rank**
IntegrityAI	0.986	0.990	0.979	0.984	1	CMI-AIGCX	0.999	0.999	0.999	0.999	1
USTC-BUPT	0.976	0.983	0.963	0.972	2	starlight	0.997	0.998	0.996	0.997	2
starlight	0.969	0.964	0.966	0.965	3	saehyunMa	0.994	0.995	0.990	0.993	3
CMI-AIGCX	0.969	0.966	0.964	0.965	4	Fsf	0.994	0.995	0.990	0.993	4
apricity	0.966	0.969	0.953	0.960	5	1-800	0.991	0.987	0.993	0.990	5
RA	0.962	0.956	0.959	0.957	6	Tesla	0.988	0.983	0.989	0.986	6
1-800	0.959	0.961	0.945	0.952	7	apricity	0.988	0.983	0.989	0.986	7
Lkminnow	0.956	0.943	0.959	0.950	8	small	0.984	0.981	0.983	0.982	8
alpaca0000001	0.949	0.937	0.948	0.942	9	jojoc	0.982	0.975	0.985	0.980	9
jojoc	0.949	0.939	0.946	0.942	10	EssayDetect	0.978	0.968	0.984	0.975	10
small	0.945	0.938	0.938	0.938	11	ShixuanMa	0.976	0.968	0.979	0.973	11
jebish7	0.945	0.945	0.929	0.937	12	RA	0.973	0.975	0.964	0.969	12
EssayDetect	0.942	0.949	0.919	0.932	13	alpaca0000001	0.956	0.940	0.967	0.951	13
nits_teja_srikar	0.922	0.943	0.882	0.904	14	Lkminnow	0.932	0.913	0.943	0.925	14
Mashixuan	0.898	0.877	0.911	0.889	15	IntegrityAI	0.880	0.864	0.911	0.873	15
Sinai	0.829	0.821	0.866	0.822	16	USTC-BUPT	0.878	0.922	0.812	0.842	16
Vasudha	0.816	0.796	0.831	0.804	17	jebish7	0.847	0.908	0.763	0.794	17
ShixuanMa	0.758	0.783	0.818	0.754	18	CNLP-NITS-PP	0.777	0.784	0.825	0.771	18
gaoyf	0.608	0.720	0.707	0.607	19	Mashixuan	0.742	0.778	0.809	0.739	19
CNLP-NITS-PP	0.590	0.557	0.563	0.557	20	nits_teja_srikar	0.773	0.875	0.649	0.658	20
halcyonized	0.495	0.488	0.487	0.475	21	Vasudha	0.517	0.700	0.643	0.509	21
Baseline	0.474	0.480	0.477	0.461	-	Mahavir_IIITA	0.512	0.683	0.634	0.504	22
						Baseline	0.495	0.494	0.494	0.478	-
						halcyonized	0.493	0.494	0.493	0.477	23
						gaoyf	0.391	0.523	0.514	0.374	24
						Sinai	0.354	0.602	0.519	0.298	25

Participants



Team	Lang.		Models										Misc	
	Arabic	English	LLama2	LLama3	BERT	RoBERTa	XLM-r	ALBERT	DistilBERT	DeBERTa	Electra	AraBERT	Prep.	Info.
IntegrityAI	1	15									✓		✓	✓
CMI-AIGCX	4	1	✓	✓			✓						✓	
Tesla		6												
EssayDetect	13	10			✓	✓	✓	✓	✓					
RA	6	12				✓				✓		✓		

Overview of the approaches. The numbers in the language box refer to the position of the team in the official ranking. Prep.: Preprocessing. Info.: Info. Extraction.

Summary

- A total of 56 teams registering to participate in the development and evaluation phases.
- 21 teams submitted official results on the test set for Arabic, and
- 25 teams did so for English.
- Finally, seven teams submitted task description papers

Thank You