

# KRISTEVA: Close Reading as a Novel Task for Benchmarking Interpretive Reasoning

Peiqi Sui<sup>1\*</sup> Juan Diego Rodriguez<sup>2</sup> Philippe Laban<sup>3</sup>  
Dean Murphy<sup>2</sup> Joseph P. Dexter<sup>4</sup> Richard Jean So<sup>1</sup> Samuel Baker<sup>2</sup> Pramit Chaudhuri<sup>2</sup>  
<sup>1</sup>McGill University <sup>2</sup>UT Austin <sup>3</sup>Microsoft Research <sup>4</sup>University of Macau  
\*peiqi.sui@mail.mcgill.ca

## Abstract

Each year, tens of millions of essays are written and graded in college-level English courses. Students are asked to analyze literary and cultural texts through a process known as close reading, in which they gather textual details to formulate evidence-based arguments. Despite being viewed as a basis for critical thinking and widely adopted as a required element of university coursework, close reading has never been evaluated on large language models (LLMs), and multi-discipline benchmarks like MMLU do not include literature as a subject. To fill this gap, we present KRISTEVA, the first close reading benchmark<sup>1</sup> for evaluating interpretive reasoning, consisting of 1331 multiple-choice questions adapted from classroom data. With KRISTEVA, we propose three progressively more difficult sets of tasks to approximate different elements of the close reading process, which we use to test how well LLMs may seem to understand and reason about literary works: 1) extracting stylistic features, 2) retrieving relevant contextual information from parametric knowledge, and 3) multi-hop reasoning between style and external contexts. Our baseline results find that, while state-of-the-art LLMs possess some college-level close reading competency (accuracy 49.7% - 69.7%), their performances still trail those of experienced human evaluators on 10 out of our 11 tasks.

“It is not surprising that the detailed analysis of metaphors... sometimes feels like extracting cube-roots in the head.”

— I.A. Richards, (1936)

## 1 Background

Close reading is “the detailed analysis of the complex interrelations and ambiguities (multiple meanings) of the verbal and figurative components

<sup>1</sup>Our benchmark is publicly available on huggingface (<https://huggingface.co/datasets/McGill-NLP/KRISTEVA>).

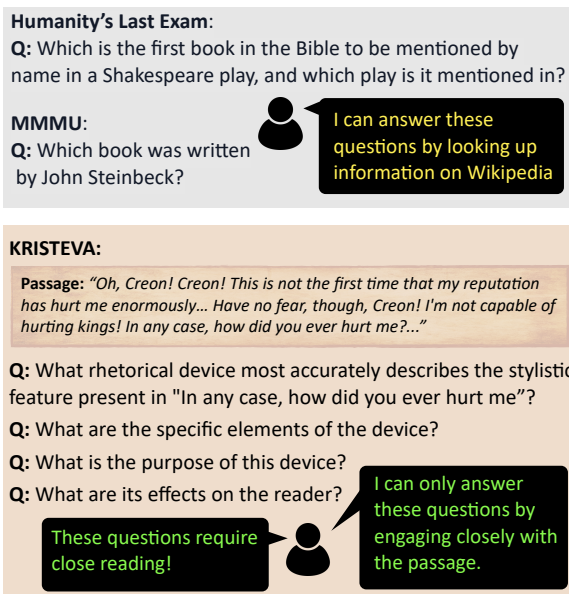


Figure 1: Examples of KRISTEVA questions that require interpretive reasoning to answer, compared to the purely informational literature questions from existing benchmarks.

within a [literary] work” (Abrams, 2009, 217). As a uniquely text-centric form of interpretive reasoning, close reading methodologies posit that aesthetic choices about literary and cultural texts are not trivially subjective or arbitrary preferences. Instead, such methodologies pay meticulous attention to how the workings of language, form, and style generate meaning, rigorously observing, analyzing, and leveraging formal and stylistic features they present as textual evidence for falsifiable claims about literary or cultural texts.

As a skill, close reading has long been considered essential for cultivating critical thinking competencies that underpin active and informed participation in the civil discourse of participatory democracies (Dewey, 1910). Recent studies argue that interpretive skills developed through close reading can be transferred to raising awareness of social jus-

tice (Hooks, 1994), recognizing disinformation and misinformation (Carillo, 2018; McGrew, 2020), developing digital and media literacy (Hayles, 2010; Hobbs, 2010), and fostering empathy (Charon, 2006). Reflecting its importance, close reading has been adopted both as a common requirement in college-level coursework (Bialostosky, 2006) and as a pedagogical benchmark for secondary education (CommonCore, 2012). This policy trajectory aligns with predominant viewpoints within the humanities. Literary and cultural studies scholars argue that close reading enables individuals to better recognize and articulate patterns of political, social, and economic meaning (Levine, 2015), thereby potentially converging on common understandings and reducing polarization. Evaluating LLMs for close reading not only measures their ability to interpret literature, but could also clarify what it means for such models to have the foundational skills in critical reasoning about complex social issues.

However, no existing benchmark directly assesses LLMs on their ability to perform close reading. This omission is reflective of a broader gap in reasoning benchmarks—specifically, a lack of benchmark tasks where there may be no definitive “right” answer but certain responses can be deemed clearly wrong or unreasonable. Notably, very few benchmarks evaluate LLMs for aesthetic judgment (Hullman et al., 2023), a form of reasoning that requires negotiating a balance between components that are subjective (with no strictly correct answers) and objective (with demonstrably wrong answers). Some pioneering work in this area has explored the ability of LLMs to display an understanding of humor (Hessel et al., 2023) or music (Yuan et al., 2024). We applaud these efforts, and we aim to further expand the scope of LLM reasoning evaluation to include what we see as close reading’s higher-order, synthetic reasoning and understanding: a task domain in which interpretations are not categorically correct or incorrect but can be judged on their plausibility for supporting arguments that a careful reader would find persuasive (Sinykin and Winant, 2025). Close reading thus exemplifies an omnipresent yet under-studied class of reasoning challenges that hold relativist instead of positivistic ground truth.

To this end, we present KRISTEVA (Close Reading and Interpretive Reasoning with Textual Evidence), the first benchmark that evaluates LLMs for 1) close reading as a form of reasoning pre-

viously overlooked by the NLP community, 2) college-level knowledge in the literary domain, and 3) figurative language understanding as multi-hop reading comprehension (Figure 1). KRISTEVA consists of 1,331 multiple-choice questions extracted from college-level exam data, along with a novel task structure adapted from UT Austin’s *Critical Reader’s Interpretive Toolkit* (CRIT),<sup>2</sup> a heuristic framework widely used to teach close reading at the college level. Our tasks are designed to evaluate how effectively LLMs can perform a sequence of intermediary analytical steps commonly presented in college literature classrooms as an essential pedagogical scaffold that guides students towards producing evidence-based literary interpretations. We find that, while competitive, numerous state-of-the-art LLMs still fall behind the top-line human performance of experienced close readers on these tasks.

For NLP, a particular strength of close reading as a data source is that it organically combines into a unified set of tasks two longstanding but isolated challenges of natural language understanding (NLU): figurative language understanding (FLU) (Chakrabarty et al., 2022b) and multi-hop reading comprehension (Welbl et al., 2018). Our study is also an initial exploration of how the vast quantities of high-quality text data produced in the routine educational activities of humanities departments might be analyzed by and leveraged for NLP.

## 2 Task Description

The KRISTEVA benchmark is adapted from CRIT, a heuristic framework developed by UT Austin’s English department for teaching close reading in literature courses required by UT’s undergraduate program. CRIT breaks close reading down into a step-by-step process: *paraphrase*, *observe*, *contextualize*, *analyze*, *argue*, and *reflect*. Each of its six sequential steps is guided by a set of exploratory questions that significantly reduces the cognitive load required for producing robust, evidence-based textual arguments. Pedagogically, humanities professors who use the CRIT tool have seen a clear positive effect on students’ ability to perform more focused, detailed, and sustained textual analysis (Bares et al., 2020).

To create KRISTEVA, we operationalize three of CRIT’s six steps into 11 distinct and sequential

<sup>2</sup><https://liberalarts.utexas.edu/english/the-critical-reader-s-toolkit.html>



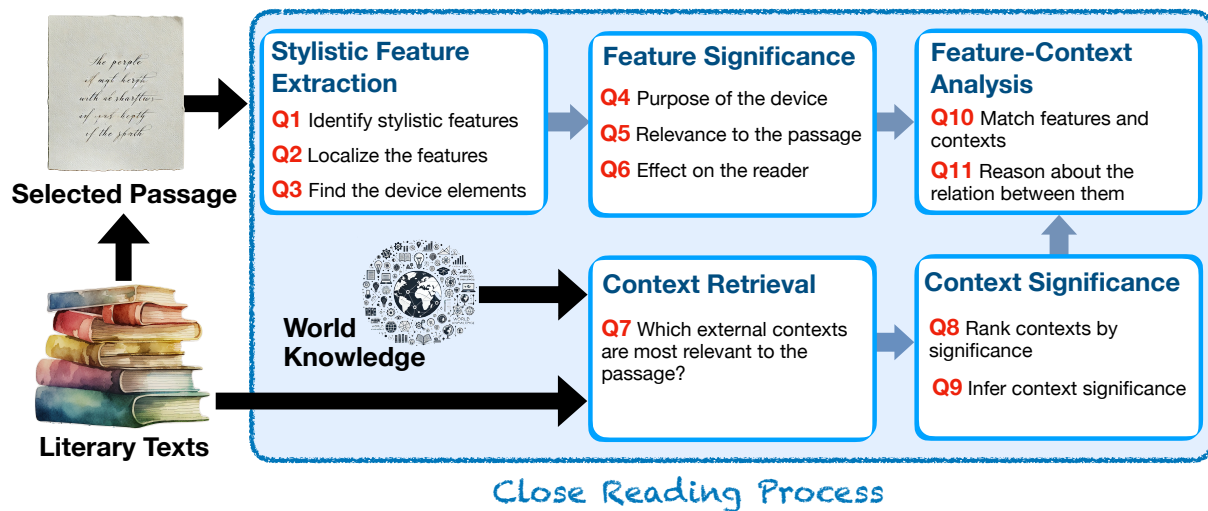


Figure 2: Question types in KRISTEVA correspond to distinct stages in the close reading process, involving both extractive tasks (e.g., stylistic feature extraction) and reasoning tasks (e.g., reasoning about the relation between features and relevant external context).

tasks designed for the evaluation of LLMs, each targeting a discrete cognitive procedure (Table 1). Seven out of the 11 tasks target novel forms of interpretive reasoning, in ascending order of complexity, that are essential building blocks for making a successful evidence-based interpretive argument (Figure 2). Detailed descriptions of each task, and of the CRIT steps to which they correspond, are presented in Appendix B. On a general level, they fall into three progressively more challenging clusters, as described in the following subsections.

**Q1-Q6: Stylistic Features** Drawing on foundational theories of close reading, we define stylistic features as qualitative measures of a literary text’s deviation from general domain language use. Writers often use expressions that have multiple, non-literal meanings and contradict standard, logical relationships; these departures can be justified on aesthetic grounds when they achieve particular effects rarely associated with functional language (Richards, 1929). Consider, for instance, the difference between reading a technical manual and a poem: both require expertise on the part of the reader, but only one creates the expectation that multiple reasonable interpretations—potentially conflicting, or referring to entirely different phenomena—are possible. Stylistic features are knowledge-based representations of such patterns, like figurative language, sonic patterns, poetic form, diction, syntax, and narrative devices, that distinguish literary from non-literary texts.

KRISTEVA focuses on these features because of their importance to close reading, which has been described as a heightened sensitivity to nuanced textual elements and patterns (Guillory, 2025). Three tasks explicitly target the extraction and mapping of such features: detection (Q1), localization (Q2), and elaboration (Q3), along with a fourth task that reasons about the possible purpose of including these features in the passage (Q4). Close reading also entails the judgment of a work’s literary merit—that is, determining whether it successfully leverages its stylistic features to conjure up a compelling enough effect that justifies the cognitive resources required to process these deviations from conventional language use. To evaluate this aspect of interpretive reasoning, two additional questions address a feature’s relative significance within its passage compared to other previously identified features (Q5), and the specific effect it achieves for the reader (Q6). Overall, these tasks follow Sravanthi et al. (2024)’s framing of FLU as pragmatics capabilities, expanding the scope of the existing evaluations (Section 7) to the interpretation of figurative language’s affordances (Q4), significance (Q5), and effectiveness (Q6) as a form of communication.

**Q7-Q9: Contextual Information** We define context as the broader external circumstances within which a literary work is positioned, circumstances that might not be immanent within the text itself but are highly pertinent to its meaning. Plausible contexts include (but are not limited to) cultural,

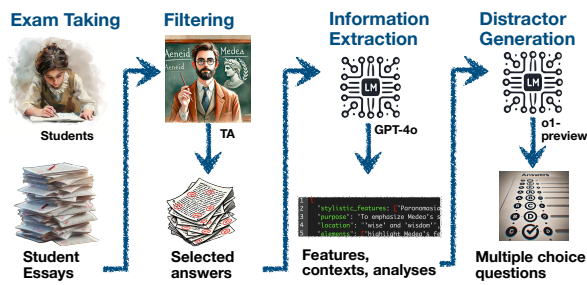


Figure 3: The dataset construction pipeline. Instructors manually filtered students’ essays. We used GPT-4o to extract literary features from essays with the best answers. Finally, we used o1 to generate distractors (incorrect choices) for each multiple-choice question.

historical, literary,<sup>3</sup> and biographical factors that could help enrich the reader’s understanding of the text. KRISTEVA evaluates models for both retrieving relevant contextual frames from parametric knowledge (Q7), and inferring their relative significance to the passage (Q8-Q9).

**Q10–Q11: Multi-hop Reasoning Between Features and Contexts** Multi-hop reasoning requires chaining together multiple pieces of information, often across multiple documents or from external knowledge sources, to perform inference that cannot be derived from any single piece alone. In the case of close reading, multi-hop reasoning more specifically involves reasoning between a passage, features extracted from it, and contexts external to it. Since the deployment of stylistic features often involves managing trade-offs between language efficiency and the potential for emergent meaning, multi-hop reasoning is crucial for interpreting how the interplay between form and content can reveal novel semantic meaning not immediately available in the text itself. Such reasoning also drives aesthetic judgments in the literary domain, insofar as such judgments parse a text’s context-grounded features in order to ascertain whether the outcome of such trade-offs renders a given text worthwhile. Although some earlier tasks could also be thought to involve multi-hop reasoning, Q10-11 directly require the combination of multiple pieces of contexts through matching features identified in previous questions to corresponding contexts (Q10), and inferring the most plausible connection between a given feature-context pair (Q11).

<sup>3</sup>Here “literary” factors include contextual information about literature as a field of social practice, such as what influences, generic conventions, or political constraints may have been relevant to the composition of the text.

## 3 Dataset Construction

### 3.1 Data Collection

#### 3.1.1 Data Source

In educational settings, a minimal but sufficient implementation of CRIT’s heuristic steps typically takes the form of a short analytical essay. As a pedagogical tool, each CRIT essay is designed as a self-contained unit of close reading, some assigned for practice and some for graded evaluation.

We collect 49 de-identified essays and grades from three exams in a university-level literature course that adapts the CRIT framework to help structure its close reading pedagogy. The course (“Introduction to Classical Literature: Forms, Cultures, Histories”) gives mostly first-year college students an introduction to world literature covering multiple historical periods, cultures, and genres. As an introductory course, it provides students with a foundational template for how to interpret literature. This template is partly based on the CRIT framework, although many of its elements are standard in the teaching of literary criticism. A major component of the course examinations is a single analytical essay focused on close-reading of a short literary passage drawn from a work students have previously studied.

#### 3.1.2 Data Processing

The ground truth on data quality is directly established from the final grade the course instructor assigned to each essay. We leverage this grade to filter the collected essays and exclude low-score entries under 80%. In addition, the course gives students the opportunity to revise and resubmit their first two exam essays for regrading, which in most cases result in significant improvements in data quality; when available, we replace the original exam essays with their revised versions.

As discussed in Section 1, close reading differs from general domain reasoning tasks in that there are no objectively correct interpretations of literary works, but rather comparatively more or less reasonable ones. Critics evaluate interpretations on multiple bases often located in very different domains (e.g., how well grounded the interpretation is in a work’s cultural context, or how it is received among some audience or other). This setting poses an epistemic challenge for validating the ground truth, since multiple answers could be correct at the same time. To address this potential issue, we ask the instructor to perform a second manual check to

CRIT Step	KRISTEVA Tasks	In NLP Terms	Reasoning	Distractors
<b>Observe:</b> Identify stylistic features from the passage and analyze their purpose and effect.	Q1 Feature Type			✓
	Q2 Feature Location			✓
	Q3 Feature Elements	Figurative Language Understanding		✓
	Q4 Feature Purpose		✓	✓
	Q5 Feature Significance Ranking		✓	
	Q6 Feature Significance Inference		✓	✓
<b>Contextualize:</b> Provide relevant pieces of cultural, historical, or literary contextual frames.	Q7 Relevant Context Retrieval	Information Retrieval		✓
	Q8 Context Significance Ranking	Multi-hop Reasoning	✓	
	Q9 Context Significance Inference		✓	✓
<b>Analyze:</b> Connect features with contexts and explain how they inform each other.	Q10 Feature-Context Matching	Multi-hop Reasoning	✓	
	Q11 Feature-Context Reasoning		✓	✓

Table 1: KRISTEVA task structure, adapted from UT Austin’s CRIT framework

ensure that for each correct answer it would also be possible to generate three distinct distractors that were less reasonable answers to the question.

### 3.2 Benchmark Construction

We build a pipeline that converts the unstructured texts from close reading essays into 1,331 multiple-choice questions ready for LLM evaluation. The specific prompts we use for the multiple choice question (MCQ) construction pipeline can be found in Appendix C.1.

#### 3.2.1 Question and Answer Extraction

Structured representations of stylistic features, external contexts, and the connection between the two are first extracted from each essay, or summarized if the essay is too long. Some combination of this information, depending on the expected input and output of each task, is then used to construct each type of question and its corresponding answer. A detailed input/output schema of each question type is presented in Table 3.

#### 3.2.2 Generation of Distractors

While some MCQs simply reuse answers from earlier tasks as less significant options, others require the creation of entirely new distractors. We use o1-preview to generate three distractors for each of the 1,178 questions (7/11 question types) that require distractors (prompts shown in Appendix C.1). Each distractor should closely mirror the structure and syntax of the correct answer to seem plausible and avoid confounders, while diverging semantically to present relatively less compelling interpretations.

Distractor generation for MCQs is a challenging problem (Stasaski and Hearst, 2017). In order to ensure the quality of distractors, we experiment with the use of other LLMs to generate distractors (such as GPT-4o and Qwen) and perform manual inspection. Our inspection leads us to conclude that o1-preview generates on average the most relevant and challenging distractors for the close-reading literary domain.

Once distractors are generated for a question, the correct answer and distractors are merged into a final list of answer options, which is then shuffled, ensuring that there are no answer positional biases in the dataset. This shuffled list of answer options is presented in this arbitrary order in all the experiments we conduct.

## 4 Experimental Settings

### 4.1 LLMs

We evaluate a range of language models—from 2B to 70B parameter models across various model families (Qwen, OLMo, Gemma, Llama, Phi, and Mistral), as well as GPT-4o and o1. We only use the instruction-tuned versions of the above models in a zero-shot setting<sup>4</sup>.

Each model is prompted to generate an answer in JSON format. We then extract the answer and performed an exact match with the ground truth to assess accuracy; notably, no outputs were unparseable. The prompt used is provided in Ap-

<sup>4</sup>We omit chain-of-thought style evaluations because these have shown to mainly benefit mathematics and logic-related tasks (Sprague et al., 2024).

	Non-reasoning					Reasoning								Overall
	Q1 (209)	Q2 (144)	Q3 (209)	Q7 (139)	avg	Q4 (166)	Q5 (53)	Q6 (160)	Q8 (42)	Q9 (76)	Q10 (66)	Q11 (67)	avg	
Random	25.2	24.7	25.6	25.0	25.2	22.2	37.7	24.1	33.3	28.2	24.5	23.1	28.5	25.5
Qwen2.5-7B	34.4	91.7	75.1	55.4	62.3	64.5	58.5	73.1	33.3	75.0	27.3	82.1	60.7	62.9
Qwen2.5-14B	40.7	95.8	76.6	54.7	65.2	66.3	58.5	71.3	28.6	77.6	34.8	82.1	61.7	64.8
Qwen2.5-32B	47.4	98.6	83.3	62.6	71.4	66.9	60.4	70.0	35.7	73.7	37.9	88.1	63.3	68.5
OLMo-2-7B	38.3	84.7	57.4	49.6	55.6	37.3	43.4	56.9	31.0	53.9	31.8	61.2	44.3	51.3
OLMo-2-13B	40.2	93.1	70.3	55.4	62.8	57.8	52.8	68.8	33.3	64.5	28.8	76.1	55.4	60.8
OLMoE-1B-7B	29.7	81.2	52.6	46.0	50.2	46.4	49.1	54.4	23.8	64.5	27.3	62.7	48.2	49.7
Gemma-2-2B	34.9	93.8	60.3	46.8	56.6	44.6	41.5	56.2	19.0	72.4	34.8	68.7	48.1	53.9
Gemma-2-9B	41.6	94.4	72.2	58.3	64.7	65.1	58.5	69.4	31.0	77.6	37.9	83.6	62.0	64.4
Gemma-2-27B	43.5	95.8	76.6	62.6	67.8	65.1	58.5	71.9	26.2	76.3	40.9	82.1	61.9	63.6
Llama-3-8B	42.1	94.4	71.8	55.4	64.1	60.2	60.4	63.7	31.0	68.4	40.9	82.1	59.9	62.5
Llama-3.1-8B	43.1	95.1	72.2	56.8	64.9	56.0	58.5	68.1	31.0	72.4	34.8	76.1	58.0	62.5
Llama-3.1-70B	46.9	96.5	78.9	54.7	67.8	63.9	60.4	70.6	31.0	71.1	36.4	88.1	61.9	66.0
Llama-3.3-70B	45.9	97.2	79.4	57.6	68.4	67.5	60.4	73.1	35.7	68.4	37.9	89.6	63.2	67.2
Phi-4	49.3	97.9	83.3	64.7	72.2	67.5	62.3	75.6	35.7	76.3	37.9	83.6	64.3	69.7
Mistral-7B	34.4	92.4	70.8	54.0	60.8	54.2	52.8	65.6	35.7	63.2	33.3	74.6	54.6	59.1
GPT-4o-mini	45.0	94.4	78.0	62.6	68.3	61.5	58.5	73.8	33.3	75.0	40.9	91.0	62.5	66.9
GPT-4o	41.2	96.5	77.5	64.8	67.9	66.3	56.6	75.6	33.3	78.9	43.9	85.1	63.4	67.5
o1-mini	43.5	95.8	77.0	54.0	66.0	57.8	60.4	68.1	35.7	77.6	36.4	79.1	60.4	64.1
o1-preview	40.7	97.2	74.6	67.6	67.8	63.9	49.1	72.5	35.7	77.6	47.0	91.0	61.5	66.8
Evaluator 1	43.5	100.0	52.2	72.2	63.7	63.6	75.0	61.1	0.0	77.8	71.4	100.0	68.7	65.4
Evaluator 2	66.7	100.0	91.7	75.0	82.5	71.4	0.0	71.4	66.7	100.0	40.0	60.0	50.5	74.7
Evaluator 3	65.2	94.1	69.6	64.0	72.0	52.9	0.0	50.0	28.6	70.0	41.7	66.7	39.0	61.5
Weighted Average	57.1	97.5	67.3	69.3	70.8	60.7	28.8	58.5	25.1	78.9	52.8	78.2	50.0	65.6

Table 2: Performance (Acc) of LLMs on the KRISTEVA benchmark in zero-shot setting alongside a human baseline. We use **green** to highlight the best model performance for each question type, and **blue** to highlight where human evaluators equal or outperform the best model. For all models, we report the direct match performance. The number of each type of questions is included in parentheses in the header.

pendix C.

All experiments are conducted with the Language Model Evaluation Harness (Gao et al., 2024)<sup>5</sup> to ensure that our baseline results are reproducible.

## 4.2 Human Evaluation

To approximate a human baseline, we construct a subset of three unit tests, one for each exam by selecting MCQs from three essays per exam. We believe the evaluation results on the subset, which accounts for percentage of the dataset, are an unbiased estimate of the human performance over the whole benchmark.

We employ three experienced close readers (PhD students in the humanities) to answer these questions. Each evaluator completed one or two unit tests, with partial overlap to enable computation of inter-annotator agreement metrics (Section 6). Although most questions can be answered solely on the basis of the passage, we also provide the

evaluators with the same class materials available to students to ensure subject-matter familiarity.

## 5 Results

Table 2 presents the performance of LLMs on the KRISTEVA benchmark, alongside a competitive human baseline. Phi-4 achieves the highest overall accuracy of 69.7, as well as the highest scores in both the reasoning (64.3) and non-reasoning (72.2) categories. Meanwhile, the o1-preview model, the largest reasoning model included in our experiment, stands out on most questions that require multi-hop reasoning. For most model families, larger variants generally outperform smaller ones (e.g., Qwen2.5-32B vs. Qwen2.5-14B), with the exception of Gemma-2-27B.

The top-line human performance surpasses the best-performing LLMs on 10 out of 11 tasks, generally by a wide margin. In addition, the best-performing human overall (evaluator 2) outperforms the best model (Phi-4) on 8 tasks. Notably, there is greater variability in performance among evaluators (average pairwise standard deviation

<sup>5</sup><https://github.com/EleutherAI/lm-evaluation-harness>, version 0.4.7



of 29.3) compared to LLMs (5.47). In addition, human performance varies more across question types, with a coefficient of variation of 0.434 versus 0.322 for LLMs.

## 6 Discussion

### Do LLMs outperform experienced humans on close reading?

While some models, most notably Phi-4, can approximate human-level performance on overall accuracy, the more fine-grained task-by-task breakdown shows that humans maintain a clear advantage in most aspects of close reading. The best performing models trail behind their human counterparts on 10 out of 11 tasks, eight of them more than 8% in accuracy. This gap is likely even more pronounced in other evaluation settings, as our human baseline reported in Table 2 likely represents a conservative estimate: although our evaluators are experienced close readers, they still need a period of adjustment to KRISTEVA’s MCQ format, which differs significantly from the actual practice of close reading. Two of the three evaluators reported considerable initial difficulty in cognitively adapting to the format of the questions, while another believed that time constraints limited their performance. These factors suggest that under more natural, open-ended evaluation settings of close reading, human performance would likely be even higher.

### Do human evaluators agree with each other?

To assess the consistency of human judgments, we computed Krippendorff’s  $\alpha$  for evaluator pairs overlapping on the same unit tests. Agreement scores range from 0.523 (unit test 3) to 0.644 (unit test 2), indicating moderate consensus. Notably, the lower agreement score corresponds to evaluators from different departments (English and Classics), while evaluators from the same department (Classics) display higher agreement.

In contrast to the relative consistency observed among LLMs, human performance exhibits more significant variability. For some question types, one evaluator achieved high accuracy while another scored zero (Q5, Q8). It is important to acknowledge that such differences may simply be the result of the small sample size (i.e., only three evaluators). We hypothesize, however, that these discrepancies could also be influenced by domain expertise. The English literature PhD student (evaluator 1) excels on the more complex reasoning questions, perhaps due to greater familiarity with the CRIT framework

or the theory of criticism underlying the approach. Meanwhile, the two Classics PhD students, who may have greater subject-matter familiarity with the particular types of literary texts or their historical circumstances, perform better on the extraction-based questions, as well as on the specific reasoning tasks involving external contexts. Our results suggest that, in small-sample settings, diversity in academic backgrounds and areas of specialization may drive volatility in human performance—a factor that future work should consider when defining human baselines for close reading.

### What makes an LLM good at close reading?

Although the best performing Phi-4 is a smaller model (14B), its high-quality, textbook-based training data might have a closer affinity to the college classroom data source from which KRISTEVA is collected. While larger models generally outperform their smaller variants, most tasks exhibit a more significant gap between Phi-4 and much larger models like Llama-3.1-70B. This difference suggests that data quality could be a more significant factor for interpretive reasoning ability than model scale, which further supports our call to explore the scalability of ethical data collection from college classroom settings.

Despite outperforming on the three out of four tasks that require multi-hop reasoning (Q8, Q10, Q11), reasoning models like o1-preview do not exhibit any advantage in most tasks. This result is consistent with the findings of recent studies that chain-of-thought mainly improves mathematics and logic-related tasks (Sprague et al., 2024), while having a very limited impact on common-sense, knowledge, and soft reasoning tasks that are more relevant to the setting of KRISTEVA.

## 7 Related Work

### College and PhD-level LLM Evaluations

Since OpenAI’s popularization of the term, “PhD-level intelligence” has rapidly caught on in the public discourse of AI as a tangible signpost for artificial general intelligence (AGI). Building on earlier general LLM evaluations at the college (MMLU) and graduate-levels of reasoning (Rein et al., 2023; Sawada et al., 2023), subsequent efforts have introduced domain-specific assessments in mathematics (Liu et al., 2024; Tsoukalas et al., 2024), computer science (son), biology (Laurent et al., 2024), history (Hauser et al., 2024), and psychology (Zhang et al., 2024). However, very few of these single or

multidisciplinary benchmarks includes literature as a test subject<sup>6</sup>—a surprising omission given OpenAI’s own results, which indicate that ChatGPT, GPT-4, and o1 all significantly underperform on AP English tests compared to other AP exams.<sup>7</sup> The causes of this discrepancy have not yet been explored. We introduce KRISTEVA to begin to address this gap.

### Multi-hop Reading Comprehension (MRC)

Since CosmosQA (Huang et al., 2019), there has been a growing interest in the evaluation of deeper reading comprehension capabilities that require reasoning components to extend beyond the literal understanding of the text (Dua et al., 2019; Sun et al., 2019; Yu et al., 2020). Evaluating such capabilities departs from earlier MRC benchmarking efforts that do not require reasoning (Rajpurkar et al., 2016; Chen et al., 2016), or where the involvement of reasoning might even lead to a drop in performance (Jia and Liang, 2017). A hallmark of multi-hop MRC is its reliance on external information, either explicitly provided across multiple documents (Welbl et al., 2018) or implicitly elicited via common sense (Huang et al., 2019), to fully understand the passage at hand. Incorporating reasoning-based MRC into domain-specific continued pre-training has been shown to enhance performance both within specialized domains and on general benchmarks (Cheng et al., 2023).

Recent benchmarks further align MRC with more complex reasoning tasks, such as natural language inference (NLI) (Liu et al., 2023), deep text understanding (Yao et al., 2023), critical reasoning (Kawabata and Sugawara, 2023), and extractive question answering (Basmov et al., 2024). In keeping with this research direction, we formulate context-dependent close reading as a uniquely challenging form of multi-hop MRC: to successfully reason between literary form and content (Q9), models must first correctly extract from the passage both components of the logical connection: stylistic features and external context. To the best

<sup>6</sup>Examples include “Humanity’s Last Exam” (Phan et al., 2025), which contains eight questions on English literature and 15 on poetry, and MMMU. MMMU (Yue et al., 2024) also technically has a “literature” category, but most questions listed in that category are purely informational, concerning book covers and illustrations rather than the literary text itself (Figure 1). Some other benchmarks also concern literature, but exclusively in the Chinese language (Li et al., 2024; Cao et al., 2024).

<sup>7</sup><https://openai.com/index/learning-to-reason-with-llms/>

of our knowledge, ours is the first MRC benchmark to be based on the challenging domain of literary texts and to require reasoning on figurative elements. Additionally, our dataset is sourced from college-level long-range documents (essays), which offer higher volumes and text quality compared to the standardized testing venues of existing benchmarks, like Chinese ESL (Sun et al., 2019) and LSAT (Yu et al., 2020).

### Figurative Language Understanding (FLU)

Initial benchmarks have evaluated FLU through QA (Rakshit and Flanigan, 2022) and NLI (Stowe et al., 2022). Moving beyond simpler tasks like metaphor detection, more recent studies have approached FLU as a form of reasoning. However, the scope of their reasoning tasks remains limited to rationales (Chakrabarty et al., 2022b), i.e., why something is a metaphor; explanations (Liu et al., 2022; Comşa et al., 2022), i.e., what the metaphor means and a breakdown of its implications; or literal rewordings (Tong et al., 2024). Few frame FLU as a pragmatics capability (Sravanthi et al., 2024), and none require models to interpret figurative language’s broader significance (as when our Q4 and Q6 investigate whether a given metaphor is needed, what its representational affordances might be, and how the passage would be different without it), or judge its relative effectiveness (as when our Q5 inquires whether one metaphor might be considered as more successful than another).

In addition, most existing benchmarks address FLU at the sentence-level, with far less focus on figurative language in context (Chakrabarty et al., 2022a). These cleanly parsed datasets do not align with the real-world complexity of figurative language as cognitive processes—metaphors, for instance, rarely exist in isolation, but are embedded in larger bodies of surrounding texts that often themselves function as part of broader networks of figural causations (Auerbach, 1953; White, 1999). Consequently, KRISTEVA introduces more complex reasoning tasks with purpose, effect, and context that are necessary for understanding how figurative language contributes to a passage’s overall meaning or the literary work’s narrative flow. As far as we know, KRISTEVA is the first benchmark to explicitly formulate FLU as a multi-hop reasoning task situated in the framework of multi-hop MRC.

**Literary NLP** The benchmark gap for close reading, the gold standard of evidentiary claims in literary studies, limits the advancement of NLP

research in the literary domain, where the performance of general-domain NLP models tend to “drop precipitously” (Bamman et al., 2019). Despite the utilization of literary corpora for tasks such as event extraction (Sims et al., 2019), information retrieval (Thai et al., 2022), and pretraining data detection (Chang et al., 2023), the literary domain remains relatively overlooked within the broader NLP community. While part of this challenge is inherent in the semantic ambiguity and pragmatic ineffability of literature, the development of literary NLP is more directly constrained by the bottleneck of standardized benchmark tasks, expert-annotated datasets, and generalizable evaluation metrics.

## 8 Conclusion and Future Work

We present KRISTEVA, the first close reading benchmark that evaluates the interpretive reasoning abilities of LLMs, featuring a novel task structure and a competitive human baseline. On comprehensive experiments with 19 models, we show that close reasoning presents several challenging tasks, and that LLMs still lag behind human performance.

Beyond the literary domain, the interpretive reasoning evaluated by KRISTEVA could be applicable to a range of NLP tasks. The close reading skills that KRISTEVA quantifies can serve as a proxy for evaluating long-range document understanding and help enhance the hermeneutic capabilities of LLMs. Additionally, close reading hones the reasoning ability to recognize narrative fidelity and coherence, which often serve as pragmatics protecting against harmful confabulation in human-to-human interaction; likewise, close reading abilities could potentially provide cultural guardrails against LLM confabulation (Sui et al., 2024). The ability to reason across latent stylistic and contextual spaces, addressed by Q10-Q11 in particular, also underpins the potential political affordances of close reading, for instance in the cases of disinformation recognition and media literacy. Moreover, KRISTEVA’s emphasis on evidence-based interpretive judgment could enrich style transfer and other human-centered creative NLP domains by facilitating more robust and justified preference judgments. We plan to continue updating KRISTEVA to better address these evolving research needs.

As the first benchmark of its type, KRISTEVA follows the NLP community’s common use of MCQ format. Our subsequent work seeks to

broaden the benchmark to open-ended evaluations of LLM free-text responses when directly prompted to perform close reading.

## Limitations

### 8.1 Data

Our current data source and collection process has several limitations. First, the scope of the study is limited to the classroom data of one course. Second, the course is at the introductory level, which limited the quality of the close readings gathered; its exams also do not implement the full CRIT (foregoing the “argument” step, which would have extended the length of the exam beyond the time available). Third, although two out of the three primary texts on the exams are translated from other languages, this work is performed entirely on texts in English.

### 8.2 Human Evaluation

The human evaluators who produced our current human baseline may have been disadvantaged by KRISTEVA’s MCQ format, as discussed in Section 6. This effect possibly reflects the fact that its MCQs contain LLM-generated distractors, which could bias the benchmark towards being more LLM-solvable. Both factors may be emblematic of the more general problem of complex and partly subjective reasoning tasks being adapted to typical approaches to AI evaluation (Crawford, 2021). To address these disadvantages, in our future work we plan to explore both human-annotated distractors and open-ended evaluations of close readings.

#### 8.2.1 “Expertise”

The term “domain expertise” is used in this paper to refer to literary critical competency at the level of a graduate student at a major research institution. Recent high-profile research has used a similar standard; for instance, the Ithaca tool for reconstructing damaged inscriptional texts was benchmarked against human annotations by “graduate students of ancient history, with 7 years of historical and linguistic training and specializing in Greek history and epigraphic documents” (Assael et al., 2022). As acknowledged by the Ithaca researchers, even experts at this level are “not yet equivalent to (the very small number) of established specialists in the field.” A further challenge for close reading of literary texts is the frequent need to draw on expertise beyond a single humanistic domain. Our results already suggest, for instance, potential differences in evaluation connected to disciplinary



background. In addition, annotators had insufficient time to read through the full course materials, hence they focused their preparation primarily on the works from which the relevant close-reading passages were drawn. Future research, therefore, might fruitfully explore the impact of greater opportunities for annotators to prepare, different forms of preparation, and a larger number of annotators from varied disciplinary backgrounds.

## Ethics Statement

The study protocol was submitted to the Institutional Review Board (IRB) at the UT Austin (ID: STUDY00006946). The IRB determined that this protocol meets the criteria for exemption from IRB review under 45 CFR 46.104 (1) Educational settings. The protocol required that the course data was de-identified by the course instructors before being shared with other members of the research team. Students consenting to participate in the study received \$20 in cash. Graduate research assistants were paid \$30 per hour for their contributions to the human evaluation data.

## Acknowledgments

This work was supported by Good Systems, a research grand challenge at UT Austin. CRIT, which helped inspire the format for this research, was developed in the Department of English at UT Austin by Professors Phillip Barrish, Evan Carton, Coleman Hutchison, and Frank Whigham, and PhD students Sydney Bufkin, Jessica Goudeau, and Jennifer Sapio. CRIT is a product of a Course Transformation Grant generously funded by the Office of the Executive Vice President and Provost. CRIT is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

We thank Ziang Xiao for his valuable feedback on the experimental design of our benchmark. Finally, we thank our anonymous reviewers for their generous time and attention.

## References

Cs-Bench: A comprehensive benchmark for large language models towards computer science mastery, author=Song, Xiaoshuai and Diao, Muxi and Dong, Guanting and Wang, Zhengyang and Fu, Yujia and Qiao, Runqi and Wang, Zhexu and Fu, Dayuan and Wu, Huangxuan and Liang, Bin and others, journal=arXiv:2406.08587, year=2024.

M.H. Abrams. 2009. *A Glossary of Literary Terms*. Wadsworth Cengage Learning.

Yannis Assael, Thea Sommerschild, Brendan Shillingford, Mahyar Bordbar, John Pavlopoulos, Marita Chatzipanagiotou, Ion Androutsopoulos, Jonathan Prag, and Nando de Freitas. 2022. [Restoring and attributing ancient texts using deep neural networks](#). *Nature*, 603(7900):280–283.

Erich Auerbach. 1953. *Mimesis: The Representation of Reality in Western Literature*. Princeton University Press.

David Bamman, Sejal Papat, and Sheng Shen. 2019. [An annotated dataset of literary entities](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2138–2144, Minneapolis, Minnesota. Association for Computational Linguistics.

Annie Bares, Daniel F. Keefe, and Francesca Sam-sel. 2020. Close reading for visualization evaluation. *IEEE Computer Graphics and Applications*, 40(4):84–95.

Victoria Basmov, Yoav Goldberg, and Reut Tsarfaty. 2024. LLMs’ reading comprehension is affected by parametric knowledge and struggles with hypothetical statements. *arXiv preprint arXiv:2404.06283*.

Don Bialostosky. 2006. Should college English be close reading? *College English*, 69(2):111–116.

Jiahuan Cao, Yang Liu, Yongxin Shi, Kai Ding, and Lianwen Jin. 2024. Wenmind: A comprehensive benchmark for evaluating large language models in Chinese classical literature and language arts. *Advances in Neural Information Processing Systems*, 37:51358–51410.

Ellen C. Carillo. 2018. *Teaching Readers in Post-Truth America*. University Press of Colorado.

Tuhin Chakrabarty, Yejin Choi, and Vered Shwartz. 2022a. [It’s not rocket science: Interpreting figurative language in narratives](#). *Transactions of the Association for Computational Linguistics*, 10:589–606.

Tuhin Chakrabarty, Arkadiy Saakyan, Debanjan Ghosh, and Smaranda Muresan. 2022b. [FLUTE: Figurative language understanding through textual explanations](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7139–7159, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Kent Chang, Mackenzie Cramer, Sandeep Soni, and David Bamman. 2023. [Speak, memory: An archaeology of books known to ChatGPT/GPT-4](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7312–7327, Singapore. Association for Computational Linguistics.



- Rita Charon. 2006. *Narrative Medicine: Honoring the Stories of Illness*. Oxford University Press.
- Danqi Chen, Jason Bolton, and Christopher D. Manning. 2016. [A thorough examination of the CNN/Daily Mail reading comprehension task](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2358–2367, Berlin, Germany. Association for Computational Linguistics.
- Daixuan Cheng, Shaohan Huang, and Furu Wei. 2023. Adapting large language models via reading comprehension. In *The Twelfth International Conference on Learning Representations*.
- Samuel Taylor Coleridge. 1849. *Notes and Lectures Upon Shakespeare and Some of the Old Poets and Dramatists: With Other Literary Remains of ST Coleridge*, volume 13. W. Pickering.
- CommonCore. 2012. Revised publishers’ criteria for the common core state standards in English language arts and literacy, grades 3–12.
- Iulia Comşa, Julian Eisenschlos, and Srinu Narayanan. 2022. [MiQA: A benchmark for inference on metaphorical questions](#). In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 373–381, Online only. Association for Computational Linguistics.
- Kate Crawford. 2021. *The Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. Yale University Press.
- John Dewey. 1910. *How We Think*. DC Heath.
- Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. [DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2368–2378, Minneapolis, Minnesota. Association for Computational Linguistics.
- Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac’h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2024. [A framework for few-shot language model evaluation](#).
- John Guillory. 2025. *On Close Reading*. University of Chicago Press.
- Jakob Hauser, Dániel Kondor, Jenny Reddish, Majid Benam, Enrico Cioni, Federica Villa, James S. Bennett, Daniel Hoyer, Pieter Francois, Peter Turchin, et al. 2024. Large language models’ expert-level global history knowledge benchmark (HiST-LLM). In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
- N. Katherine Hayles. 2010. How we read: Close, hyper, machine. *ADE Bulletin*, 150(18):62–79.
- Jack Hessel, Ana Marasovic, Jena D. Hwang, Lillian Lee, Jeff Da, Rowan Zellers, Robert Mankoff, and Yejin Choi. 2023. [Do androids laugh at electric sheep? Humor “understanding” benchmarks from the New Yorker Caption Contest](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 688–714, Toronto, Canada. Association for Computational Linguistics.
- Renee Hobbs. 2010. *Digital and Media Literacy: A Plan of Action. A White Paper on the Digital and Media Literacy Recommendations of the Knight Commission on the Information Needs of Communities in a Democracy*. ERIC.
- Bell Hooks. 1994. *Teaching to Transgress: Education as the Practice of Freedom*. Routledge.
- Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. [Cosmos QA: Machine reading comprehension with contextual commonsense reasoning](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2391–2401, Hong Kong, China. Association for Computational Linguistics.
- Jessica Hullman, Ari Holtzman, and Andrew Gelman. 2023. Artificial intelligence and aesthetic judgment. *arXiv:2309.12338*.
- Robin Jia and Percy Liang. 2017. [Adversarial examples for evaluating reading comprehension systems](#). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031, Copenhagen, Denmark. Association for Computational Linguistics.
- Akira Kawabata and Saku Sugawara. 2023. [Evaluating the rationale understanding of critical reasoning in logical reading comprehension](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 116–143, Singapore. Association for Computational Linguistics.
- Jon M. Laurent, Joseph D. Janizek, Michael Ruzo, Michaela M. Hinks, Michael J. Hammerling, Sidharth Narayanan, Manvitha Ponnampati, Andrew D. White, and Samuel G. Rodrigues. 2024. Lab-bench: Measuring capabilities of language models for biology research. *arXiv:2407.10362*.

- Caroline Levine. 2015. *Forms: Whole, Rhythm, Hierarchy, Network*. Princeton University Press.
- Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. 2024. **CMMLU: Measuring massive multitask language understanding in Chinese**. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11260–11285, Bangkok, Thailand. Association for Computational Linguistics.
- Emmy Liu, Chenxuan Cui, Kenneth Zheng, and Graham Neubig. 2022. **Testing the ability of language models to interpret figurative language**. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4437–4452, Seattle, United States. Association for Computational Linguistics.
- Hanmeng Liu, Jian Liu, Leyang Cui, Zhiyang Teng, Nan Duan, Ming Zhou, and Yue Zhang. 2023. **Logiqa 2.0—an improved dataset for logical reasoning in natural language understanding**. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.
- Hongwei Liu, Zilong Zheng, Yuxuan Qiao, Haodong Duan, Zhiwei Fei, Fengzhe Zhou, Wenwei Zhang, Songyang Zhang, Dahua Lin, and Kai Chen. 2024. **MathBench: Evaluating the theory and application proficiency of LLMs with a hierarchical mathematics benchmark**. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6884–6915, Bangkok, Thailand. Association for Computational Linguistics.
- Sarah McGrew. 2020. Learning to evaluate: An intervention in civic online reasoning. *Computers & Education*, 145:103711.
- Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, Sean Shi, Michael Choi, Anish Agrawal, Arnav Chopra, et al. 2025. **Humanity’s last exam**. *arXiv:2501.14249*.
- Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. **SQuAD: 100,000+ questions for machine comprehension of text**. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.
- Geetanjali Rakshit and Jeffrey Flanigan. 2022. **FigurativeQA: A test benchmark for figurativeness comprehension for question answering**. In *Proceedings of the 3rd Workshop on Figurative Language Processing (FLP)*, pages 160–166, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
- David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2023. **GPQA: A graduate-level Google-proof Q&A benchmark**. *arXiv:2311.12022*.
- Ivor Armstrong Richards. 1929. *Practical Criticism*. Kegan Paul, Trench, Trubner, London.
- Ivor Armstrong Richards. 1936. *The Philosophy of Rhetoric*. Bryn Mawr College. Mary Flexner Lectures. Oxford University Press.
- Tomohiro Sawada, Daniel Paleka, Alexander Havrilla, Pranav Tadepalli, Paula Vidas, Alexander Kranias, John J. Nay, Kshitij Gupta, and Aran Komatsuzaki. 2023. **ARB: Advanced reasoning benchmark for large language models**. *arXiv:2307.13692*.
- Matthew Sims, Jong Ho Park, and David Bamman. 2019. **Literary event detection**. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3623–3634, Florence, Italy. Association for Computational Linguistics.
- Dan Sinykin and Johanna Winant. 2025. *Close Reading for the Twenty-First Century*. Princeton University Press.
- Zayne Sprague, Fangcong Yin, Juan Diego Rodriguez, Dongwei Jiang, Manya Wadhwa, Prasann Singhal, Xinyu Zhao, Xi Ye, Kyle Mahowald, and Greg Durrett. 2024. **To CoT or not to CoT? Chain-of-thought helps mainly on math and symbolic reasoning**. *CoRR*, abs/2409.12183.
- Settaluri Sravanthi, Meet Doshi, Pavan Tankala, Rudra Murthy, Raj Dabre, and Pushpak Bhattacharyya. 2024. **PUB: A pragmatics understanding benchmark for assessing LLMs’ pragmatics capabilities**. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12075–12097, Bangkok, Thailand. Association for Computational Linguistics.
- Katherine Stasaski and Marti A Hearst. 2017. Multiple choice question generation utilizing an ontology. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 303–312.
- Kevin Stowe, Prasetya Utama, and Iryna Gurevych. 2022. **IMPLI: Investigating NLI models’ performance on figurative language**. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5375–5388, Dublin, Ireland. Association for Computational Linguistics.
- Peiqi Sui, Eamon Duede, Sophie Wu, and Richard So. 2024. **Confabulation: The surprising value of large language model hallucinations**. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14274–14284, Bangkok, Thailand. Association for Computational Linguistics.
- Kai Sun, Dian Yu, Jianshu Chen, Dong Yu, Yejin Choi, and Claire Cardie. 2019. **DREAM: A challenge data set and models for dialogue-based reading comprehension**. *Transactions of the Association for Computational Linguistics*, 7:217–231.

- Katherine Thai, Yapei Chang, Kalpesh Krishna, and Mohit Iyyer. 2022. **RELiC: Retrieving evidence for literary claims**. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7500–7518, Dublin, Ireland. Association for Computational Linguistics.
- Xiaoyu Tong, Rochelle Choenni, Martha Lewis, and Ekaterina Shutova. 2024. **Metaphor understanding challenge dataset for LLMs**. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3517–3536, Bangkok, Thailand. Association for Computational Linguistics.
- George Tsoukalas, Jasper Lee, John Jennings, Jimmy Xin, Michelle Ding, Michael Jennings, Amitayush Thakur, and Swarat Chaudhuri. 2024. Putnambench: Evaluating neural theorem-provers on the putnam mathematical competition. *arXiv:2407.11214*.
- Ted Underwood. 2019. *Distant Horizons: Digital Evidence and Literary Change*. University of Chicago Press.
- Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. 2018. **Constructing datasets for multi-hop reading comprehension across documents**. *Transactions of the Association for Computational Linguistics*, 6:287–302.
- Hayden White. 1999. *Figural Realism: Studies in the Mimesis Effect*. Johns Hopkins University Press.
- Zijun Yao, Yantao Liu, Xin Lv, Shulin Cao, Jifan Yu, Juanzi Li, and Lei Hou. 2023. **KoRC: Knowledge oriented reading comprehension benchmark for deep text understanding**. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11689–11707, Toronto, Canada. Association for Computational Linguistics.
- Weihao Yu, Zihang Jiang, Yanfei Dong, and Jiashi Feng. 2020. Reclor: A reading comprehension dataset requiring logical reasoning. *arXiv:2002.04326*.
- Ruibin Yuan, Hanfeng Lin, Yi Wang, Zeyue Tian, Shangda Wu, Tianhao Shen, Ge Zhang, Yuhang Wu, Cong Liu, Ziya Zhou, Liumeng Xue, Ziyang Ma, Qin Liu, Tianyu Zheng, Yizhi Li, Yinghao Ma, Yiming Liang, Xiaowei Chi, Ruibo Liu, Zili Wang, Chenghua Lin, Qifeng Liu, Tao Jiang, Wenhao Huang, Wenhui Chen, Jie Fu, Emmanouil Benetos, Gus Xia, Roger Dannenberg, Wei Xue, Shiyin Kang, and Yike Guo. 2024. **ChatMusician: Understanding and generating music intrinsically with LLM**. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6252–6271, Bangkok, Thailand. Association for Computational Linguistics.
- Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. 2024. **MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI**. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567.
- Junlei Zhang, Hongliang He, Nirui Song, Zhanchao Zhou, Shuyuan He, Shuai Zhang, Huachuan Qiu, Anqi Li, Yong Dai, Lizhi Ma, and Zhenzhong Lan. 2024. **ConceptPsy: A benchmark suite with conceptual comprehensiveness in psychology**. *arXiv:2311.09861*.

## A Humanistic Mission Statement

Automating close reading is not an end goal of this work: for many reasons, including the basic fact that close reading is a practice whose significance largely derives from its status as a method of personal deliberation. However, developing computational models for close reading could help clarify and facilitate what we humans do when we engage in this practice, describe it, and teach it. For example, machine learning could enable new kinds of systematic comparisons among critical and literary-theoretical approaches.

Such comparisons could inform practitioners of literary criticism about a range of issues relating to their craft: about what kinds of task are prone to error, where human creativity excels, how disciplinary conventions shape analysis, and which skills warrant particular attention in our pedagogy. At the same time, the field of literary studies, with its distinctively subjective and associative forms of reasoning, can provide resources for the development of the next generation of language models, and a crucial test for them. Here as elsewhere, the craft of traditional disciplines has much to offer computational research (Underwood, 2019).

In the course of connecting these fields, we try to avoid attributing “understanding,” “judgment,” or other forms of critical consciousness to LLMs. When such attributions do happen, we consider them to make the same kind of sense as commonly seen attributions of human qualities to traditional human-created works, like books, or latent author figures composed from our experience of such works: as when we say that Coleridge’s *Lectures on Shakespeare* change our understanding of the Bard, “himself a nature humanized, a genial understanding directing self-consciously a power and an implicit wisdom deeper even than our consciousness” (Coleridge, 1849).

## B KRISTEVA Tasks Details

Please see Table 3 for the full input/output schema of KRISTEVA’s task structure.

Question	Input / Output	Full Question
Q1 Feature Type	Input: {location} Output: {feature_type}	What rhetorical device is present in {location}?
Q2 Feature Location	Input: {feature_type} Output: {location}	
Q3 Feature Elements	Input: {feature_type}, {location} Output: {feature_elements}	In the {feature_type} that occurs in {location}, what are the specific elements of the device?
Q4 Feature Purpose	Input: {feature_type}, {location}, {feature_elements} Output: {purpose}	In the {feature_type} that occurs in {location}, {feature_elements}. What is the purpose of this device?
Q5 Feature Relative Importance	Input: {features} (answers of Q1-4) Output: {significant_feature}	Which of the following stylistic features is the most significant to the passage?
Q6 Feature Significance	Input: {significant_feature} Output: {feature_significance}	In the {feature_type} that occurs in {location}, {feature_elements}. Which of the following best describes the significance of this device, and what are its effects on the reader?
Q7 Relevant Context Retrieval	Input: {passage} Output: {context_type}, {context_elements}	Which external context is the most relevant to the following passage?
Q8 Context Relative Importance	Input: {contexts} (answers of Q7) Output: {significant_context}, {context_elements}	Which of the following contextual information is the most significant to the passage?
Q9 Context Significance	Input: {significant_context} Output: {context_significance}	In the {context_type} that occurs in {location}, {context_elements}. Which of the following best describes the significance of this device, and what are its effects on the reader?
Q10 Feature-Context Matching	Input: {context_type}, {context_elements}, {features} Output: {corresponding_feature}	Please identify the stylistic feature that the following {context_type} information best helps to contextualize: {context_elements}.
Q11 Feature-Context Reasoning	Input: {context_type}, {corresponding_feature}, {selected_passage} Output: Rationale for the {feature_context_pair}	What is the most reasonable connection between the {context_type} context and the use of {feature_type} feature in the following passage?

Table 3: KRISTEVA task structure and question formats, adapted from UT Austin’s CRIT framework.

## B.1 CRIT Step: Observe

In this step of the CRIT framework that helped formalize the close reading pedagogy this study engages, students are asked to perform literary pattern recognition (So and Long, 2016): the observation of how the passage stands out as a literary one through its deviation from normal everyday language use, identified as features of form and style that help a passage accomplish its literary aims. These stylistic features, including figurative language, sonic patterns, poetic form, diction and syntax, narrative devices, etc., are widely considered to be characteristic of literary texts and distinguish them as a unique domain. Operationally, the “observe” step could be broken down into two components: 1) listing and explicating potentially significant stylistic features in the selected passage, and 2) inferring the purpose they serve in the passage and evaluate their rhetorical success in terms of what unique effect they have achieved (i.e., hy-

pothesizing how the text would read differently if a given feature were removed).

**Q1 (Feature Type), Q2 (Feature Location), Q3 (Feature Elements)** Mostly following the existing framework of FLU, these three tasks collectively evaluate LLMs’ ability to accurately detect, locate, and explain figurative language and other stylistic features embedded in a given passage of literary text. Q1 asks models to detect the type of features present in a given passage (e.g., metaphor, alliteration, or symbolism). Q2 is the reversal of Q1, where models are given a feature type and asked to identify the part of the passage where it occurs. Q3 builds on the answers of Q1 and Q2 to require a higher level of stylistic feature understanding, prompting models to describe the specific elements of a given feature and location.

For student essays that do not provide a specific location through direct quotes, Q1 and Q2 are collapsed into one question: “Which of the following



is the most prominent use of rhetorical device in the passage?”

Q1 and Q3 are standard figurative language identification tasks that structurally align with extant forms of evaluation in FLU. Q2, however, represents a departure by testing for open stylistic feature extraction from unstructured text: existing work in FLU tends to represent figurative language with “subject-relation-object” triples (Liu et al., 2022) in clean, information-extraction style datasets. In contrast, complex rhetorical devices in literary texts (e.g, a Homeric simile) are often not neatly separable from the surrounding language. This scenario entails a more challenging form of FLU beyond only identifying and explaining the component of a device, requiring models to also parse out the language around it that aids in its construction.

**Q4 (Feature Purpose), Q5 (Feature Significance Ranking), and Q6 (Feature Significance Inference)** Given a fully extracted and localized stylistic feature (outputs of Q2–Q4), Q5 asks the model to interpret why a particular device is used and how it influences the text. This question concerns the functional role of a given feature within a text (e.g., is it creating suspense, emphasizing an emotional tone, making a certain imagery more concrete, etc.) and its rhetorical success (i.e., what effect it could have on the reader’s interpretation or emotional response). Q5 & Q6 builds on this analysis of effect to evaluate the model’s ability to weigh the relative importance of features. Specifically, it asks the model to compare the influence of multiple stylistic features (all outputs of Q1–Q4) within the selected passage, and rank them in descending order of significance by assessing their effects on the reader (how well they direct reader engagement and manage their attention and expectations). Together, Q4, Q5, and Q6 push the reasoning components of FLU to interpretations regarding figurative language’s broader significance – *what they are for*, and *why certain features are especially impactful* for the text.

## B.2 CRIT Step: Contextualize

This step expands the interpretive scope of close reading by introducing elements external to the literary passage: pertinent configurations of facts and boarder circumstances, be it literary, biographical, cultural, or historical, that could serve as contextual frames for the passage and contribute to its

overall meaning. Students are asked to draw on either their own world knowledge or course materials to extract a list of relevant contextual information, such as cultural artefacts, intellectual history, authorial influences, or wider societal developments, and consider how these external factors could be potentially significant for the passage.

**Q7 (Relevant Context Retrieval)** To evaluate the “contextualize” step, Q7 requires models to identify the most relevant piece of contextual information to a literary passage, and classify its type (literary, biographical, cultural, or historical). It tests if the model could effectively leverage its parametric knowledge to perform zero-shot world knowledge retrieval, and understand the relative significance of what it retrieved.

## C Prompts

### C.1 Benchmark Construction Prompts

The following prompts were used to extract structured representations from CRIT essays and generate distractors:

You are given the response to an exam that had four questions (Question 1-4) that analyze the following passage:

[passage]

Your goal is to extract the individual answers to each of the questions.

**Rules:**

- If there are errors due to text extraction, such as excessive new lines, you can fix that.
- You can fix clear and unambiguous typos.
- Your response should be a JSON object, with the following schema:

"q1\_observe": [{"id": 1, "type": "...", "location": "...", "elements": "...", "purpose": "..."}, ...]  
(list of dictionaries, each with the following keys: id, type, location, elements, purpose)

- The id corresponds to a unique number for each rhetorical device (1, 2, 3, etc.) so it can be identified in the response.
- The type corresponds to a type of rhetorical device, such as: Allusion, Metaphor, Personification, etc.
- The location corresponds to an exact excerpt from the selected passage that corresponds to the rhetorical device. Include only what the student presents in direct quotations from the literary passage. If the student does not present any direct quotations, leave this key blank.
- The elements corresponds to the rhetorical device's components. In one sentence, simply outline the elements of what is being described as what (if the device is a metaphor, the elements are the tenor and the vehicle, etc.), or the specific language that carries the device (if the device is an alliteration, the elements are the letters that are alliterated, etc.). Some students might also offer an explanation for how the device's components differ from their literal meaning. If so, include the explanation as a second sentence.
- The purpose is the purpose and effect of the rhetorical device, extracted from the response. For instance, the purpose of "Word choice: words such as 'gnawing', 'rage' and 'crying' create an eerie and dark tone that fits throughout the passage" is "to create an eerie and dark tone that fits throughout the passage". Please summarize the purpose and effect in one sentence if the student's description is too long. Some responses could be excessively short and not contain a purpose, e.g., "epithets ('gem of Raghuis', 'best of Raghuis')", in which case, leave the purpose blank.
- You must only extract rhetorical devices listed in the response and not invent ones that are not present. You should only extract elements from the response for the elements/purpose, and not infer or create new ones.

"q2\_context": [{"id": 1, "type": "...", "elements": "..."}, ...]  
(list of dictionaries, each with the following keys: id, type, elements)

- The id corresponds to a unique number for each context element (1, 2, 3, etc.) so it can be identified in the response.
- The type corresponds to a type of contextual element, such as: Historical, Cultural, Biographical, etc.
- The elements are a high-level description of the contextual element that is relevant to the selected passage, extracted from the response of the student. Focus on the factual information about the context external to the literary work. Please summarize the content in one sentence if the student's response is too long.

"q3\_analyze\_i": [{"id": 1, "type": "...", "corresponding\_id": 1, "significance": "..."}, ...]  
(list of dictionaries, each with the following keys: id, type, corresponding\_id, significance)

- The id corresponds to a unique number for each significance (1, 2, 3, etc.) so it can be identified in the response.
- The type describes if the significance is about a rhetorical device or a contextual element. Return "feature" for rhetorical devices and "context" for contextual elements.
- The corresponding\_id is the id of the rhetorical device or contextual element that the significance corresponds to, drawn from the id of either "q1\_observe" or "q2\_context".
- The significance is the significance of a rhetorical device or contextual element, extracted from the response. Please summarize the significance in one sentence if the student's description is too long.

"q4\_analyze\_ii": [{"feature\_context\_pair": "...", "feature\_id": 1, "context\_id": 1, "feature\_context\_conn": "..."}, ...]  
(list of dictionaries, each with the following keys: feature\_context\_pair, feature\_id, context\_id, feature\_context\_conn)

- The feature\_context\_pair corresponds to the rhetorical device and the context connected together. Each pair should be listed as a string with the following structure: "{feature\_type}, {feature\_location}; {context\_type}, {context\_elements}".
- The feature\_id corresponds to the id of the pair's rhetorical device identified in "q1\_observe". If multiple rhetorical devices are described in the response, only select the first one. If no rhetorical device is described, make it -1.
- The context\_id corresponds to the id of the pair's context identified in "q2\_context". If multiple rhetorical devices are described in the response, only select the first one. If no rhetorical device is described, make it -1.
- The feature\_context\_conn is a description in the response of how the rhetorical device is connected to its corresponding contextual element and what makes this connection significant.

**Overall, the JSON response you produce should therefore adhere to the following schema:**

{"q1\_observe": [...], "q2\_context": [...], "q3\_analyze\_i": [...], "q4\_analyze\_ii": [...]}

**Now proceed with the extraction for the following response:**

You are given a JSON array of question objects, each containing the following fields:

- "question\_number" (string; one of: "Q1", "Q2", "Q4", "Q5", "Q7", "Q9", or "Q11")
- "answer" (the correct answer or interpretation for the question)
- "selected\_passage" (a snippet of the full poem, if relevant)
- "full\_passage" (the entire poem or the relevant portion of it, if needed)
- "location" (location reference in the poem, if relevant)
- "r\_type" (the rhetorical device type, if relevant)
- "elements" (the interpretation or content of the rhetorical device, if relevant)
- "purpose" (the purpose/effect of the rhetorical device, if relevant)
- "ctype" (the type of external context: historical, cultural, biographical, or literary, if relevant)
- "celements" (the factual external context details, if relevant)
- "corresponding\_feature" (the rhetorical device to which the context is connected, if relevant)
- "pair" (a short string describing the rhetorical device + external context pairing, if relevant)

**Your task:**

1. For each question object in the input JSON, generate exactly **three (3) distractors**.
2. Append those three distractors plus the correct answer (in the fourth position) to a new array "choices" within that same question object.
3. **Do not** add any additional commentary or fields; only add "choices" to each question object, containing [distractor1, distractor2, distractor3, correctAnswer].

**The way you generate the three distractors depends on question\_number:**

**Q1**

Here is a snippet of a poem: {selected\_passage}, selected for literary analysis from the full poem: {full\_passage}. One interpretation of how this selected passage stands out from the full work is: {answer}. I want three one-sentence distractors in the context of how this passage might differ from or connect to the rest of the poem. Number them 1, 2, 3, and do not provide other commentary.

**Q2**

Here is a snippet of a poem: {selected\_passage}. In {location}, there is a {r\_type}. I want three rhetorical devices as distractors that might also appear in {location}. Number them 1, 2, 3, and do not provide other commentary.

**Q4**

Here is a snippet of a poem: {selected\_passage}. In {location}, there's a {r\_type}. One interpretation is: {elements}. I want three one-sentence distractors for this interpretation. Number them 1, 2, 3, and do not provide other commentary.

**Q5**

Here is a snippet of a poem: {selected\_passage}. In the {r\_type} that occurs in {location}, {elements}. One interpretation of this device's purpose/effect is: {purpose}. I want three one-sentence distractors for that interpretation. Number them 1, 2, 3, and do not provide other commentary.

**Q7**

Here is a snippet of a poem: {selected\_passage}. We know it is relevant to this external context: "{ctype}, {celements}." I want three distractors for other possible contexts that might be relevant. Structure each distractor similarly as 'context\_type, context\_content.' Number them 1, 2, 3, and do not provide other commentary.

**Q9**

Here is a snippet of a poem: {selected\_passage}. It uses {corresponding\_feature}, and is relevant to this {ctype} context: {celements}. One interpretation for how the rhetorical device connects with that context is: {answer}. I want three one-sentence distractors for that interpretation. Number them 1, 2, 3, and do not provide other commentary.

**Q11**

Here is the full poem: {full\_passage}. It contains a connection between a rhetorical device and an external context: {pair}. One interpretive argument for that connection is: {answer}. I want three distractors for that argument. Number them 1, 2, 3, and do not provide other commentary.

**Any other question\_number**

Here is a snippet of a poem: {selected\_passage}. We have a question, and the correct answer is {answer}. I want three one-sentence distractors relevant to this question. Number them 1, 2, 3, and do not provide other commentary.

**Final Output:**

After generating these three distractors for each question, insert them plus the correct answer into a new array field "choices" (with the correct answer as the last item) for each question object.

Finally, output the entire updated JSON array of questions, where each question has the form:

```
{
  "question_number": "...",
  "answer": "...",
  ...
  "choices": ["distractor1", "distractor2", "distractor3", "correct answer"]
}
```

**Now proceed with distractor generation for the following JSON array of questions:**

[[QUESTIONS\_JSON]]

## C.2 Evaluation Prompts

The following prompt is used for the evaluation of LLMs, reported in Table 2:

Consider the following literary passage:

[[PASSAGE]]

Question:

[[QUESTION]]

Select the correct answer from the following choices:

[[CHOICES]]

**Response Format:**

You must answer the question using JSON format, with the following schema:

```
{"answer": "A" | "B" | "C" | "D"}
```

You should not include any other text in your response.

## D Details on Data Source

### D.1 Exams

Please see [Figure 4](#).

### D.2 Grading Rubrics

Please see [Figure 5](#).

## E Details on Human Evaluation

Please see [Figure 6](#) for the interface we used to perform the human evaluation.



## Unit Test 2

Keep Editing This Quiz

This is a preview of the published version of the quiz

Started: Feb 16 at 12:28am

### Quiz Instructions

#### Questions

- Question 4
- Question 5
- Question 6
- Question 7
- Question 8
- Question 9
- Question 10
- Question 11

Time Running:

Attempt due: Nov 7, 2024 at 3:40pm  
0 Minutes, 0 Seconds



#### Question 11

25 pts

##### Basics

Provide the work's author, title, genre, original language, approximate date of creation, and place of origin. (3)

##### CRIT

###### Step 1-Paraphrase

Read the passage below carefully. In your own words, give a summary of the factual content of the passage—what the text directly states—as it proceeds from beginning to end. What situation is being described here and by whom? What happens in that situation? Respond to this prompt in no more than three complete sentences. (2)

###### Step 2-Observe

Read the passage again, this time thinking about what it seeks to accomplish. Then, identify and list any potentially significant features of the passage's language or form—that is, those textual elements that contribute to the passage's overall meaning, purpose, or effect. Your list of observations should include specific examples of various kinds of textual elements, such as: descriptive details; word choice; repetition of phrases, sounds, or ideas; imagery or figurative language; syntactical structure; changes in vocabulary, rhythm, or tone; characteristics of the narrative voice or perspective. Respond to this prompt with a list of features. \*Remember that your passage is a translation, so some of the features listed above may not be reflected in English (and, conversely, some of the features of the translation may not belong to the original). Focus only on those features which you feel the English translation is likely to reflect with reasonable accuracy.\* (5)

###### Step 3-Contextualize

Think about contexts for the passage. (Contexts are facts or broader circumstances external to a literary work that are important to its production, reception, or understanding; for instance: literary, biographical, political, or historical information.) From your own knowledge of any relevant contextual facts or circumstances, or from information provided by your instructor, identify and list any potentially significant contexts for the passage—that is, those contextual frames that contribute to the passage's overall meaning, purpose, or effect. Respond to this prompt with a list of contexts. (5)

Figure 4: Exam interface (using unit test 2 as an example).

### **Rubric for grading**

Rubric for Basics component of essay (/3). ½ pt for each item (total of 6, hence 3 marks total).

Rubric for CRIT (/22):

Paraphrase (/2):

Accurate summary of content of passage and situation – 2 marks

Partial summary or Detailed summary containing errors – 1 mark

Vague or erroneous summary – 0 marks

Observe (/5):

2 marks for first correct formal feature, 1 mark for each feature thereafter up to a max of 5 (i.e., 4 features total)

Contextualize (/5):

2 marks for first correct context, 1 mark for each context thereafter up to a max of 5 (i.e., 4 contexts total)

Analyze (/5):

2 marks for first plausible claim about significance, 1 mark for each plausible claim thereafter up to a maximum of 5 (i.e., 4 claims total)

Alternatively, 5 marks can also be achieved through substantial developed answers dealing with one or two larger themes using multiple examples for each theme: 2 themes with a total of 4 good examples, or (in rare cases) 1 theme with a total of 4 good examples.

Analyze II (/5):

2 marks for first plausible claim about *paired* significance, 1 mark for each plausible claim thereafter up to a maximum of 5 (i.e., 4 claims total)

Alternatively, 5 marks can also be achieved through substantial developed answers dealing with one or two larger themes using multiple *paired* examples for each theme: 2 themes with a total of 4 good examples, or (in rare cases) 1 theme with a total of 4 good examples.

Figure 5: Rubric used to grade CRIT essays.

**Progress**  
Exam ID: 1  
Annotator: peiq\_sui

---

Completed: 0  
Remaining: 66  
Total: 66

## MCQ Annotation Interface

### Passage

"Oh, Creon! Creon! This is not the first time that my reputation has hurt me enormously. The wise man must never bring up his children to be too wise. Because, not only will the crowds call them "useless" but that same crowd will also turn their back on them and treat them with envy and hostility. If you were to teach the uneducated something new and beautiful, not only will they not appreciate it but they'll call you "good for nothing." And if those who think themselves to be wise, consider you even wiser than them, then you'll hurt them most awfully. And that's where my own suffering stems from: They call me "wise" yet, in reality I am not very "wise" at all. Some people scorn me, others simply hate me and you are now afraid that I might do you some enormous harm!

Have no fear, though, Creon! I'm not capable of hurting kings! In any case, how did you ever hurt me? You simply married your daughter off to the man you liked. I hate my husband whereas what you did, you did with your mind intact. So, on my part now, I do not at all envy your good luck. By all means, have the wedding and may you rejoice by them. But let me stay here, in this land and, even though I've been dishonoured, I'll say no more of it because you are by far my superior!"

### Question

What rhetorical device most accurately describes the stylistic feature present in "wise" and "wisdom"?

### Choices

- Select your answer:
- A. Paronomasia
  - B. Alliteration
  - C. Repetition
  - D. Polypoton

Submit

Figure 6: The online interface used for collecting the performance of evaluators on a subset of KRISTEVA to create the human baseline.