Argumentative essay assessment with LLMs: A critical scoping review

Lucile Favero[1*], Gabrielle Gaudeau[2], Juan Antonio Pérez-Ortiz[3], Tanja Käser[4], Nuria Oliver[1]

[1*]ELLIS Alicante, Muelle de Poniente 5, Distrito Digital 5, Edificio A, Puerto de Alicante, Alicante, 03001, Spain.
[2]ALTA Institute, Computer Laboratory, University of Cambridge, Cambridge, CB3 0FD, United-Kingdom.
[3]Universitat d'Alacant, Carretera San Vicente del Raspeig s/n, San Vicente del Raspeig - Alicante, 03690, Spain.
[4]École Polytechnique Fédérale de Lausanne (EPFL), Station 14, Lausanne, 1015, Switzerland.

*Corresponding author(s). E-mail(s): lucile@ellisalicante.org;
Contributing authors: gjg34@cam.ac.uk; japerez@dlsi.ua.es ;
tanja.kaeser@epfl.ch; nuria@ellisalicante.org ;

Abstract

Large Language Models are rapidly reshaping Automated Essay Scoring (AES), yet the methodological, conceptual, and ethical foundations of Argumentative Automated Essay Scoring (AAES) remain underdeveloped. This critical review synthesizes 46 studies published between 2022–2025, following PRISMA 2020 guidelines and a preregistered protocol. We map the landscape of LLM-based AAES across six dimensions—datasets, traits, models, methods, evaluation, and analytics. Our findings show that AAES research remains fragmented and insufficiently grounded in argumentation theory. The field relies on non-comparable datasets which vary in availability, prompt diversity, rater configuration, and linguistic background. Trait analysis reveals substantial overrepresentation of rhetorical and linguistic features and sparse coverage of reasoning-oriented constructs (*e.g.*, logical cogency, dialectical quality). Studies mainly rely on proprietary GPT-family models and rubric-based prompting, while only a minority employ fine-tuning, multi-agent approaches, or reasoning LLMs. Evaluation practices remain uneven: although studies report high human-model agreement, robustness analyses expose sensitivity to prompting, score distributions, and

learner proficiency. FATEN analyses reveal recurrent concerns regarding fairness (*e.g.*, style and L1 bias), transparency, randomness sensitivity, limited pedagogical alignment, and an absence of work on privacy or deployment safety. Taken together, the evidence suggests that while LLMs can approximate human scoring on several traits, current systems insufficiently model core argumentative reasoning and lack the validity, interpretability, and accountability required for high-stakes assessment. We conclude by proposing a research agenda focused on construct-valid datasets and rubrics, psychometric modeling, transparent evaluation protocols, and responsible design frameworks.

Keywords: Argumentative Automated Essay Scoring, Educational Assessment, Large Language Models, Multi-trait Assessment, Fairness, Scoring Rubrics

1 Introduction

Argumentative writing is a core academic and civic competency: it requires learners to formulate claims, support them with relevant evidence, and articulate coherent lines of reasoning Crossley et al. (2024). Assessing such writing is therefore central to educational evaluation and instruction. Unfortunately, manual scoring of argumentative essays remains labor-intensive, time-consuming, and susceptible to inter-rater variability when judgment must extend beyond surface linguistic features to the coherence, rigor, and persuasiveness of students' reasoning Wang et al. (2025b).

Automated Essay Scoring (AES) has long been proposed as a scalable, timely, and potentially more consistent alternative Wang et al. (2025b). Early AES systems, ranging from hand-engineered linguistic pipelines—namely Natural Language Processing—to deep learning architectures such as LSTM Yu et al. (2019) and BERT Koroteev (2021), achieved moderate success in predicting holistic or trait-level scores Xu et al. (2024). However, the AES literature has historically focused on general writing quality rather than argumentative writing: most established AES systems evaluate linguistic accuracy, coherence or overall proficiency, with limited reasoning-oriented constructs. But assessing argumentation requires validity measures and annotation schemes that differ from holistic or language-proficiency scoring Wachsmuth et al. (2017); Romberg et al. (2025). A parallel body of work in computational argumentation provides formal and empirical tools for argument mining Lawrence and Reed (2019), including evidence detection Rinott et al. (2015), stance detection Küçük and Can (2020), implicit warrant reconstruction Habernal et al. (2018). This area of work explores ways of analyzing the internal structure and validity of arguments beyond surface linguistic features. While insights from computational argumentation have been partially integrated into AAES pipelines (*e.g.*, Stab and Gurevych (2017)), core argumentative properties such as evidential sufficiency, inference validity, or counterargument handling, are rarely modeled explicitly.

The emergence of Large Language Models (LLMs)—from GPT-3 OpenAI (2023), GPT-4 OpenAI (2024) to Llama Dubey et al. (2024), Deepseek Liu et al. (2024), or Gemini Yoshida (2025)— has reshaped expectations for AES and AAES. Unlike earlier models, LLMs exhibit fluency in natural language understanding, discourse

modeling, and context-sensitive reasoning, thus can reason over an extended essay, integrate specific rubric descriptions, and generate structured scoring rationales Wang et al. (2025b). As a result, LLM-based AES has rapidly expanded into a distinct, fast-moving research domain. In particular, the "LLM-as-a-Judge" paradigm Zheng et al. (2023), where LLMs are being used as evaluators to approximate human judgments that would otherwise be costly and time-consuming to obtain Gu et al. (2024), has enormous potential for AES where expert annotation is expensive. Despite this rapid growth, the field remains conceptually unsettled and methodologically fragmented. Yet this growth has occurred without a consolidated understanding of the field's foundations. Studies operationalize AAES in heterogeneous ways—zero-shot grading, rubric-aware prompting, chain-of-thought reasoning, few-shot learning, fine-tuning, or multi-agent evaluation—while relying on datasets that vary widely in prompt design, rater configurations, trait definitions, and learner populations Emirtekin (2025).

Many scoring rubrics emphasize rhetorical effectiveness and linguistic fluency, whereas traits aligned with argument quality—such as logical cogency, evidential sufficiency, and dialectical engagement—remain rare. This misalignment raises questions about construct validity and the extent to which current systems truly assess argumentative competence rather than stylistic proficiency.

Moreover, while early results suggest that LLMs can approach or even surpass traditional machine learning approaches Yoo et al. (2025); Eltanbouly et al. (2025), critical concerns persist. Reliability varies across the essay topics and populations; robustness analyses reveal sensitivity to prompting strategies, randomness in generation, and imbalances in score distributions; and construct-level validity remains insufficiently examined Huang et al. (2025); Emirtekin (2025). These issues are amplified when considering the FATEN principles—a set of ethical principles and practical dimensions (fairness, pedagogical augmentation, transparency, beneficence, and non-maleficence) designed to ensure that data-driven decision making is responsible, trustworthy, and beneficial for society—which highlight ethical, instructional, and operational risks associated with deploying AAES in classroom contexts Favero et al. (2025). Preliminary evidence points to potential style bias Farzi (2024), L1 sensitivity Liu et al. (2025b), compressed score ranges Jordan et al. (2025), and inconsistent pedagogical alignment of explanations Da Silva et al. (2025); Ormerod and Kwako (2024). Questions of data privacy, model safety, and environmental sustainability are even less frequently addressed Emirtekin (2025).

1.1 Limitation of current reviews

To date, no systematic or scoping review has synthesized the emerging literature of LLMs for AAES, while critically evaluating its methodological, psychometric, and ethical foundations. Existing reviews focus primarily on pre-LLM approaches to automated writing evaluation or on general LLM-based feedback generation without addressing scoring Yildiz Durak and Onan (2025); Sun et al. (2025a); Huang et al. (2025); Emirtekin (2025); ElMassry et al. (2025); Wang et al. (2025b); Xu et al. (2024). None of these reviews focuses on an argumentative essay. As a result, researchers and practitioners lack a coherent overview of:

1. How LLMs are currently employed to score argumentative essays;
2. the datasets, techniques, and evaluation practices shaping the field;
3. the validity, reliability, and fairness of LLM-based scoring; and
4. the extent to which current systems align with educational, psychometric, and ethical standards.

1.2 Research questions

This review addresses these gaps by conducting a systematic scoping and critical analysis of LLM-based AAES research published between January 2022 and October 2025, following PRISMA 2020 guidelines Page et al. (2021) and a preregistered protocol Van den Akker et al. (2025). From an initial corpus of 3,467 records, 46 studies met the inclusion criteria. We organize our analysis around two research questions:

- **RQ1:** How are LLMs currently employed for the automated scoring of argumentative essays and the provision of feedback in educational settings? Specifically, what techniques, datasets, and evaluation methodologies are used, and what methodological gaps or unresolved challenges remain?
- **RQ2:** To what extent do LLM-based AAES approaches align with human judgment, both in terms of psychometric validity and in relation to the FATEN principles Oliver (2019) guiding responsible educational assessment?

1.3 Contributions

The four major contributions of this scoping and critical review are the following:

- **A comprehensive mapping of the methodological landscape.** The review provides a structured synthesis across six dimensions: (1) datasets, (2) scoring traits, (3) LLM families, (4) technical approaches, (5) evaluation practices, and (6) analytical frameworks and findings. It reveals fragmentation in data design, trait definitions, and evaluation methodologies in current AAES research.
- **A theoretically grounded analysis of argumentative constructs.** By mapping 82 essay traits to the Argument Quality (AQ) framework Romberg et al. (2025), the paper demonstrates that existing datasets overwhelmingly emphasize rhetorical and linguistic features, while logical cogency, dialectical engagement, and reasoning quality remain severely underrepresented, raising concerns about construct validity in LLM-based AAES systems.
- **A critical demonstration of validity, reliability, robustness, and FATEN-alignment.** While LLMs often reach at least substantial agreement with human raters, their performance is highly sensitive to prompts, score distributions, proficiency levels, and sampling variance. The FATEN analysis identifies recurrent issues involving fairness (e.g., L1 and style bias), transparency, sensitivity to randomness, weak pedagogical alignment, and the concerning absence of privacy or deployment-safety considerations.

- **A research agenda for responsible AAES** emphasizing the need for: (1) construct-valid openly available datasets; (2) theoretically grounded, reasoning-oriented scoring traits; (3) psychometric modeling and transparent evaluation; (4) robust assessment protocols; and (4) responsible, equitable design frameworks for AAES deployment.

The remainder of the paper proceeds as follows. We first present the methodology, then report our results, and conclude with a discussion of the findings. A containing a comprehensive glossary of definitions and abbreviations is provided to support clarity and consistency throughout the manuscript.

2 Methodology

Methodologically, this manuscript follows the *Preferred Reporting Items for Systematic Reviews and Meta-Analyses* (PRISMA) 2020 guidelines Page et al. (2021). These guidelines are designed to improve the credibility of research outcomes by promoting standardized and transparent reporting in systematic reviews. PRISMA offers a checklist and flow diagram that help researchers clearly document each step of the review process, enhancing the methodological robustness of the review and facilitating the reproducibility of its findings. To further support transparency and reproducibility, the review protocol was preregistered on the Open Science Framework Van den Akker et al. (2025) and made publicly available.[1] Additionally, the review process was supported by *Rayyan*[2]—a web-based platform that facilitates the screening and management of search results—during the screening phase, and *Notion*,[3] a platform for data extraction and management, during the extraction phase.

The PRISMA methodology requires defining inclusion and exclusion criteria, followed by three steps: identification, screening and selection, which are detailed next.

2.1 Inclusion and Exclusion Criteria

Table 1 summarizes the inclusion and exclusion criteria applied in this review, which are elaborated below.

2.1.1 Inclusion criteria

When selecting relevant manuscripts for this literature review, we considered the following inclusion criteria:

1. Publication period

The publication period was between January 2022 and October 2025. Note that the starting date aligns with the emergence of LLMs, marked by the release of popular and widely adopted models such as GPT-3 and GPT-4, which fundamentally transformed the landscape of AAES and feedback generation.

[1] https://osf.io/kd9f4/overview?view_only=9141a724dc8d4591b453e243df33c588.
[2] https://www.rayyan.ai.
[3] https://www.notion.so.

Table 1: Summary of the most relevant inclusion and exclusion criteria of manuscripts.

Criterion	Include	Exclude
Date range	Between January 2022 and October 2025	Before 2022
Target population	Students in secondary school, high school, undergraduate, or graduate education and adult learners	Primary school students
Essay type	Argumentative essays written in first (L1) or second (L2) language	Non-essay format (*e.g.*, narrative, descriptive essays, or short-answer responses)
Language	Manuscripts published in English and where the methods are tested on at least one English dataset	Manuscripts written in other languages, and/or testing only on non-English dataset(s)
Assessment type	Multi-trait rubric-based essay scoring	No scoring rubrics, scoring at the argument level only, rewriting tasks, essays dependent on external sources
LLM-focused	LLMs are used for AAES	Other models, such as BERT, traditional deep learning or other machine learning techniques are used for AAES

2. Population focus

The target population are learners in secondary school, high school, undergraduate, graduate, or adult education contexts, including professional or lifelong learning scenarios. Studies involving primary school students are excluded due to the developmental simplicity of young learners' argumentative writing, which limits comparability in assessing reasoning quality and rhetorical complexity.

3. Essay type

The focus are argumentative essays written in either a first language (L1) or second language (L2), reflecting the growing societal demand for critical thinking and argumentation skills in an era shaped by LLMs Favero et al. (2024, 2025). Argumentative writing requires the assessment of rhetorical coherence, logical cogency, and persuasiveness—dimensions that extend beyond surface-level linguistic quality. However, current AAES methods often lack reliability, interpretability, and validity in evaluating such higher-order reasoning skills.

4. Argumentative essay dataset selection

To determine whether a dataset qualifies as argumentative, we applied the following two conditions: (1) the dataset explicitly mentions argumentative, opinion, or persuasive writing in the title or abstract; or (2) the essay prompts require participants to take a position, persuade, or express an opinion; the scoring traits include content- or structure-related dimensions (*e.g.*, coherence, organization, thesis development) rather than focusing solely on surface-level features such as grammar or spelling.

5. Language

Eligible manuscripts were required to be written in English, performing the assessment of essays also written in English, to ensure comparability across corpora and to reflect the predominance of English as both the primary language of science and the main working language in this research domain.

6. Assessment

Regarding the assessment criteria, manuscripts should focus on multi-trait, rubric-based essay scoring, acknowledging that argumentative writing assessment must capture multiple dimensions (*e.g.*, content, structure, argument quality, and reasoning depth) rather than relying on single holistic scores.

7. Technical approach

This review is devoted to the use of Large Language Models (LLMs)—including but not limited to GPT, Llama, Deepseek, Mistral, or similar transformer-based architectures—as the main technical approach to perform automated argumentative essay scoring. Earlier approaches based on other machine learning methods, including LSTM or BERT architectures, were excluded, as they have already been thoroughly reviewed in prior AAES literature Yu et al. (2019). Focusing on LLMs allows this review to capture not only scoring performance but also the emergent capability of LLMs to generate formative feedback, interpret student reasoning, and align with human evaluative judgment.

2.1.2 Exclusion criteria

Conversely, we defined as exclusion criteria:

1. Format-related

Theses, book chapters, opinion pieces, purely theoretical papers, longitudinal studies, and student surveys without computational analysis.

2. Content-related

We excluded: (1) studies focused on vision-dependent, coding-intensive, or highly technical domains (*e.g.*, physics, mathematics, L2 translation); (2) studies dealing exclusively with general LLM feedback, feedback mechanisms without explicit reference to essay scoring, or tools centered on dialogue systems, chatbots, writing assistants, AI-driven scaffolding, essay revision, or rewriting tasks; (3) comparative studies of AI text generation systems not directly addressing assessment or scoring; and (4) studies emphasizing student or teacher perspectives on feedback without including automated scoring components.

3. LLM-use

We excluded any manuscripts where LLM-generated feedback is not integrated with AAES or related evaluative processes.

2.2 Identification

Once the inclusion and exclusion criteria were defined, we proceeded with the first step in the PRISMA methodology, namely the identification of suitable and relevant studies. This step consists of three sub-steps: defining the query string, selecting suitable databases and refining the search procedure.

2.2.1 Definition of the query string

Aligning with the inclusion and exclusion criteria, the search was conducted using the Boolean query string whose characteristics are detailed in Table 2.

Table 2: Properties of the Boolean query string used to retrieve studies in this systematic review. see Section A for a comprehensive glossary of definitions and abbreviations.

Search dimension	**Keywords and Boolean operators**
Date range	2022–2025
Essay	"essay" OR "text"
Essay type	"persuasive" OR "argumentative"
Language	"english"
Assessment type	"scoring" OR "grading" OR "automated writing evaluation" OR "assessment" OR "feedback" OR "AES" OR "AWS"
LLM-focused	"LLM" OR "GPT" OR "language models"

2.2.2 Selection of databases

A comprehensive search was performed on 8 major academic databases: arXiv, ERIC, PubMed, SpringerLink, ACM Digital Library, Web of Science, ScienceDirect and Google Scholar. These databases were selected for their broad coverage of educational technology, computer science, and applied linguistics research.[4] While we recognize the limitations of non-peer-reviewed content, we considered grey literature, such as ArXiv, university repositories, and technical reports, to capture the most recent and relevant developments in a rapidly evolving research area. Given the emergent and dynamic nature of the field, many valuable contributions—particularly novel methods and emerging findings—are available as preprints prior to formal peer-reviewed publication. Including grey literature helps ensure comprehensive coverage of the state of the art and avoids overlooking impactful work that may not yet appear in traditional academic venues.

[4] More details can be found in B.1.

2.2.3 Search validation and refinement

To ensure completeness and recall, the search results were cross-validated against a predefined set of benchmark publications known to be relevant and representative of the domain, according to the following procedure:

1. Benchmark identification

Prior to executing the final search queries, a reference list of approximately 50 key studies on LLM-based argumentative automated essay scoring was compiled. These studies were identified through preliminary scoping searches, backward and forward citation tracking, and expert recommendations, ensuring a diverse range of contributions to the field.

2. Coverage verification

After retrieving relevant manuscripts from the databases, the resulting corpus was examined to confirm the inclusion of all benchmark studies. Missing references prompted refinement of the search terms, Boolean operators, and/or database selection to enhance recall while preserving precision.

3. Iterative refinement

Pilot searches were conducted iteratively until the strategy consistently retrieved all benchmark publications alongside additional relevant works.

One week after the initial searches, a complementary verification was performed on Google Scholar by reviewing the first two pages of results for each query. Furthermore, a deep search was conducted using GPT-5's Deep Research capabilities OpenAI (2025). In this case, the prompt consisted of a concise description of the objectives of the literature review and exact search queries to identify potentially relevant studies not captured through traditional databases.

The last search was performed on October 16th 2025, yielding a total of 3,467 manuscripts that met the specified query established in Table 2 from the following sources (the number of manuscripts is included in parenthesis): Google Scholar (2,029), Springer (755), ACM Digital Library (382), ScienceDirect (186), ArXiv (82), Web of Science (24), ERIC (9), and PubMed (0).

2.3 Screening

All retrieved records were exported to *Rayyan*,[2] and duplicates were removed both automatically and manually. Manuscripts were screened based on the predefined inclusion and exclusion criteria (Table 2). Full-text articles that passed initial screening were reviewed in detail to ensure methodological and topical relevance.[5] From the initial set of 3,467 records, 592 were removed due to duplication, 2,999 were excluded after manual abstract screening, and 200 full-text articles were assessed for eligibility, resulting in a final number of 46 included studies.

Figure 1 depicts the flow diagram of the PRISMA methodology applied for our purpose.

[5] More details about this process can be found in B.2.

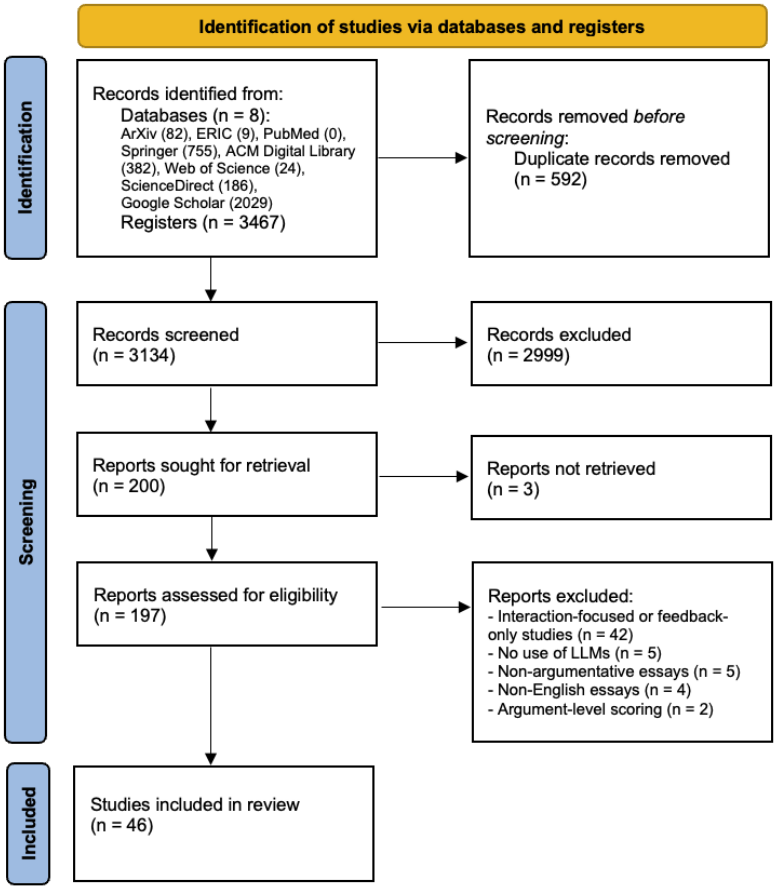

Fig. 1: PRISMA flow diagram of the adopted methodology for data collection as per Page et al. (2021).

2.4 Data extraction

Finally, to ensure a comprehensive and reproducible synthesis of the literature, the following entities were systematically extracted from each source using a structured coding framework: (1) manuscript's metadata; (2) manuscript's context and objectives; (3) methodological characteristics; (4) LLM-specific technical information; (5) target tasks and argumentative essay scoring dimensions; (6) outcome measures and (7) indicators related to risk of bias and trustworthiness. The aim is to enable both quantitative and qualitative meta-synthesis across diverse study designs, methodologies, and reporting standards.[6]

[6] More details can be found in B.3.

3 Results

Among the 46 studies included in this review, 4 (9%) were published in 2023, 20 (43%) in 2024, and 22 (48%) in 2025, indicating a sharp increase in research activity over the past three years. Nine studies (20%) appeared in education-focused journals, 7 (15%) in venues primarily dedicated to artificial intelligence, 17 (37%) in interdisciplinary outlets bridging education and technology or AI, 9 (20%) were preprints at the time of analysis, and the remaining 4 (9%) were published in other types of venues. These publication patterns underscore both the rapid growth and the inherently multidisciplinary nature of research on automated argumentative essay scoring using LLMs, situated at the intersection of educational assessment, natural language processing, and artificial intelligence Emirtekin (2025). Figure 2 summarizes the main dimensions addressed in this literature review, namely, (1) essay datasets, (2) traits, (3) LLMs, (4) techniques, (5) metrics, and (6) analytical perspectives. Each branch highlights key subcategories (*e.g.*, established datasets, prompt engineering technique, reliability and agreement metrics, or FATEN analysis) that inform the in-depth analyses presented in the following sections.[7]

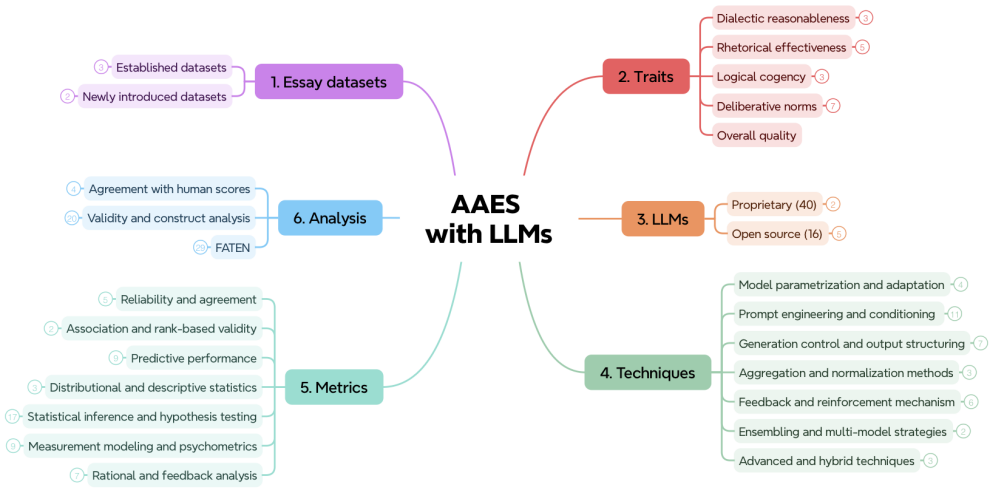

Fig. 2: Taxonomy of the six core dimensions that structure current research on Large Language Model–based Automated Argumentative Essay Scoring (AAES). The taxonomy synthesizes how existing studies vary in (1) essay datasets, (2) scoring traits, (3) LLM families, (4) technical approaches, (5) evaluation metrics, and (6) analytical perspectives, providing a high-level map of the methodological landscape reviewed.[7]

[7]The full taxonomy can be found here: https://app.xmind.com/share/5sShNF4G.

3.1 Essay datasets and learner populations

Tables 3 and 4 summarize all annotated essay datasets employed across the included studies. For each dataset, the tables report:

1. The year of publication;
2. The availability status: yes (Y), paid license required (P), maybe/upon request (M) or not available/proprietary (N);
3. The number of essays (or number of argumentative essays when the dataset also contains no argumentative essays);
4. The number of essay prompts, *i.e,* the topic of each essays;
5. The minimum number of raters per essay;
6. The number of essay traits or H when the essays are only scored holistically (see definitions in 3.2);
7. The characteristics of the writer population;
8. The first-language (L1) of the writers when available (noting that the all essays are written in English) ;
9. The included studies that report using the dataset;
10. When applicable, the original publication introducing the dataset.

Datasets that include rater rationales or feedback are explicitly marked with ‡, and missing information is denoted as *U*. In the following subsections, we provide a more detailed description of the datasets.

3.1.1 Essay datasets

We identified 29 different benchmark datasets, summarized in Tables 3 and 4, that were used in the 46 studies included in this review. We observe two types of datasets: (1) *established* datasets, which have been used by other works in the literature; and (2) *newly introduced* datasets, which were only used in the study where they were introduced.

Table 3: Characteristics of established argumentative essay datasets used in AAES research. For each dataset, the table reports availability, number of essays, prompts, raters and scoring traits, learner population, and L1 distribution. Availability status: Y: Yes, P: Paid license required, M: Maybe/upon request, N: Not available/proprietary. U: Unknown/not reported, H: Holistic.

Dataset	Year	Avail.	Nb. essays	Nb. prompts	Nb. raters	Traits	Population	L1[1]	Selected studies	Ref.
Student with English as a First Language										
ASAP **ASAP++**	2012 2018	Y	4,877	4[2]	1	H,4-6	Secondary[+]	–	Cai et al. (2025), Eltanbouly et al. (2025), Hou et al. (2025), Kundu and Barbosa (2024), Lee et al. (2024), Mansour et al. (2024), Oketch et al. (2025), Ormerod and Kwako (2024), Shermis (2025), Stahl et al. (2024), Tang et al. (2024), Wang et al. (2025a), Xiao et al. (2025)	Hamner et al. (2012); Mathias and Bhattacharyya (2018)
ELLIPSE	2023	Y	6,482	29	2	H,6	Secondary	–	Chen et al. (2024), Eltanbouly et al. (2025), Hou et al. (2025)	Crossley et al. (2023)
Student with English as a Foreign Language										
TOEFL 11	2013	P	12,100	8	2	H	Uni. entrance	*	Lee et al. (2024), Liu et al. (2025b), Liu et al. (2025a), Mizumoto and Eguchi (2023), Yeung (2025), Yoshida (2025)	Blanchard et al. (2013)
ICNALE[3]	2013	Y	5,600	1	1	H	Undergrad.	Asian[4]	Lin and Pu (2024), Bui and Barrot (2025b), Bui and Barrot (2025a)	Ishikawa (2013)
ICNALE GRA[‡]	2020	Y	200	1	80	H,10	Undergrad.	Asian[4]	Uchida (2024), Yamashita (2024)	Ishikawa (2020)
DET-Coh[‡]	2022	N	500	U	1	H	Test-taker	**	Naismith et al. (2023)	Cardwell et al. (2022)
FCE	2011	Y	2,466	5[5]	U	H	Test-taker	U	Oketch et al. (2025)	Yannakoudakis et al. (2011)
Pathway 2.0	2020	N	344	2	2	H,4	Secondary	U	Tate et al. (2024)	Olson et al. (2020)
PERSUADE 2.0	2024	Y	25,000	15	2	H	Secondary	U	Tate et al. (2024)	Crossley et al. (2024)

[‡] : The dataset contains feedback from the annotators
[+] : Secondary school refers to both middle and high school
[*] : Arabic, Chinese, French, German, Hindi, Italian, Japanese, Korean, Spanish, Telugu, and Turkish
[**] : Chinese, Arabic, Spanish, Telugu, English, Bengali, Gujarati
[1] L1: Student's first language
[2] : Prompts *7* for Tang et al. (2024), Prompts *1* and *2* for Wang et al. (2025a) and Prompts *1*, *2*, *7*, and *8* for the other studies using ASAP or ASAP++
[3] : ICNALE Written Essays and ICNALE Edited Essays module
[4] Asian: indicates L1s from multiple Asian languages (details vary by subset)
[5] : Prompts *2*, *3*, *4*, *5a* and *5b*

Table 4: Characteristics of newly introduced datasets appearing in the reviewed studies. For each dataset, the table reports availability, number of essays, prompts, raters, and scoring traits, learner population, and L1 distribution. Availability status: Y: Yes, P: Paid license required, M: Maybe/upon request, N: Not available/proprietary. U: Unknown/not reported, H: Holistic.

Study[1]	Avail.	Nb. essays	Nb. prompts	Nb. raters	Traits	Population	L1[2]
Student with English as a First Language							
ASAP++ (custom version)							
Altamimi (2023)	N	50	1	U	4	Secondary^{+}	–
Graduate Record Examination (GRE)							
Jordan et al. (2025)‡	M	48	8	U	H	High school, undergrad.	English, U
Student with English as a Foreign Language							
International English Language Testing Systems (IELTS)							
Chen et al. (2025)	P	23	23	U	4	Test-taker	U
Xu et al. (2025)‡	N	5,088	U	2	4	Test-taker	U
Uyar and Büyükahıska (2025)	N	50	U	1	4	Preparatory	Turkish
Test of English as a Foreign Language (TOEFL)							
DREsS Yoo et al. (2025)	Y	2,279	22	1	3	Undergrad.	Korean
National English Ability Test (NEAT)							
Shin and Lee (2024)‡	M	50	1	1	4	High school	Korean
Other English Foreign Language test							
Yavuz et al. (2025)	Y	3	3	15	5	Preparatory	Turkish
Pack et al. (2024)	M	119	1	2	H	Entrance	Asian[3]
Kim (2025)	N	300	5	2	4	Entrance	*
CSEE Xiao et al. (2025)	Y	13,372	2	U	3	Entrance	Chinese
Gao et al. (2025)‡	N	84	1	4	4	Entrance	Chinese
Bouziane and Bouziane (2024)	N	100	3	U	10	Undergrad.	U
Jin et al. (2025)	M	117	1	2	9	Undergrad.	Chinese
Farzi (2024)‡	N	60	2	1	4	Undergrad.	**
Mahdi and Alkhateeb (2025)	N	33	U	2	8	Undergrad.	Arabic
Geckin et al. (2023)	N	43	1	5	H	Undergrad.	Turkish
Tekin and Aydogdu (2024)	M	59	1	1	8	Undergrad.	Turkish
Arif Cem Topuz et al. (2025)	N	210	1	2	5	Undergrad.	Turkish
Albuquerque Da Silva et al. (2024); Da Silva et al. (2025)	N	33	1	1	6	Military students	Portuguese

‡ : The dataset contains feedback from the annotators
$^{+}$: Secondary school refers to both middle and high school
* : Twenty-five L1 backgrounds, with the top three L1s being Japanese ($n = 83$), Chinese ($n = 68$), and Korean($n = 61$)
** : Twelve L1s including Chinese, French, Cantonese, Arabic, Spanish, Japanese, Somali, and Dari
[1] : First author, year of publication, reference
[2] L1: Student's first language
[3] Asian: indicates L1s from multiple Asian languages (details vary by subset)

From all 29 datasets, 9 (31%) were established, namely, ASAP and ASAP++ Hamner et al. (2012); Mathias and Bhattacharyya (2018), ELLIPSE Crossley et al. (2023), TOEFL11 Blanchard et al. (2013), ICNALE Ishikawa (2013, 2020), DET-Coh Cardwell et al. (2022), FCE Yannakoudakis et al. (2011), Pathway 2.0 Olson et al. (2020), and PERSUADE 2.0 Crossley et al. (2024). These datasets were used in 31 studies (67%), reflecting their strong anchoring in the AAES literature Emirtekin (2025). In contrast, 20 datasets[8] (69%) were newly introduced within the primary studies and were not used by any other study in our corpus (Table 4). These datasets tend to be smaller in scale, more heterogeneous in design, and more constrained by task- or context-specific aims.

The **Automated Student Assessment Prize** (ASAP, Hamner et al. (2012); ASAP++, Mathias and Bhattacharyya (2018); or other variants Altamimi (2023)) emerged as the most frequently used benchmark dataset, appearing in 14 studies (30%). ASAP++ is an expanded and refined version of the original ASAP corpus Hamner et al. (2012), containing thousands of student essays annotated for multiple traits including argumentation, coherence, and holistic quality. It comprises 8 prompts (*i.e,* essay topics): 4 source-dependent and 4 source-independent prompts, whereby *1*,*2*,*7*, and *8* are argumentative. This dataset is relevant for argumentative essay scoring because it provides large-scale, rubric-aligned annotations that support both analytic and holistic evaluation, making it the most widely used benchmark for AAES with LLMs Sun et al. (2025a); Huang et al. (2025). Other datasets were used less consistently. Among them, **ELLIPSE** (n = 3, 7%) consists of secondary-school essays annotated for multiple analytic traits, including argumentation and elaboration quality Crossley et al. (2023). **TOEFL11** (n = 6, 13%) contains essays written by English learners from 11 different L1 backgrounds, scored for holistic proficiency Blanchard et al. (2013). The **ICNALE** (GRA) corpus (n = 5, 11%) provides essays from Asian learners across proficiency levels, annotated for holistic quality depending on the subset with many argumentative prompts Ishikawa (2013, 2020).

3.1.2 Dataset size and availability

Established datasets contain at least a few hundred essays, with notable exceptions such as PERSUADE 2.0 (25,000 essays) and TOEFL11 (12,100 essays). In contrast, newly introduced datasets are substantially smaller, with many (n = 14, 70%) containing fewer or equal to 120 essays (*e.g.,* Uyar and Büyükahıska (2025); Tekin and Aydogdu (2024), though some reach 200-300 essays (*e.g.,* Arif Cem Topuz et al. (2025); Tate et al. (2024)). Outliers include a large IELTS-derived dataset Xu et al. (2025), with 5,088 essays, the TOEFL DReSS dataset Yoo et al. (2025) with 2,279 essays, and a major entrance-exam dataset: CSEE Xiao et al. (2025) with 13,372 essays.

Regarding availability, among well-established datasets, six (67%) datasets are freely available (Y), one (TOEFL11) requires a paid license (P), and two (DET-Coh and Pathway 2.0) are not available. Among the newly introduced datasets, four (20%) are freely available, four (20%) are available upon request or subject to conditional access (M), and 10 (50%) are proprietary or not publicly accessible (N). This uneven

[8] Note that some studies use multiple datasets.

availability limits reproducibility and represents a structural barrier to comparative AAES research.

3.1.3 Essay prompt coverage

Most datasets contain a limited number of essay prompts: 11 datasets (44% of the datasets for which the number of prompts is known) rely on a single prompt, and 17 datasets (68%) use five or fewer prompts, indicating a still-prevailing dependence on narrow-topic, single-prompt assessment. Only a few benchmarks exhibit broad prompt coverage: PERSUADE 2.0 (15 prompts), DREsS, a TOEFL-derived dataset (22 prompts), an IELTS-derived dataset Chen et al. (2025) (23 prompts), and ELLIPSE (29 prompts). Four datasets (14%) did not report the number of prompts used.

3.1.4 Rater configuration

Rater configurations vary substantially across datasets. Roughly one third of datasets ($n = 10$, 34%) relies on a single trained or expert rater, consistent with common practice in large-scale assessments. Another third employs two raters per essay, and several datasets use multiple raters only on sub-samples to estimate inter-rater agreement Lin and Pu (2024); Lee et al. (2024). Two studies adopt high–rater-count designs—15 raters in Yavuz et al. (2025) and 80 raters in the ICNALE GRA dataset Ishikawa (2020)—to analyze demographic variability in scoring practices. Five datasets (17%) do not report the number of raters.

3.1.5 Population characteristics

The populations represented in the datasets are dominated by undergraduate students ($n = 11$, 38%) and high-school, preparatory, or college-entrance test takers ($n = 9$, 31%). A smaller proportion of datasets focus on secondary-school students ($n = 4$, 14%) or general English test takers ($n = 4$, 14%). Notably, although the search strategy included lifelong-learning and adult-education contexts, no datasets targeting these populations were identified.

3.1.6 First-language (L1) backgrounds

Aside from the ASAP and ELLIPSE datasets—where the writers' L1 can be reasonably assumed to be English although this is not explicitly reported—the remaining datasets focus on essays written in English as a foreign language, allowing an analysis of L1 distribution. Twelve datasets (74%) include Asian L1s (Chinese, Korean, Japanese, and/or mixed Asian categories). Within this group, Chinese L1 writers appear in 10 datasets (37%), mostly in corpora newly introduced by the studies. Turkish L1 writers are represented in 6 datasets (22%), again largely stemming from newly created datasets. Across all datasets, L1 information is missing or incomplete in 11 cases (41%), while lower-frequency L1s include Arabic, Spanish, French, German, Portuguese, and others. Overall, coverage of global L1 diversity remains highly uneven.

3.2 Traits

In AES, there are two main approaches for assessing writing quality: holistic and trait-based scoring. Holistic scoring "employs a reader's full impression of a text without trying to reduce her judgment to a set of recognizable skills" (Huot 1990, p.201), and results in one single letter grade or score. An example is used to evaluate the Writing Section of the Test of English as a Foreign Language (TOEFL) Internet-based Test (iBT) examination.[9] In contrast, essay traits break down an essay's quality into individually-evaluated components. For example, the Writing section of the B2 First English examination evaluates Content, Communicative achievement, Organisation and Language (Cambridge University Press & Assessment 2023, p.31-33). In this section, we explore how essay traits are defined and used across existing AAES datasets, and analyze the extent to which they capture different dimensions of argumentative quality.

3.2.1 Diversity of essay traits

Out of the 29 datasets presented in Tables 3 and 4, 22 include essay traits different from holistic scores (73%). Collectively, the datasets define 82 distinct essay trait names. A large number of these focus on surface-level and linguistic dimensions, particularly the detection of grammatical and orthographical errors (*e.g.*, *Grammar and Spelling* in Tekin and Aydogdu (2024)). Other frequently studied dimensions include vocabulary usage and lexical sophistication (*e.g.*, *Vocabulary and Phraseology* in ELLIPSE Crossley et al. (2023)), relevance to the prompt (*e.g.*, *Relevance* in Bouziane and Bouziane (2024)), organizational structure, coherence and cohesion (e.g., *Organisation* in Prompts *7* and *8* of ASAP; Hamner et al. (2012)), thesis clarity (*e.g.*, *Content* in Mahdi and Alkhateeb (2025)), and argument persuasiveness (*e.g.*, *Task completion* in Shin and Lee (2024)).

3.2.2 Essay trait ambiguity

Although some of these datasets share common essay trait names (*e.g.*, 9 datasets, or 30%, have *Content* as a trait), each dataset comes with its own scoring rubric and definitions. As a result, two traits with the same name can ultimately evaluate slightly different constructs. For instance, *Content* as defined in CSEE Xiao et al. (2025) only specifies that arguments should be "complete" and "related to the topic," whereas Mahdi and Alkhateeb (2025) additionally asks for arguments to be "clear" and accompanied with "relevant supporting details" (*i.e.,* evidence). Conversely, two different traits may in fact be defined similarly. For example, ELLIPSE's *Vocabulary* Crossley et al. (2023) is comparable to the IELTS's *Lexical resources* component Chen et al. (2025); Xu et al. (2025); Uyar and Büyükahıska (2025). Conversely, two traits that differ in name can in fact be defined very similarly. For example, Farzi (2024)'s *Grammatical accuracy and range* and Mahdi and Alkhateeb (2025); Yavuz et al. (2025)'s *Grammar* are closely aligned: both capturing concepts like "complex structures" or "complex sentences" (*i.e.*, syntax), and "clarity". As a consequence of this, we cannot easily compare studies which draw from different sources of essay

[9] For the rubric, see here: https://www.ets.org/pdfs/toefl/toefl-ibt-performance-descriptors.pdf.

data. The dependency to a scoring rubric, and ambiguity of essay traits in general, is a clear limitation of trait-based datasets, and an obstacle to AAES research.

While there have been some efforts to rationalize this ambiguity (*e.g.*, the authors of DREsS Yoo et al. (2025) propose a way to standardize and unify the ASAP Prompts *7* and *8* Hamner et al. (2012), ASAP++ Prompts *1* and *2* Mathias and Bhattacharyya (2018), and ICNALE EE Ishikawa (2013) datasets with their own rubric), we identify a clear consistency gap in the characterization and description of these essay traits.

3.2.3 From essay traits to argument quality categories

To provide a common framework, we classified the essay traits of the 29 datasets into the five Argument Quality (AQ) categories defined in Romberg et al. (2025, Table 5), namely *Dialectical reasonableness*, *Rhetorical effectiveness*, *Logical cogency*, *Deliberative norms*, and *Overall quality*. For reference we include a copy of the definitions of the AQ categories in Table 11. We adopt this specific framework because: (1) it provides a theoretically grounded taxonomy of AQ that is well aligned with the goals of trait-based essay scoring, that is, to decompose writing quality into interpretable and meaningful dimensions; (2) instead of conflating writing quality with linguistic proficiency, this model distinguishes between rhetorical form, reasoning quality, and deliberative engagement, thereby capturing dimensions that are central to argumentative writing but that we found to be lacking in existing AAES datasets. Ultimately, doing this enables us to evaluate not only which linguistic properties are assessed, but whether and to what extent existing datasets capture argumentative competence. Our final classification can be found in Figure 8,[10] and provides a framework for comparing the datasets.

Figure 3 depicts the coverage of each surveyed dataset in terms of the AQ categories, noting that a trait can belong to more than one category. On average, the essay traits are heavily concentrated in the *Rhetorical effectiveness* (86%) category, which signifies a strong focus on language, style, and organization. In contrast, *Dialectical reasonableness* and *Deliberative norms* receive much lower and more variable coverage (both averaging at 25%), suggesting that engagement with the prompt and deliberative aspects of argument are inconsistently represented. *Logical cogency* is almost entirely absent (5%), with only a few datasets including any traits which evaluate the reasoning process at the fine-grained level of individual arguments and their components (premises, conclusion). Similarly, *Overall quality* is also sparsely covered (6%), reflecting the general emphasis on analytic rather than holistic assessment. The essays and metadata corresponding to Prompts 1&2 of ASAP++ Mathias and Bhattacharyya (2018), Arif Cem Topuz et al. (2025); Albuquerque Da Silva et al. (2024); Da Silva et al. (2025) stand out as the data with the best coverage of the AQ categories. Finally, Figure 4 shows the aggregate distribution of the five AQ categories across the 29 datasets. Again, this analysis reveals a strong skew towards *Rhetorical effectiveness* which is evaluated by 56% of a dataset's traits on average. Overall, these results highlight a certain bias towards rhetorical form and away from reasoning and deliberation in current datasets, which is what AAES should be focusing on instead.

[10] A graphical visualization of can be found here: https://app.xmind.com/share/jzN3r6Lh.

	Dialectic reasonableness	Rhetorical effectiveness	Logical cogency	Deliberative norms	Overall quality
ASAP++ [Prompts 1&2]	20%	100%	20%	60%	20%
ASAP [Prompt 7]	25%	75%	0%	0%	0%
ASAP [Prompt 8]	17%	100%	17%	67%	17%
ICNALE GRA	10%	60%	0%	50%	0%
Pathway 2.0*	25%	75%	0%	0%	0%
Altamimi (2023)*	25%	75%	25%	0%	0%
ELLIPSE	0%	100%	0%	0%	0%
Tekin and Aydoğdu (2024)	13%	100%	0%	13%	0%
Farzi (2024)	33%	100%	0%	33%	0%
Bouziane and Bouziane (2024)	30%	100%	0%	20%	0%
Shin and Lee (2024)	50%	100%	0%	50%	0%
Albuquerque Da Silva et al. (2024) and Da Silva et al. (2025)	17%	83%	17%	50%	33%
Yavuz et al. (2025)	20%	100%	0%	20%	0%
DREsS (Yoo et al., 2025)	33%	100%	0%	33%	0%
Mahdi and Alkhateeb (2025)	25%	75%	13%	38%	0%
IELTS (X. Chen et al., Xu et al., Uyar and Büyükahıska, 2025)	0%	75%	0%	0%	25%
Jordan et al. (2025)*	20%	60%	0%	20%	20%
Arif Cem Topuz et al. (2025)	40%	100%	20%	60%	20%
Kim (2025)*	25%	75%	0%	0%	0%
Jin et al. (2025)	33%	89%	0%	33%	0%
CSEE (Xiao et al., 2025)	33%	67%	0%	0%	0%
Gao et al. (2025)*	50%	75%	0%	0%	0%
Average	25%	86%	5%	25%	6%

Fig. 3: Coverage of the essay traits of the datasets presented in Tables 3 and 4 across the five Argument Quality (AQ) categories introduced in Table 11. Each cell reports the proportion of traits of a dataset that belong to a given AQ category (rounded to 2 decimal places) according to our classification in Figure 8. Asterisks (*) denote datasets for which no formal definitions were found; resulting mappings were inferred from trait names only.

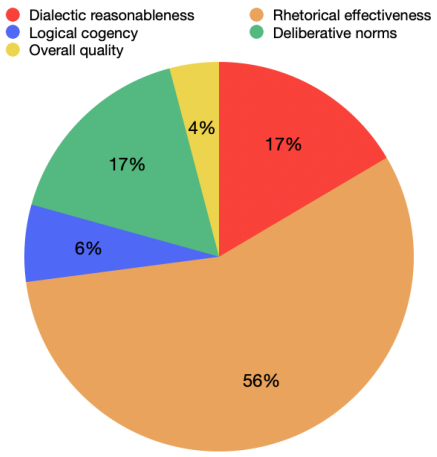

Fig. 4: Average weight of each AQ category (as defined in Table 11) in the trait sets of the datasets presented in Tables 3 and 4 (rounded to the nearest whole number) given our classification in Figure 8.

3.2.4 Fine-grained coverage

While the category-level analysis highlights broad differences across AQ dimensions, the sub-category breakdown exposes additional structural asymmetries within each dimension (Figure 8).

For instance, the coverage within *Rhetorical effectiveness* is highly uneven. Traits linked to *Clarity* (*e.g., Grammar*, *Syntax*, *Conventions*, and *Vocabulary*) and *Arrangement* (*e.g., Cohesion*, *Coherence*, *Organisation*) dominate this category, being present in every single trait-based dataset. On the other hand, sub-categories such as *Credibility*, *Appropriateness*, and *Emotional appeal* are sparsely represented. This suggests that *Rhetorical effectiveness* is largely evaluated in terms of linguistic correctness and discourse-level organization, rather than rhetorical effect.

A comparable imbalance is observed in *Deliberative norms*, where only one sub-category—namely *Rationality*—is consistently represented across datasets. Traits related to support, evidence, and logical analysis appear in a small number of recent datasets (*e.g.,* ICNALE GRA; Ishikawa (2020)), whereas other deliberative dimensions such as *Interactivity*, *Equality*, *Civility*, *Common good reference*, *Constructiveness* and *Alternative forms of communication* are (almost) entirely absent. This indicates that even when deliberative norms are included, they are operationalized narrowly and primarily in terms of justification rather than dialogical engagement.

These results show that even when AQ categories are nominally covered, such as in the case of *Rhetorical effectiveness*, their internal structure can still be unevenly represented, with substantial gaps in dimensions related to persuasion and deliberation.

3.3 LLMs

Regarding the LLMs that are used in the studies, we analyze their licensing model, scale and reasoning capabilities.

3.3.1 Licensing models: proprietary vs open

Among the 46 studies included in this survey, 40 (87%) employed one or more variants of the proprietary GPT family of models, including GPT-3.5 Kundu and Barbosa (2024), GPT3 text-davinci-003 Mizumoto and Eguchi (2023), GPT-4 Bouziane and Bouziane (2024), GPT-4o Jin et al. (2025), and GPT-4o mini Wang et al. (2025a). In addition, 8 studies (17%) also incorporated other proprietary LLMs, such as Claude (*e.g.,* versions 2; Pack et al. (2024); Tang et al. (2024)), 3.5 Sonnet Jin et al. (2025); Wang et al. (2025a), 3.5 Haiku Wang et al. (2025a); Yoshida (2025)), Gemini families Mahdi and Alkhateeb (2025); Yoshida (2025); Bui and Barrot (2025a), and Bard Yavuz et al. (2025). Only 6 studies (13%, *e.g.,* Cai et al. (2025); Eltanbouly et al. (2025)) did not rely on any GPT-based model. Sixteen studies (35%) used open-weighted LLMs either exclusively or in combination with proprietary ones. Llama was the most frequently used open model, appearing in 12 studies (26%), with versions ranging from Llama 2 Kundu and Barbosa (2024); Oketch et al. (2025) to Llama 3.3 Jordan et al. (2025) and parameter sizes spanning from 8B Xu et al. (2025); Yoo et al. (2025), to 405B Oketch et al. (2025). Other open-weighted models included

Mistral families ($n = 5$, 11%, *e.g.,* Lee et al. (2024); Stahl et al. (2024)) and Gemma families ($n = 3$, 7%, Eltanbouly et al. (2025); Jordan et al. (2025); Ormerod and Kwako (2024)). Notably, the Deepseek models (versions V3 and R1) were used in only two studies—namely Gao et al. (2025); Oketch et al. (2025)—suggesting that their adoption within automated argumentative essay scoring research remains limited. Figure 5 provides an overview of this distribution, distinguishing between proprietary and open models and their respective subfamilies and Figure 6 details the diversity of GPT variants adopted within individual studies, showing that two-thirds of studies relied on a single GPT version, and only a minority combined multiple ones.

3.3.2 Model size

In line with widely used capacity classifications,[11] models up to ∼10B parameters are typically categorized as "small", whereas models in the 10–70B range are often considered "medium". According to this definition, only 14 studies (30%) employ small LLMs (*e.g.*, Eltanbouly et al. (2025); Ormerod and Kwako (2024); Xu et al. (2025)), and 7 additional studies (15%) rely on medium-sized models (*e.g.*, Jordan et al. (2025); Kundu and Barbosa (2024)). Overall, the literature remains heavily skewed toward the use of large, proprietary systems, which typically have a potentially significant environmental, monetary and social cost.

3.3.3 Reasoning capabilities

A recent methodological trend is the adoption of reasoning-enhanced LLMs that are explicitly optimized for multi-step inference, chain-of-thought consistency, or tool-augmented reasoning Sun et al. (2025a), such as GPT-4o/GPT-4o mini, Claude 3.5 Sonnet, Haiku, Gemini 1.5 Flash, Deepseek-R1, and Phi-3 reasoning. Nine studies (20%) incorporate at least one LLM with reasoning capabilities. The emergence of reasoning LLMs reflects a shift toward architectures designed to enhance analytical decomposition and interpretability capabilities, aligning closely with the cognitive demands of argumentative essay scoring.

3.4 Technical approaches

The 46 studies included in this survey report a variety of technical methods to use LLMs for the automated argumentative essay scoring, which can be characterized according to six dimensions: (1) model parameterization and adaptation, (2) prompt engineering and conditioning, (3) generation control and output structuring, (4) aggregation and normalization, (5) feedback and reinforcement mechanisms, and (6) ensembling and multi-model and (7) advanced techniques (see Table 6). These characteristics reflect the different stages at which researchers can influence an LLM's behavior. This structure enables a systematic view of the diversity of technical approaches in the field while also clarifying points of convergence in current practice.

[11] https://huggingface.co/blog/jjokah/small-language-model.

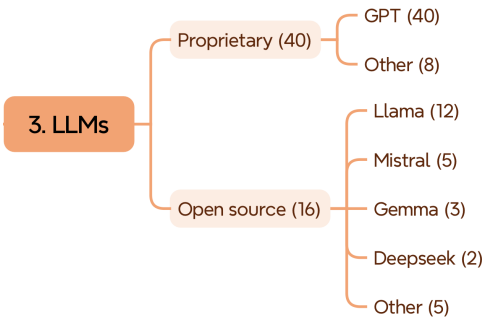

Fig. 5: Licensing models of the LLMs used in the 46 studies reviewed. Numbers in parentheses indicate the number of studies using each model.

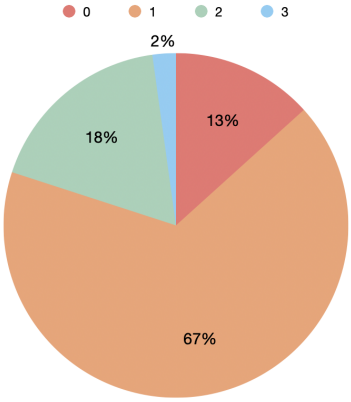

Fig. 6: Distribution of the number of GPT variants used across the 46 studies. Most studies (67%) employed a single GPT variant, while 18% used two, 2% three, and 13% did not use any GPT model. This pattern indicates a predominant reliance on a single GPT configuration rather than multi-variant comparative approaches.

3.4.1 Model parametrization and adaptation

Only a subset of works ($n = 10$, 22%) modify the model parameters beyond the base LLM. These studies rely on **partial fine-tuning**, often via parameter-efficient methods such as **LoRA adapters** Xiao et al. (2025), and occasionally combine this with **quantization** to reduce memory and deployment costs Ormerod and Kwako (2024). Fine-tuning is typically used to align model scores more closely with standardized rubrics; however, it remains minor compared with purely prompt-based approaches.

3.4.2 Prompt engineering and conditioning.

Prompt design is the dominant control mechanism of the LLM's behavior. All studies implement **rubric-based prompting**, explicitly encoding the human scoring criteria, which are different in each study, into the input prompt so that the LLM can access this information at inference time. Twenty-one studies (46%) further structure prompts through **chain-of-thought** instructions ($n = 4$, 9%) and **few-shot** or **in-context learning** ($n = 17$, 37%), thereby encouraging the model to reason step-by-step and to anchor scoring decisions in concrete examples. Contextual framing is another recurrent strategy used by 10 studies (22%): **persona conditioning** (*e.g.*, instructing the model to act as an examiner or teacher; Stahl et al. (2024)) and **additional contextual information about the assessment setting** are used to stabilize responses and

approximate human raters. Nine studies (20%) perform prompt optimization, either by **augmenting prompts** with engineered (*e.g.,* generic or prompt-specific features Eltanbouly et al. (2025) or linguistic features Hou et al. (2025); Kim (2025)) or by performing **iterative manual refinement** Yoshida (2025); Xu et al. (2025).

3.4.3 Generation control and structured outputs

Generation-level controls are mostly limited to standard decoding and sampling strategies: 10 studies (22%) tune the temperature or top-k/p to balance determinism and variability in scoring. More substantive innovation lies in structured output design. **Rationale and feedback elicitation** is adopted by 21 studies (46%): many systems explicitly request justifications, diagnostic comments, or rubric-aligned explanations Xiao et al. (2025); Jordan et al. (2025); Yoo et al. (2025). Together with **multi-trait scoring decomposition**—where the model outputs separate scores for each dimensions such as content, organization, or language Lin and Pu (2024)–and JSON based scoring output, these techniques appear in 35 studies (76%). They are used to increase the LLM's performance, alignment, and transparency.

3.4.4 Aggregation and normalization

Aggregation and normalization procedures are emerging as a relevant layer in the AAES pipeline. **Statistical aggregation** (*e.g.,* averaging across multiple runs utilizing the same prompt Kim (2025); Uyar and Büyükahıska (2025)) and **score normalization** (*e.g.,* rescaling to human score distributions; Liu et al. (2025a); Yoshida (2025)) are reported in 10 studies (22%). These methods are explicitly motivated by the need to reduce LLM output variance, mitigate prompt- or decoding-induced biases, and improve calibration against human raters.

3.4.5 Feedback and reinforcement mechanisms

Only 2 studies (4%) explore feedback-driven optimization. One study leverages **Reinforcement Learning from Human Feedback** (RLHF) to align scoring behavior with human preferences, using rater judgments as a reward signal Xu et al. (2025). The other study formulates AAES as a comparative judgment problem, using **pairwise ranking and reward modeling** to train models that choose between alternative scores or responses Cai et al. (2025). These approaches target finer-grained control and better calibration but are still limited to experimental prototypes—*i.e.,* they are not common practice.

3.4.6 Ensembling, multi-agent, and advanced hybrid strategies

Finally, 7 studies (15%) move beyond single-model pipelines. **Multi-agent evaluation and modular architectures** decompose the task into specialized roles or stages (*e.g.,* adherence, persuasiveness, organization, vocabulary, and grammar scorers Jordan et al. (2025)), whose outputs are then combined (*e.g.,* with an orchestrator Jordan et al. (2025) or a regression model Eltanbouly et al. (2025)). More **advanced and hybrid techniques** (such as feature generation and extraction Eltanbouly et al.

(2025) or fast and slow thinking framework Xiao et al. (2025)) appear in only 6 studies (13%). While heterogeneous in implementation, these studies share the goal of improving robustness, interpretability, and score reliability by orchestrating multiple models, decision stages, or training signals. Their comparatively limited adoption highlights both the methodological complexity of such systems and a promising direction for more systematic, calibrated, and auditable LLM-based AAES.

3.5 Metrics

Across the included studies, performance evaluation metrics can be clustered into four broad types: (1) reliability and agreement; (2) association and predictive validity; (3) inferential and psychometric evaluation; and (4) qualitative and textual analyses. Tables 7 and 6 provide an overview of the different performance evaluation metrics used for AAES.

3.5.1 Reliability and agreement

Twenty four studies (52%) measure rater–model consistency with **quadratic weighted kappa (QWK)** (*e.g.,* Cai et al. (2025); Lee et al. (2024); Eltanbouly et al. (2025)), including pairwise and macro-averaged variants Oketch et al. (2025). Ten studies (22%) use **intraclass correlation coefficients (ICC)** instead of QWK or to complement it (*e.g.,* Bui and Barrot (2025a); Kim (2025)), **Krippendorff's** α Gao et al. (2025); Stahl et al. (2024); Yamashita (2024); and 9 studies (20%) leverage **Cohen's/Fleiss'** κ (*e.g.,* Naismith et al. (2023)) for multi-rater scenarios. Fine-grained comparisons appear through **exact** (*e.g.,* Shermis (2025)), **adjacent** (*e.g.,* Tate et al. (2024)), or **exact-plus-adjacent** Lin and Pu (2024) agreement measures in 11 studies (24%).

3.5.2 Association and predictive validity

Linear and rank-order correspondence between human and model scores is assessed via **Pearson's** r in 15 studies (33%, *e.g.,* Kundu and Barbosa (2024); Uchida (2024) and **Spearman's** ρ in 10 studies (22%, *e.g.,* Geckin et al. (2023); Shin and Lee (2024). When scores are treated numerically, predictive performance is evaluated with regression metrics in 6 studies (13%), such as **Mean Absolute Error (MAE)** (*e.g.,* Xu et al. (2025)), **Root Mean Square Error (RMSE)** (*e.g.,* Altamimi (2023)), and (adjusted/pseudo) R^2 Mizumoto and Eguchi (2023). Under an ordinal classification framing, 5 studies report **accuracy, precision, recall,** F_1 (11%, *e.g.,* Albuquerque Da Silva et al. (2024)), and 7 studies include **confusion matrices** (15%, *e.g.,* Tang et al. (2024)). Thirty-one studies (67%) use **distributional statistics** (*e.g.,* means, variances, standard deviations, minima/maxima), and provide descriptive comparisons (*e.g.,* Jin et al. (2025); Farzi (2024); Mahdi and Alkhateeb (2025)). Four studies (9%) further quantify divergence through **standardized mean differences (SMD)** (*e.g.,* Shermis (2025)).

Table 5: Overview of the techniques used in LLM-based automated argumentative essay scoring research (see Section 3.4)

1. Model parametrization and adaptation		
Fine-tuning		Cai et al. (2025), Eltanbouly et al. (2025), Gao et al. (2025), Liu et al. (2025b), Liu et al. (2025a), Ormerod and Kwako (2024), Shin and Lee (2024), Xiao et al. (2025), Yavuz et al. (2025), Yoo et al. (2025)
2. Prompt engineering and conditioning		
Prompt structuring	– Rubric based-prompt	All studies
	– Chain-of-thought	Jordan et al. (2025), Wang et al. (2025a), Xiao et al. (2025), Xu et al. (2025)
	– Few-shots and in-context learning	Chen et al. (2024), Farzi (2024), Hou et al. (2025), Jin et al. (2025), Jordan et al. (2025), Kim (2025), Kundu and Barbosa (2024), Lin and Pu (2024), Liu et al. (2025a), Mansour et al. (2024), Naismith et al. (2023), Oketch et al. (2025), Shermis (2025), Shin and Lee (2024), Xiao et al. (2025), Yoo et al. (2025)
Contextual framing	– Persona conditioning	Hou et al. (2025), Lin and Pu (2024), Mansour et al. (2024), Shermis (2025), Stahl et al. (2024), Uchida (2024) Yeung (2025)
	– Contextual framing	Chen et al. (2025), Shin and Lee (2024), Tekin and Aydogdu (2024), Uyar and Büyükahıska (2025)
Prompt optimization	– Feature-augmented	Eltanbouly et al. (2025), Hou et al. (2025), Kim (2025), Liu et al. (2025b)
	– Manual prompt refinement	Farzi (2024), Kim (2025), Kundu and Barbosa (2024), Lee et al. (2024), Mansour et al. (2024), Naismith et al. (2023) Shin and Lee (2024), Tang et al. (2024), Uchida (2024) Xu et al. (2025), Yoshida (2025)
3. Generation control and output structuring		
Decoding and sampling strategies	– Temperature and top-k/p	Hou et al. (2025), Liu et al. (2025b) Mansour et al. (2024), Tang et al. (2024), Tate et al. (2024), Uchida (2024), Wang et al. (2025a), Xiao et al. (2025), Yavuz et al. (2025), Yeung (2025)
Structured output generation	– JSON based scoring output	Hou et al. (2025), Uchida (2024)
	– Rationale and feedback elicitation	Bouziane and Bouziane (2024), Chen et al. (2025), Albuquerque Da Silva et al. (2024), Farzi (2024), Gao et al. (2025), Hou et al. (2025) Jin et al. (2025), Jordan et al. (2025), Kim (2025), Kundu and Barbosa (2024), Lin and Pu (2024), Naismith et al. (2023), Ormerod and Kwako (2024), Shin and Lee (2024), Da Silva et al. (2025), Stahl et al. (2024), Tang et al. (2024), Arif Cem Topuz et al. (2025) Xiao et al. (2025), Xu et al. (2025), Yoo et al. (2025)
	– Multi-trait scoring decomposition	Altamimi (2023), Bouziane and Bouziane (2024), Bui and Barrot (2025b), Bui and Barrot (2025a), Chen et al. (2025), Chen et al. (2024), Da Silva et al. (2025), Albuquerque Da Silva et al. (2024), Eltanbouly et al. (2025), Farzi (2024), Jin et al. (2025), Jordan et al. (2025), Kim (2025), Kundu and Barbosa (2024), Lee et al. (2024) Lin and Pu (2024), Mahdi and Alkhateeb (2025), Mizumoto and Eguchi (2023), Naismith et al. (2023), Shermis (2025), Shin and Lee (2024), Stahl et al. (2024), Albuquerque Da Silva et al. (2024) Tang et al. (2024), Tekin and Aydogdu (2024), Arif Cem Topuz et al. (2025), Uchida (2024), Uyar and Büyükahıska (2025), Wang et al. (2025a), Xiao et al. (2025), Xu et al. (2025), Yamashita (2024), Yavuz et al. (2025), Yoo et al. (2025)
4. Aggregation and normalization methods		
Statistical aggregation		Kim (2025), Pack et al. (2024), Arif Cem Topuz et al. (2025), Uchida (2024), Uyar and Büyükahıska (2025)
Score normalization		Lee et al. (2024), Liu et al. (2025b), Liu et al. (2025a), Yeung (2025), Yoshida (2025)
5. Feedback and reinforcement mechanism		
Reinforcement Learning from Human Feedback		Xu et al. (2025)
Comparative judgment optimization	– Pairwise ranking and reward modeling	Cai et al. (2025)
6. Ensembling and multi-model strategies		
Multi-agent evaluation and modular pipelines		Cai et al. (2025), Eltanbouly et al. (2025), Hou et al. (2025), Jordan et al. (2025), Xiao et al. (2025)
7.Advanced techniques		
Advanced and hybrid techniques		Cai et al. (2025), Eltanbouly et al. (2025), Hou et al. (2025), Shermis (2025), Xiao et al. (2025), Xu et al. (2025)

3.5.3 Statistical inference and hypothesis testing

Twenty seven studies (59%) conduct parametric hypothesis testing, including **normality checks** Lin and Pu (2024); Arif Cem Topuz et al. (2025), ***t*-tests** ($n = 6$, 13%, *e.g.,* Pack et al. (2024)), **ANOVA-MANOVA** ($n = 8$, 17%, *e.g.,* Mahdi and Alkhateeb (2025)), and effect-size reporting using η^2 Arif Cem Topuz et al. (2025) or **Cohen's** d or p**-values** ($n = 22$, 48%, *e.g.,* Gao et al. (2025); Yavuz et al. (2025)) for significance assessment. Non-parametric alternatives such as **Wilcoxon signed-rank tests** are used when distributional assumptions are violated Geckin et al. (2023); Uyar and Büyükahıska (2025). Uncertainty quantification appears in the form of **bootstrap** confidence intervals Naismith et al. (2023); Yoshida (2025) and, more rarely, **information-theoretic criteria** Mizumoto and Eguchi (2023).

3.5.4 Measurement modeling and psychometrics

Ten studies (22%) integrate psychometric frameworks, including **linear or ordinal regression models** for score calibration ($n = 4$, 9%, *e.g.,* Kim (2025)), **multi-faceted Rasch measurement (MFRM;** ($n = 5$, 11%, *e.g.,* Wang et al. (2025a)) for rater–task–prompt disentanglement, and proportional bias analyses to diagnose systematic scoring deviations. Broader generalization is assessed through **leave-one-out cross-validation information criteria (LOOIC;** Mizumoto and Eguchi (2023)) or **generalizability-theory (G-theory;** Wang et al. (2025a); Gao et al. (2025)) approaches.

3.5.5 Rationale and feedback analysis

Complementing quantitative metrics, ten studies (22%, *e.g.,* Stahl et al. (2024); Yoo et al. (2025)) qualitatively examine generated rationales or feedback using **grounded or thematic content analysis**. Two studies (4%), Stahl et al. (2024); Jordan et al. (2025), use **LLMs "as judges"** approaches to evaluate the models' meta-reasoning. Finally, textual similarity between human and model explanations or summaries is occasionally assessed using **ROUGE, BLEU, or cosine similarity** on the text embeddings Xu et al. (2025).

Overall, the landscape reveals a heterogeneous but increasingly multi-layered evaluation ecosystem: traditional agreement indices remain foundational, yet several studies enrich analyses with inferential statistics, psychometric modeling, and qualitative assessments to better capture the reliability, validity, and interpretability of LLM-based AAES systems.

3.6 Analysis

Regarding the analytical dimension, we examine the inter-rate agreement between predictions and human scores, the validity and construct analysis carried out by the studies, their consideration of the FATEN principles—a set of ethical principles and practical dimensions designed to ensure that data-driven decision making is responsible, trustworthy, and beneficial for society Oliver (2019)—and the deployment and educational integration of the proposed solutions.

Table 7: Overview of the performance metrics used in LLM-based automated argumentative essay scoring research (see Section 3.5).

	Reliability and agreement	
	QWK, pairwise QWK, mQWK	Cai et al. (2025), Chen et al. (2024), Eltanbouly et al. (2025), Hou et al. (2025), Jordan et al. (2025), Kim (2025), Lee et al. (2024), Liu et al. (2025b), Liu et al. (2025a), Mansour et al. (2024), Mizumoto and Eguchi (2023), Naismith et al. (2023), Oketch et al. (2025), Ormerod and Kwako (2024), Shermis (2025), Stahl et al. (2024), Tang et al. (2024), Tate et al. (2024), Xiao et al. (2025), Xu et al. (2025), Yamashita (2024), Yeung (2025), Yoo et al. (2025), Yoshida (2025)
	Intraclass Correlation Coefficient (ICC)	Bui and Barrot (2025b), Bui and Barrot (2025a), Farzi (2024), Gao et al. (2025), Kim (2025), Pack et al. (2024), Shin and Lee (2024), Tate et al. (2024), Tekin and Aydogdu (2024), Yavuz et al. (2025)
	Krippendorff's α	Gao et al. (2025), Stahl et al. (2024), Yamashita (2024)
	Cohen's κ or Fleiss' κ for >2 raters	Farzi (2024), Gao et al. (2025), Geckin et al. (2023), Jordan et al. (2025), Mahdi and Alkhateeb (2025), Mizumoto and Eguchi (2023), Naismith et al. (2023), Shermis (2025), Tate et al. (2024)
	Exact, Adjacent (± 1 band) or Exact Plus Adjacent Agreement	Gao et al. (2025), Jin et al. (2025), Kim (2025), Lin and Pu (2024), Liu et al. (2025b), Liu et al. (2025a), Naismith et al. (2023), Shermis (2025), Shin and Lee (2024), Tate et al. (2024), Yamashita (2024)
	Association and rank-based validity	
	Pearson's r (linear association)	Bui and Barrot (2025b), Bui and Barrot (2025a), Chen et al. (2024), Farzi (2024), Kim (2025), Kundu and Barbosa (2024), Liu et al. (2025b), Liu et al. (2025a), Mahdi and Alkhateeb (2025), Oketch et al. (2025), Pack et al. (2024), Stahl et al. (2024), Arif Cem Topuz et al. (2025), Uchida (2024), Yeung (2025)
	Spearman's ρ (rank association)	Chen et al. (2025), Geckin et al. (2023), Jin et al. (2025), Lin and Pu (2024), Liu et al. (2025b), Liu et al. (2025a), Naismith et al. (2023), Oketch et al. (2025), Shin and Lee (2024), Uyar and Büyükahıska (2025)
	Predictive performance	
Regression (ordinal treated as numeric)	– Mean Absolute Error (MAE)	Oketch et al. (2025), Xu et al. (2025), Yeung (2025)
	– (Root) Mean Square Error ((R)MSE)	Altamimi (2023), Chen et al. (2024), Oketch et al. (2025), Xu et al. (2025)
	– R^2, adjusted R^2 and pseudo R^2	Mizumoto and Eguchi (2023)
Ordinal classification	– Accuracy, precision, recall, F1	Albuquerque Da Silva et al. (2024), Liu et al. (2025b), Liu et al. (2025a), Da Silva et al. (2025), Yeung (2025)
	– Confusion matrix	Jordan et al. (2025), Liu et al. (2025a), Mizumoto and Eguchi (2023), Shermis (2025), Tang et al. (2024), Yeung (2025), Yoshida (2025)
	Distributional and descriptive statistics	
	Distribution, (mean, variance, standard deviation, min, max...)	Bouziane and Bouziane (2024), Bui and Barrot (2025b), Bui and Barrot (2025a), Chen et al. (2024), Albuquerque Da Silva et al. (2024), Farzi (2024), Jin et al. (2025), Jordan et al. (2025), Kim (2025), Kundu and Barbosa (2024), Lee et al. (2024), Lin and Pu (2024), Mahdi and Alkhateeb (2025), Mizumoto and Eguchi (2023), Naismith et al. (2023), Oketch et al. (2025), Pack et al. (2024), Shermis (2025), Shin and Lee (2024), Da Silva et al. (2025), Tang et al. (2024), Tate et al. (2024), Tekin and Aydogdu (2024), Uchida (2024), Uyar and Büyükahıska (2025), Xiao et al. (2025), Xu et al. (2025), Yamashita (2024), Yavuz et al. (2025), Yoo et al. (2025), Yoshida (2025)
	Standardized Mean Difference (SMD)	Farzi (2024), Lin and Pu (2024), Liu et al. (2025b), Shermis (2025)

3.6.1 Agreement with human scores

Inter-rater agreement between the model predictions and human scores was evaluated using Quadratic Weighted Kappa (QWK) in 24 studies (52%). Table 7 synthesizes the distribution of reported QWK values across these studies following standard interpretive ranges Landis and Koch (1977).

Only 2 studies (8%) report *fair* agreement (QWK $\in [0.21, 0.40]$); 8 studies (33%) fall within the *moderate* range (QWK $\in [0.41, 0.60]$), reflecting partially reliable but inconsistent scoring performance. The largest group of studies ($n = 9$, 38%) achieve

Table 6: Overview of the performance metrics used in LLM-based automated argumentative essay scoring research (see Section 3.5). Continued.

Statistical inference and hypothesis testing			
Parametric	Assumption diagnostics	– Normality	Lin and Pu (2024), Arif Cem Topuz et al. (2025)
	Group mean comparison	– t-test (paired or independent)	Bouziane and Bouziane (2024), Chen et al. (2025), Farzi (2024), Pack et al. (2024), Arif Cem Topuz et al. (2025), Yoshida (2025)
		– ANOVA (one way or factorial)	Bouziane and Bouziane (2024), Lin and Pu (2024), Oketch et al. (2025), Tate et al. (2024), Tekin and Aydogdu (2024), Arif Cem Topuz et al. (2025), Yoo et al. (2025)
		– MANOVA	Mahdi and Alkhateeb (2025)
	Effect size	– η^2, or partial η^2 – Cohen's d and p-value	Arif Cem Topuz et al. (2025) Bui and Barrot (2025b), Bui and Barrot (2025a), Chen et al. (2025), Farzi (2024), Gao et al. (2025), Geckin et al. (2023), Jin et al. (2025), Kim (2025), Kundu and Barbosa (2024), Lin and Pu (2024), Liu et al. (2025b), Mahdi and Alkhateeb (2025), Mizumoto and Eguchi (2023), Pack et al. (2024), Shin and Lee (2024), Tate et al. (2024), Arif Cem Topuz et al. (2025), Uchida (2024), Uyar and Büyükahıska (2025), Yamashita (2024), Yavuz et al. (2025), Yoshida (2025),
Non-parametric	– Wilcoxon signed-rank		Geckin et al. (2023), Uyar and Büyükahıska (2025)
Uncertainty estimation	– Bootstrap confidence intervals – Uncertainty of Information Criterion		Naismith et al. (2023), Yoshida (2025) Mizumoto and Eguchi (2023)
Measurement modeling and psychometrics			
Model-based measurement	– Linear and ordinal regression		Kim (2025), Kundu and Barbosa (2024), Mizumoto and Eguchi (2023), Tate et al. (2024)
	– Multifaceted Rasch (MFRM)		Gao et al. (2025), Jin et al. (2025), Shin and Lee (2024), Wang et al. (2025a), Yamashita (2024)
Bias diagnostic	– Proportional bias analysis		Pack et al. (2024)
Generalization	– Leave-One Out cross validation information criterion (LOOIC)		Mizumoto and Eguchi (2023)
	– Generalizability (G-) theory frameworks		Gao et al. (2025), Wang et al. (2025a)
Rationale and feedback analysis			
Qualitative analysis	– Grounded approach, thematic content analysis		Albuquerque Da Silva et al. (2024), Gao et al. (2025), Jordan et al. (2025), Kundu and Barbosa (2024), Naismith et al. (2023), Ormerod and Kwako (2024), Da Silva et al. (2025), Stahl et al. (2024), Xu et al. (2025), Yoo et al. (2025)
	– LLM-as a judge		Jordan et al. (2025), Stahl et al. (2024)
Textual similarity	– ROUGE, BLEU, cosine similarity		Xu et al. (2025)

substantial agreement (QWK $\in [0.61, 0.80]$), demonstrating comparatively strong alignment between LLM predictions and human judgments. Finally, 5 studies (21%) report *almost perfect* agreement (QWK $\in [0.81, 1]$), though these are unevenly distributed across prompts, traits, and datasets, suggesting potential sensitivity to task design and data characteristics.

Based on the reported validity analyses, LLM-based AAES systems can reach **high levels of scoring reliability**, but performance **varies substantially across tasks, prompting conditions, and model configurations**.

.

Table 7: Ranges of Quadratic Weighted Kappa (QWK) agreement scores reported across the included studies, together with their standard interpretive categories as defined by Landis and Koch (1977)

QWK Range	Interpretation	Count	Studies
0.00 – 0.20	Poor agreement	–	–
0.21 – 0.40	Fair agreement	2 (8%)	Mansour et al. (2024), Mizumoto and Eguchi (2023)
0.41 – 0.60	Moderate agreement	8 (33%)	Hou et al. (2025), Kim (2025), Lee et al. (2024), Stahl et al. (2024), Tang et al. (2024), Tate et al. (2024), Xu et al. (2025), Yeung (2025)
0.61 – 0.80	Substantial agreement	9 (38%)	Chen et al. (2024), Eltanbouly et al. (2025), Jordan et al. (2025), Liu et al. (2025b), Shermis (2025), Xiao et al. (2025), Yamashita (2024), Yoo et al. (2025), Yoshida (2025)
0.81 – 1.00	Almost perfect agreem.	5 (21%)	Cai et al. (2025), Liu et al. (2025a), Naismith et al. (2023), Oketch et al. (2025), Ormerod and Kwako (2024)

3.6.2 Validity and construct analysis

families (*e.g.,* Oketch et al. (2025)). Eight studies (17%) report that smaller open models can achieve substantial to almost perfect agreement (*e.g.,* Yoshida (2025)), while in 6 studies (13%) the agreement appears to be moderate (*e.g.,* Kundu and Barbosa (2024)).

2. Comparison with off-the-shelf models

Ablation analyses examine the contribution of prompting and different modeling components to the performance of the AAES system. Across the 18 studies (39%) reporting **improvements over a vanilla setup**, three consistent patterns emerge: (1) **few-shot prompting** reliably improves scoring accuracy ($n = 6$, 13%, *e.g.,* Stahl et al. (2024)); (2) in two studies—Liu et al. (2025a); Yoo et al. (2025) (4%)—**fine-tuning** yields notable gains; (3) other **advanced techniques**, such as pairwise ranking or multi-agent pipeline, further enhance performance ($n = 5$, 11%, *e.g.,* Eltanbouly et al. (2025); Cai et al. (2025)).

By contrast, the **impact of rationale or feedback elicitation is mixed**: while 6 studies (13%) report performance gains (*e.g.,* Jordan et al. (2025); Kim (2025)), two (4%) do not identify a clear benefit or even report a degradation in performance Xiao et al. (2025); Yoo et al. (2025). Similarly, providing **additional information** or context in the prompt **generally improves results** ($n = 11$, 24%, *e.g.,* Mizumoto and Eguchi (2023); Mansour et al. (2024)), though 4 studies (9%) document no effect or reduced performance, often attributed to verbosity or misalignment with scoring criteria (*e.g.,* Chen et al. (2025); Yoshida (2025)).

3. Robustness and sensitivity

Robustness analyses report **varied stability profiles**. Nine studies (20%, *e.g.,* Gao et al. (2025)) find performance to be consistent across repeated runs, and 3 studies (7%) report even higher stability than that of human raters (*e.g.,* Tate et al. (2024)). Yet 6 studies (13%) document substantial run-to-run variability, revealing sensitivity to sampling randomness (*e.g.,* Geckin et al. (2023)). Temperature is fixed at zero in 8 studies (17%) to promote deterministic scoring (*e.g.,* Uchida (2024)). Additional analyses identify memorization-related behaviors—scoring fatigue Bui and Barrot (2025a) and halo effects Lee et al. (2024); Wang et al. (2025a)—and reduced performance on minority score distributions, underscoring the influence of data imbalance Bui and Barrot (2025a); Liu et al. (2025a); Xu et al. (2025).

4. Generalization and transferability

Sixteen studies (35%, *e.g.,* Kundu and Barbosa (2024); Hou et al. (2025)) perform generalization analyses and evaluate the performance of the LLMs on different datasets, *i.e.,* argumentative essay genres, prompts, topics, or scoring trait types. Only Liu et al. (2025a) explicitly tests for overfitting during fine-tuning training via proportional sampling.

3.6.3 FATEN analysis

Fourty one studies (89%) also examine LLM-based AAES systems through the **FATEN** framework lens Oliver (2019): (1) **F**airness and equity, (2) **A**ugmentation:

Table 8: Summary of agreement, validity, and construct-analytic findings reported across the reviewed studies. Green-colored font is used to indicate supportive or positive evidence; red-colored font indicates negative or contradictory evidence. Roman numerals denote areas where no clear consensus emerges, which are visualized in Figure 7 (see Sections 3.6.1, 3.6.2 and 3.6.3)

Analysis	Findings	Studies
1 Agreement with human scores (Further details in Table 7)		
QWK agreement	– [I] LLM achieves QWK >0.60	Chen et al. (2024), Eltanbouly et al. (2025), Jordan et al. (2025), Liu et al. (2025b), Shermis (2025), Xiao et al. (2025), Yamashita (2024), Yoo et al. (2025), Yoshida (2025),Cai et al. (2025), Liu et al. (2025a), Naismith et al. (2023), Oketch et al. (2025), Ormerod and Kwako (2024)
	– [I] LLM achieves QWK <0.60	Mansour et al. (2024), Mizumoto and Eguchi (2023) Hou et al. (2025), Kim (2025), Lee et al. (2024), Stahl et al. (2024), Tang et al. (2024), Tate et al. (2024), Xu et al. (2025), Yeung (2025)
2.1 Benchmark and comparative evaluation		
Against SOTA	– [II] Outperforms or similar perf. to SOTA	Eltanbouly et al. (2025), Naismith et al. (2023), Yeung (2025), Yoo et al. (2025)
	– [II] Underperforms relative to SOTA	Cai et al. (2025), Chen et al. (2024), Hou et al. (2025), Mansour et al. (2024), Ormerod and Kwako (2024), Stahl et al. (2024), Xiao et al. (2025)
Cross-LLM comparisons	– [III] GPT outperforms others LLMs	Bui and Barrot (2025a), Jin et al. (2025), Liu et al. (2025a), Pack et al. (2024), Tang et al. (2024), Tate et al. (2024)
	– [III] GPT performs comparably to other LLMs	Hou et al. (2025), Kim (2025), Lee et al. (2024), Stahl et al. (2024), Tang et al. (2024), Tate et al. (2024), Xu et al. (2025), Yeung (2025)
	– [IV] Smaller open LLM achieves QWK > 0.60	Cai et al. (2025), Eltanbouly et al. (2025), Jordan et al. (2025), Oketch et al. (2025), Ormerod and Kwako (2024), Xiao et al. (2025), Yoo et al. (2025), Yoshida (2025)
	– [IV] Smaller open LLM achieves QWK < 0.60	Hou et al. (2025), Kundu and Barbosa (2024), Lee et al. (2024), Mansour et al. (2024), Stahl et al. (2024), Xu et al. (2025)
2.2 Comparison with off-the-shelf models		
Techniques improving the perf.	– Few-shot exemplars	Chen et al. (2024), Farzi (2024), Jin et al. (2025), Kundu and Barbosa (2024), Lin and Pu (2024), Stahl et al. (2024)
	– Fine-tuning	Liu et al. (2025a), Yoo et al. (2025)
	– Other advanced methods	Cai et al. (2025), Eltanbouly et al. (2025), Hou et al. (2025), Jordan et al. (2025), Xiao et al. (2025)
Feedback/rational elicitation	– [V] Improves performance	Jordan et al. (2025), Kim (2025), Naismith et al. (2023), Stahl et al. (2024), Tang et al. (2024), Xiao et al. (2025)
	– [V] No effect or decreases performance	Xiao et al. (2025), Yoo et al. (2025)
Extra info. in prompt	– [V] Extra prompt info. helps	Eltanbouly et al. (2025), Farzi (2024) Hou et al. (2025), Kim (2025), Kundu and Barbosa (2024), Lee et al. (2024), Lin and Pu (2024), Liu et al. (2025b), Mansour et al. (2024), Mizumoto and Eguchi (2023), Tekin and Aydogdu (2024)
	– [V] No effect or decreases of the perf.	Chen et al. (2025), Mansour et al. (2024), Wang et al. (2025a), Yoshida (2025)
2.3 Robustness and sensitivity		
Sensitivity to randomness	– [VI] Consistent across repeated runs	Gao et al. (2025), Kim (2025), Liu et al. (2025a), Mizumoto and Eguchi (2023), Pack et al. (2024), Tate et al. (2024), Arif Cem Topuz et al. (2025), Xiao et al. (2025), Yavuz et al. (2025)
	– [VI] More consistent than human raters	Jin et al. (2025), Tate et al. (2024), Arif Cem Topuz et al. (2025)
	– [VI] Inconsistent across repeated runs	Bui and Barrot (2025b), Bui and Barrot (2025a), Geckin et al. (2023), Lin and Pu (2024), Pack et al. (2024), Xu et al. (2025)
Temperature variance	– Temperature fixed at 0 for scoring stability	Liu et al. (2025b), Mansour et al. (2024), Tang et al. (2024), Tate et al. (2024), Uchida (2024), Wang et al. (2025a), Xiao et al. (2025), Yavuz et al. (2025)
Memorization & Contamination	– Scoring fatigue effects	Bui and Barrot (2025a)
	– Halo effect	Lee et al. (2024), Wang et al. (2025a)
Data imbalance	– Lower performance on minority score distributions	Bui and Barrot (2025a), Liu et al. (2025a), Xu et al. (2025)
2.4 Generalization and transferability		
Type and dataset generalization	– Ability to generalize across essay type, topics and traits	Altamimi (2023), Cai et al. (2025), Eltanbouly et al. (2025), Hou et al. (2025), Kundu and Barbosa (2024), Lee et al. (2024), Mansour et al. (2024), Oketch et al. (2025), Ormerod and Kwako (2024), Shermis (2025), Stahl et al. (2024), Tang et al. (2024), Tate et al. (2024), Wang et al. (2025a), Xiao et al. (2025), Yoo et al. (2025)
Overfitting diagnostics	– Behavior under different test-set sampling	Liu et al. (2025a)

cognitive and pedagogical alignment, (3) **T**ransparency and explainability, (4) b**E**nficence and (5) **N**on-maleficence. Table 9 summarizes these analyses, which highlight dimensions of system performance that extend beyond predictive accuracy alone.

1. Fairness and equity

Fairness analyses examine score bias, distributional behavior, stylistic sensitivity, and demographic or proficiency-related disparities. Across studies, **no consistent pattern of systematic score bias** emerges: 12 studies (26%) identify a positive bias (the tendency of a scoring model to give higher grades to essays, *e.g.,* Mahdi and Alkhateeb (2025)) whereas 13 studies (28%) report a negative bias (the tendency of a scoring model to give lower grades to essays, *e.g.,* Kim (2025)). Only Farzi (2024) and Gao et al. (2025) report no significant bias in the scores provided by the LLMs.

Regarding distributional behavior, only Pack et al. (2024) reports full utilization of the scoring range; the remaining 13 studies (28%) that provide this information document **conservative scoring patterns** (*e.g.,* Da Silva et al. (2025)), suggesting a systematic compression of score variance.

A **style bias** is identified in 5 studies (11%, *e.g.,* Wang et al. (2025a)), which show that LLM scores tend to correlate with surface-level linguistic features rather than argumentative quality.

Eight studies (17%) report clear **performance sensitivity** to the **learners' proficiency in English** (*e.g.,* Yavuz et al. (2025)), while only Yamashita (2024) finds no such effect.

Demographic analyses in four studies (9%) conclude that there are no gender- or age-based differences in the performance of the LLMs (*e.g.,* Oketch et al. (2025)), while Liu et al. (2025b) report significant **sensitivity to linguistic background**, indicating that LLM-based AAES systems may disadvantage writers from certain L1 groups.

2. Cognitive and pedagogical alignment

Eleven studies (24%) evaluate whether LLM-generated feedback **aligns with educational objectives**. Ten studies (22%, *e.g.,* Gao et al. (2025)) consider such feedback to be pedagogically useful, supporting learning-oriented interpretations of AAES outputs. However, five studies (9%) report the feedback to be misaligned or not pedagogically meaningful (*e.g.,* Bouziane and Bouziane (2024), where feedback contains vocabulary that is too complex for students), highlighting limitations in the coherence, accuracy, or instructional value of generated explanations.

3. Transparency and explainability

Trait-level analyses in 16 studies (35%) (*e.g.,* Uchida (2024)) provide finer-grained insights into **how LLMs score specific dimensions**—typically stylistic and argumentative traits[12]—rather than (only) holistic outcomes. These analyses help clarify which aspects of the writing drive model judgments and at which level an essay can be improved, thereby supporting greater transparency in scoring.

[12] See Sec 3.2 for more details.

In addition, 10 studies (22%) conduct feature-based explainability analyses (*e.g.,* Yeung (2025)), examining the linguistic, argumentative, or structural cues that influence LLM scoring decisions. Across studies, models appear to rely on identifiable and interpretable feature sets, though the specific features vary substantially by model architecture and prompt design.

4. Beneficence

The beneficence dimension in the FATEN framework includes concepts such as sustainability, diversity, veracity and contribution to progress. **Only 3 studies** (7%) conduct analyses that consider the **beneficence** dimension in LLM-based AAES systems. Eltanbouly et al. (2025) and Xiao et al. (2025) examine inference efficiency, reporting on runtime, while only Oketch et al. (2025) evalutes the effectiveness. These emerging results suggest that practical deployment considerations remain underexplored in the current literature.

5. Non-maleficence

Finally, the non-maleficence dimension entails aspects related to the preservation of privacy, the reliability, security, and safety of the methods and the application of a principle of prudence to minimize potential unintended negative consequences. Although few studies mention risks related to ethical concerns Jordan et al. (2025), privacy Shermis (2025), data security Oketch et al. (2025), or model safety, none provide empirical analysis of these issues. **No study** evaluates the **privacy-preserving mechanisms, nor the security and safety** of the proposed approach, highlighting a critical gap within current LLM-based AAES research.

Table 9: Summary of the **FATEN** Oliver (2019) analytic dimensions and reported findings across studies. Green-colored font is used to indicate supportive or positive evidence; red-colored font indicates negative or contradictory evidence. Roman numerals denote areas where no clear consensus emerges, which are also depicted in Figure 7.

Analysis	Findings	Papers
3.1 Fairness and equity analysis		
Score bias	– No systematic bias	Farzi (2024), Gao et al. (2025)
	– Positive bias	Albuquerque Da Silva et al. (2024) Bouziane and Bouziane (2024), Da Silva et al. (2025), Liu et al. (2025a), Liu et al. (2025b), Mahdi and Alkhateeb (2025), Pack et al. (2024), Shermis (2025), Shin and Lee (2024), Tekin and Aydogdu (2024), Yavuz et al. (2025), Yoshida (2025)
	– Negative bias	Bui and Barrot (2025b), Bui and Barrot (2025a), Chen et al. (2025), Jin et al. (2025), Kim (2025), Kundu and Barbosa (2024), Naismith et al. (2023), Pack et al. (2024), Tang et al. (2024), Arif Cem Topuz et al. (2025), Uyar and Büyükahıska (2025), Wang et al. (2025a), Yeung (2025)
Distribution bias	–Full scoring-range utilization	Pack et al. (2024)
	– Conservative scoring behavior	Da Silva et al. (2025), Jin et al. (2025), Jordan et al. (2025), Kim (2025), Lin and Pu (2024), Naismith et al. (2023), Pack et al. (2024), Shin and Lee (2024), Tang et al. (2024), Tate et al. (2024), Wang et al. (2025a), Yamashita (2024), Yavuz et al. (2025)
Style bias	Biased by surface-level linguistic features	Bui and Barrot (2025a), Farzi (2024), Lin and Pu (2024) Naismith et al. (2023), Wang et al. (2025a)
Proficiency sensitivity	– No sensitivity to proficiency level	Yamashita (2024)
	– Sensitivity detected	Bui and Barrot (2025b), Liu et al. (2025a), Liu et al. (2025b), Mizumoto and Eguchi (2023), Tate et al. (2024), Yavuz et al. (2025), Yeung (2025), Yoo et al. (2025)
Demographic sensitivity	– No gender-based differences	Jin et al. (2025), Oketch et al. (2025), Yamashita (2024)
	– No age-related differences	Oketch et al. (2025)
	– Sensitivity to linguistic background	Liu et al. (2025b)
3.2 Augmentation: cognitive and pedagogical alignment		
Alignment with educational objectives	– [VII] Feedback judged pedagogically useful	Bouziane and Bouziane (2024), Chen et al. (2025), Da Silva et al. (2025), Albuquerque Da Silva et al. (2024), Gao et al. (2025), Kundu and Barbosa (2024), Naismith et al. (2023), Stahl et al. (2024), Xiao et al. (2025), Xu et al. (2025)
	– [VII] Feedback judged not pedagogically useful	Bouziane and Bouziane (2024), Da Silva et al. (2025), Albuquerque Da Silva et al. (2024), Kundu and Barbosa (2024), Ormerod and Kwako (2024)
3.3 Transparency and explainability		
Multi-trait scoring decomposition	– Trait-level analysis	Bouziane and Bouziane (2024), Bui and Barrot (2025a), Chen et al. (2024), Eltanbouly et al. (2025), Farzi (2024), Jordan et al. (2025), Kim (2025), Mahdi and Alkhateeb (2025), Mansour et al. (2024) Shermis (2025), Shin and Lee (2024), Tang et al. (2024), Uchida (2024) Wang et al. (2025a), Yavuz et al. (2025), Yoo et al. (2025)
Feature analysis	– Identification of influential linguistic, argumentative, or structural features	Eltanbouly et al. (2025), Kim (2025), Kundu and Barbosa (2024), Liu et al. (2025b), Mizumoto and Eguchi (2023), Uchida (2024), Xu et al. (2025), Yeung (2025), Yoo et al. (2025), Yoshida (2025)
3.4 Beneficence		
Inference efficiency	– Runtime performance	Eltanbouly et al. (2025), Xiao et al. (2025)
Cost effectiveness		Oketch et al. (2025)

Derived from Table 8 and Table 9, Figure 7 provides a visual overview of the analytic findings reported across the included studies. For each evaluated outcome, it displays the number of studies *supporting* the finding (green) versus those *contradicting* it (red). This visualization complements the narrative by highlighting areas of divergence across the analytical dimensions, including: [I] LLM QWK archivements, [II] Perfromance relative to SOTA, [III] performance of GPT in comparison to others LLMs, [IV] Smaller LLM QWK archivements, [VI] Benefit of extra prompt information, [VII] Consistency across repetitions and [VII] Pedagogical usefulness of the feedback generated by the LLM. Overall, the distribution reveals several domains in

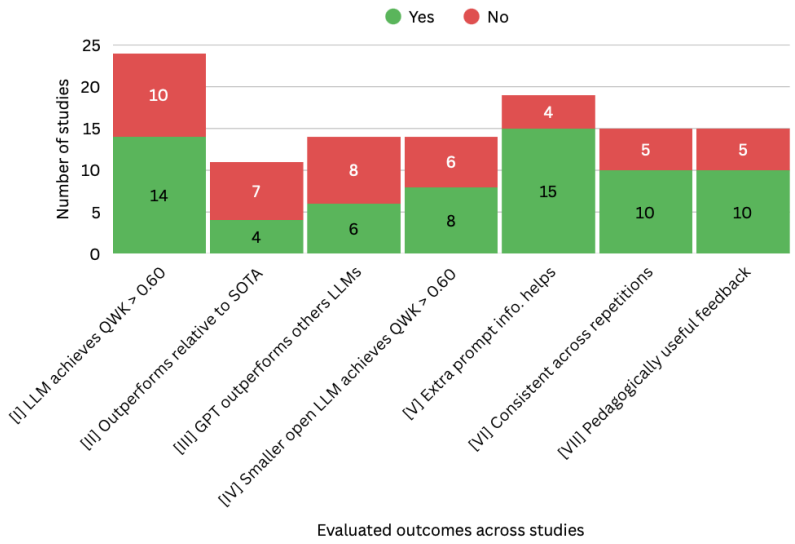

Fig. 7: Distribution of studies that support or contradict each evaluated finding. The figure highlights results from Table 8 and Table 9 for which no consensus was observed. Green bars ("Yes") depict the number of studies reporting evidence in favor of the finding, whereas red bars ("No") correspond to the number of studies presenting opposing or negative evidence.

which empirical results remain mixed, underscoring the need for more systematic and rigorous evaluation.

The data shows that while LLMs provide strong baseline performance in AAES, their consistency and reliability depend heavily on prompt design, model configuration, and dataset characteristics. The variability reported in robustness, ablation, and generalization analyses underscores the need for additional research. Furthermore, a more systematic evaluation of the FATEN dimensions, particularly educational utility, explainability, environmental impact, privacy implications, and cost, is needed to support responsible real-world adoption.

4 Discussion: Trends and open challenges

This review reveals that the AAES field is advancing rapidly in methodological innovation yet it is marked by uneven analytical depth and structural limitations. The following discussion synthesizes cross-cutting issues concerning dataset quality, LLM choices, evaluation practices, and broader methodological and ethical gaps in current AAES research.

4.1 Datasets

The first set of limitations concern the **quality and shortcomings of existing datasets** that are used to train and evaluate automated systems, including dataset

fragmentation, incomplete metadata, narrow prompt coverage, limited annotations and demographic and linguistic biases.

4.1.1 Benchmark dependence and dataset fragmentation

The field remains **disproportionately anchored to a small set of benchmarks** —namely, ASAP Hamner et al. (2012) ASAP++ Mathias and Bhattacharyya (2018), TOEFL11 Blanchard et al. (2013), ICNALE Ishikawa (2013), and ELLIPSE Crossley et al. (2023)—used in 70% of the studies included in this survey. The ASAP/ASAP++ dataset alone is used in 30% of the studies, yet it presents substantive limitations, including heterogeneous prompt genres with prompt-specific rubrics that hinder comparability, short essays written by a demographically narrow population, limited linguistic and contextual diversity, and imbalanced score distributions. Its age (ASAP was introduced in 2012, and ASAP++ in 2018) and visibility further increase the risk of data contamination and reduce its relevance to contemporary writing Sun et al. (2025b). Conversely, newly introduced datasets are numerous but typically small ($\leq$ 120 essays in 74% of cases), distinctive in design, and often inaccessible, limiting their cumulative scientific value (see Table 4).

4.1.2 Incomplete and inconsistent metadata

As shown in Tables 3 and 4, **essential metadata from existing datasets are often missing or inconsistently reported**. Gaps include a lack of prompt descriptions, rater expertise, trait definitions, demographic variables, and L1 information. In 38% of datasets, L1 is absent or only partially specified. Dataset availability is similarly uneven: while two-thirds of established benchmarks are freely accessible, most newly introduced datasets (85%) are not. Several corpora are presented as accessible but remain effectively unavailable, hindering transparency, reproducibility, and systematic comparison.

4.1.3 Narrow prompt coverage and weak generalization

Prompt coverage is limited: 44% of datasets rely on a single prompt, and two-thirds include five or fewer prompts. Such narrow topical ranges likely induce strong prompt-specific biases, yet these biases are rarely analyzed. Only a handful of datasets (*e.g.,* DReSS Yoo et al. (2025), PERSUADE 2.0 Crossley et al. (2024), ELLIPSE Crossley et al. (2023)) provide broader coverage that supports cross-topic or cross-genre evaluation. **No studies examine topic sensitivity, value alignment, or potential leakage from widely used benchmarks, leaving generalizability largely untested**.

4.1.4 Rater configuration and annotation practices

Rater practices vary widely across datasets. One-rater scoring remains common (34%), and multi-rater designs are exceptions rather than norms. Rater training Gao et al. (2025); Jin et al. (2025), calibration Pack et al. (2024); Lin and Pu (2024), disagreement resolution procedures Chen et al. (2024), or averaging Xu et al. (2025) improve

reliability estimates but obscure underlying variance in scorer judgment. Although many datasets provide trait definitions, **annotation schemes are not standardized across corpora, creating inconsistencies in how argumentative quality is operationalized**.

4.1.5 Demographic and linguistic population biases

Dataset populations are narrow, dominated by undergraduate writers (38%) and high-school or entrance-exam test takers (31%). Despite the inclusion criteria, no dataset targets adult or lifelong learners, limiting ecological validity. **L1 distributions are similarly skewed**: Asian L1s dominate (41%), especially Chinese L1 writers (34%), with limited representation of other linguistic groups. Lower-frequency L1s (*e.g.,* Arabic, Spanish, French) appear only sporadically. These imbalances are rarely acknowledged, and few studies assess their implications for LLM fairness or bias Liu et al. (2025b).

4.1.6 Heterogeneous dataset size

While a few corpora contain several thousand essays (*e.g.*, Crossley et al. (2023)), the majority of newly introduced datasets include fewer than 120 samples, **restricting statistical power, reliability of model evaluation, and exacerbating risks of overfitting and poor generalization**, especially when only one essay prompt is used.

4.1.7 Implications and open research directions

Given the previously described limitations, current datasets constrain the validity, fairness, and robustness of AAES evaluations. **High-quality** (*e.g.*, several prompts and raters), **balanced, large enough and fully documented datasets are needed not only for fine-tuning but also for credible benchmarking.**

From the perspective of datasets, advancing AAES research, would require: (1) new benchmark datasets with broad prompt coverage, fine-grained argumentative traits, and complete metadata (prompts, demographics, L1, rater expertise); (2) transparent and accessible corpora to enable reproducibility and cross-study comparability; (3) datasets with more diverse writer populations, extending beyond school-age learners and currently overrepresented Asian L1 groups; (4) systematic generalization analyses, including topic robustness, L1 fairness, and assessment neutrality; and improvements on (5) data quality, including de-duplication, balanced sampling, and unified annotation schemes.

Recent initiatives such as DReSS Yoo et al. (2025) illustrate progress toward larger, more diverse, and rubric-aligned datasets, but substantial gaps remain. Without sustained investment in high-quality and fully documented datasets, advances in LLM-based argumentative essay scoring will remain constrained by the limitations in existing corpora.

4.2 Traits

The second set of limitations concerns the definition and use of essay traits in current AAES datasets.

4.2.1 The limits of current datasets' trait coverage

In 3.2, we found that current AAES trait-based datasets disproportionately emphasize the *Rhetorical effectiveness* within the Argument Quality (AQ) framework. This AQ category evaluates "the persuasive power of an author's argument towards a target audience" Romberg et al. (2025) along five subcategories: the clarity of the argument's language, the ordering and structure of the argument, the author's credibility and appropriateness of style (similar to Artistotle's *ethos* Braet (1992)), and whether the authors purposefully appeals to emotion to persuade the audience (similar to *pathos* Braet (1992)). **The dominance of *Rhetorical effectiveness* reflects the historical focus of AAES research on language proficiency rather than argument quality assessment.**

Many of the survey datasets originate from educational contexts concerned primarily with second-language learning (19 of the surveyed datasets or 63% include essays that were written by non-native English speakers) where essay traits like grammar, vocabulary, fluency, and organisation constitute central learning objectives. These dimensions are also more amenable to modeling than higher-level reasoning constructs, as they can be operationalized using linguistic features (*e.g.*, TAACO metrics for cohesion Kyle and McNamara (2015)). In contrast, **high-level traits such as argument quality** (encapsulated by *Logical cogency* and *Dialectic reasonableness*) **or engagement with alternative viewpoints** (*Interactivity*, one of the *Deliberative norms* subcategories) **remain difficult to annotate reliably and to model computationally** Ke and Ng (2019), which likely contributes to their limited representation in existing resources. The sub-category coverage analysis in 3.2.4 further confirms this as it reveals how datasets tend to focus on linguistically tractable traits rather than on broader rhetorical or deliberative constructs.

4.2.2 The gap between AQ and AAES

Our analysis further highlights **a systematic gap between contemporary theories of AQ and what current datasets evaluate in practice**. While frameworks such as AQ emphasize reasoning, deliberation, and dialectical engagement, the datasets examined here largely privilege rhetorical form and linguistic competence. This mismatch has important implications for both dataset design and model evaluation: systems trained on existing benchmarks may perform well on stylistic or grammatical criteria while remaining fundamentally limited in argumentative assessment. Without broader coverage of *Logical cogency*, *Deliberative norms* and *Dialectic reasonableness*, progress toward truly argumentative automated essay scoring remains constrained.

4.2.3 Implications and open research directions

We hope that the proposed **AQ-based classification of essay traits, depicted in Figure 8, will provide a practical tool for comparing AAES datasets** which are otherwise different in terminology and rubric design. By mapping trait names onto one shared taxonomy, the proposed framework can help inform researchers of the scope, conceptual coverage, and evaluation focus of different datasets and help them decide which dataset to work with. It can also be used to diagnose systematic blind spots in the field (both at the category and subcategory-levels) and offer guidance for the development of future trait-based scoring rubrics and datasets that better reflect the multidimensional nature of argumentative assessment.

4.3 LLM choices

The second set of limitations arise from the choices of LLMs used to perform the automated scoring of argumentative essays.

4.3.1 Dominance of proprietary models and reproducibility concerns

The strong dominance of GPT-based systems, used by 87% of the studies, observed in this literature review **does not necessarily reflect a performance advantage but rather the convenience of readily accessible interfaces**. A substantial proportion of studies interacted with GPT directly through the ChatGPT platform—without using the API—facilitating rapid experimentation but limiting reproducibility and methodological transparency Jin et al. (2025). This reliance raises concerns both regarding traceability, versioning, and experimental control, particularly in high-stakes assessment settings, and with respect to ethical aspects, such as privacy and sustainability.

4.3.2 The potential of smaller, open models

Small to medium open LLMs, most commonly from the Llama family, were used by 35% of the studies Kundu and Barbosa (2024); Oketch et al. (2025); Jordan et al. (2025). When supported by robust methodological design, including rubric-guided prompts, exemplar-based few-shot conditioning, or structured analytical scaffolds, these smaller models frequently achieved a performance comparable to proprietary GPT variants (*e.g.,* Ormerod and Kwako (2024); Hou et al. (2025)). These results suggest that **scoring quality hinges more on the applied techniques than on raw model scale**. However, the variability across results echoes the broader pattern noted throughout this review: prompt engineering, decomposition strategies, and evaluation design exert substantial influence on the reliability of the results.

4.3.3 The rise of reasoning-optimized models

A notable observed trend is the growing use of reasoning-optimized LLMs, including GPT-4o Oketch et al. (2025), 4o-mini Uyar and Büyükahıska (2025), Claude 3.5 Sonnet Wang et al. (2025a), Gemini 1.5 Flash Yoshida (2025), and Deepseek-R1 Gao

et al. (2025). Although only one-fifth of studies adopted such models, they consistently reported strong outcomes, with high correlations, substantial-to-near-perfect QWK, and overall improved inter-rater consistency. These findings support emerging evidence that **reasoning-enhancement can benefit tasks requiring analytical decomposition and nuanced judgment** Sui et al. (2025).

4.3.4 Availability of locally run deployment ecosystems

Recent advances in deployment ecosystems (*e.g.,* Ollama[13], lightweight GPU runtimes) have made open LLMs increasingly accessible. These tools enable **local, offline inference without programming skills, directly addressing concerns surrounding privacy, data security, and institutional compliance** Marcondes et al. (2025). Such advantages are particularly relevant in educational contexts, where student writing might contain sensitive personal data and where transparency and auditability are central to the responsible deployment and assessment of AI tools Al-Zahrani (2024); Kwapisz et al. (2024).

4.3.5 Implications and open research directions

Despite these developments, open-weighted adoption remains limited relative to GPT-centric practices. This imbalance has implications for sustainability, privacy, transparency and reproducibility, especially as proprietary systems provide minimal insight into training data, alignment processes, or inherent biases. As open, small LLMs continue to improve, the field would benefit from **broader methodological diversification and more systematic evaluations of transparent**, replicable models that are locally run and hence are potentially more **privacy-preserving and with lower carbon footprint**. Strengthening these practices will be essential for establishing robust, equitable, and trustworthy AAES systems.

4.4 Evaluation

From the perspective of evaluation practices, we identify two important areas for discussion and future work: metrics and evaluation methodologies.

4.4.1 Metrics

The metric landscape is diverse: studies combine reliability indices (*e.g.,* Jordan et al. (2025)), correlation-based validity measures (*e.g.,* Kundu and Barbosa (2024)), error-based predictive metrics (*e.g.,* Oketch et al. (2025)), psychometric modeling (*e.g.,* Mizumoto and Eguchi (2023)), and qualitative evaluations (*e.g.,* Naismith et al. (2023)). However, several recurring methodological limitations warrant attention. First, **the dominant metric—quadratic weighted kappa (QWK)—is frequently misinterpreted**. QWK captures agreement structure rather than absolute accuracy; it is highly sensitive to score-range differences, distributional imbalance, rubric heterogeneity, and sample size Doewes and Pechenizkiy (2021). These properties produce well-documented "kappa paradoxes" in AAES settings, *i.e.,* high observed

[13] https://ollama.com

agreement but low QWK, or the reverse, depending on score distributions Doewes et al. (2023).

Second, **adjacent-agreement metrics** (*e.g.,* "±1 band") are **not inherently comparable** across scoring scales Williamson et al. (2012). A one-band tolerance on a 0–4 scale is not equivalent to one band on a 0–60 scale, yet several studies treat these ranges as interchangeable, obscuring cross-study comparison.

Third, some studies evaluate generated rationales or summaries using text-similarity metrics such as ROUGE or BLEU. These metrics are **poorly aligned with argumentative quality and rubric-defined constructs**. Hence, they should not be used to assess reasoning adequacy Favero et al. (2024). Construct-valid alternatives include evaluations of argument structure, warrant sufficiency, and rubric-aligned quality dimensions.

Finally, **uncertainty is rarely quantified**. Only a small number of studies report confidence intervals (*e.g.,* Mizumoto and Eguchi (2023)), despite the high variance of LLM outputs across runs, prompts, temperatures, and stochastic sampling. AAES evaluations would benefit from bootstrap confidence intervals, calibration assessments, and distributional uncertainty estimates to avoid overconfident or unstable performance claims He et al. (2025).

4.4.2 Evaluation methodology

Across the literature, two parallel methodological approaches to evaluate the AAES systems emerge. A first set of studies **prioritizes technical optimization and reports QWK as the main evaluation result**, frequently layering increasingly complex techniques, such as data augmentation Yoo et al. (2025), reinforcement learning Xu et al. (2025), multi-agent architectures Xiao et al. (2025), or highly specialised prompting pipelines Yoshida (2025); Farzi (2024). A second cluster of studies adopts **simple prompting baselines yet contributes with rich, human-centric analyses of the results** Kundu and Barbosa (2024), offering detailed examinations of scoring behavior Ormerod and Kwako (2024), construct coverage Mizumoto and Eguchi (2023), and pedagogical implications Bouziane and Bouziane (2024) without aiming to advance state-of-the-art performance or make technical contributions. This bifurcation renders the field **simultaneously methodologically innovative and analytically uneven**.

Across studies, LLM behavior is sensitive to prompt order Stahl et al. (2024), contextual leakage, and chat-session contamination—which occurs when multiple essays are scored sequentially within the same conversation Bui and Barrot (2025a). However, very few studies (3%) enforce strict separation between training data, exemplars, and evaluation samples to mitigate these issues, raising concerns about inadvertent memorization and inflated performance.

4.4.3 Implications and open research directions

These findings illustrate a maturing yet uneven evaluation ecosystem. While the field benefits from methodological diversity, greater rigor is needed in metric selection, interpretation of agreement scores, construct-aligned evaluation, and uncertainty quantification. Establishing consistent psychometric and reporting standards would

substantially enhance comparability, validity, and transparency across LLM-based AAES research. Finally, future work should adopt controlled **prompt isolation, randomized orderings, and independent single-turn evaluation protocols** to avoid cross-sample leakage and provide a more rigorous evaluation.

4.5 Technical approaches

From a technical perspective, there are several technical challenges and open directions for research that could be investigated.

4.5.1 Prompt engineering dominates but lacks methodological rigor

Most studies rely on prompt-based improvements, including few-shot examples Farzi (2024), rubric insertion Kim (2025), feature augmentation Eltanbouly et al. (2025), persona conditioning Stahl et al. (2024), or contextual framing Hou et al. (2025). Yet, several technical challenges remain, including the fact that rubric inclusion is universal, but rarely tested for its **marginal effect**; few-shot exemplars introduce **data dependence, with little analysis of selection strategies**; complex prompting artifacts (*e.g.,* chain-of-thought) are **rarely validated through ablations**, despite their centrality to published results; and JSON or schema-constrained outputs are used Hou et al. (2025) but **not studied as stabilizing mechanisms**.

4.5.2 Parameterization and fine-tuning

Fine-tuning consistently yields substantial gains Cai et al. (2025), yet only one study Liu et al. (2025a) evaluates the optimal data size and risk of overfitting when fine-tuning LLMs to automatically perform the argumentative essay scoring. Furthermore, no study has reported an analysis of the generalization capabilities of the LLMs across prompts, tasks, and populations. Existing tools[14] now allow no-code fine-tuning, increasing accessibility but also posing the **risk of poorly validated and unsafe models** (*e.g.,* Zhao et al. (2025)).

4.5.3 Implications and open research directions

Future research should test systematic prompt optimization, constrained decoding, and automatic example selection (*e.g.,* cosine similarity–based retrieval Xiao et al. (2025)), which remain unexplored despite their relevance.

Regarding fine-tuning, future work should **quantify sample efficiency, evaluate transferability, and systematically test fine-tuned models for memorization, fairness, and robustness**.

Furthermore, AAES research has not yet well leveraged many mature techniques from NLP, machine learning, and psychometrics, including: automated prompt search and optimization Marvin et al. (2023); constrained decoding or enforced schema outputs Geng et al. (2025); reinforcement learning with human or AI feedback (RLHF/RLAIF) Lee et al. (2023); comparative judgment models and probabilistic

[14] https://poe.com/FineTuning-Setup or https://platform.openai.com/docs/guides/supervised-fine-tuning

ranking Ameli et al. (2024); self-critique or self-consistency mechanisms Madaan et al. (2023); heterogeneous ensembles or multi-agent role specialization Wang et al. (2024); neuro-symbolic or rule-augmented systems integrating discourse constraints Yang and Zhao (2025); and meta-learning for adaptive scoring across domains and populations Chen and Li (2024).

Investigating the application of these approaches to AAES could strengthen the technical soundness, validity, stability, and transparency of the methods.

4.6 Human-centric aspects

Beyond technical performance, the use of LLMs in education calls for a careful consideration of the FATEN dimensions to ensure that there won't be unintended negative consequences of their deployment. However, no study in our literature review reported a comprehensive evaluation of these human-centric aspects.

4.6.1 Persistent biases

Halo effects, stylistic bias, and superficial linguistic cues continue to influence LLM scoring Wang et al. (2025a); Farzi (2024). Models often reward surface-level fluency while overlooking deeper argumentative structure—mirroring concerns that grammatical errors (common among non-native writers) unfairly depress scores Yavuz et al. (2025); Yeung (2025). Evidence shows clear performance differences across L1 groups, but only one study Liu et al. (2025b) systematically analyzes them. Research must incorporate **bias diagnostics, L1- and proficiency-aware calibration, and evaluation pipelines sensitive to multilingual learner populations**.

4.6.2 Rationale and feedback generation

Feedback-based prompting was used in 46% of the studies, yet few (33%) assess whether feedback improves alignment with human scores, reduces variance, or enhances construct validity Tang et al. (2024); Jordan et al. (2025). The generated feedback is often pedagogically useful Bouziane and Bouziane (2024); Chen et al. (2025) yet incomplete (missing argumentative dimensions), too complex for learners, or uncalibrated Da Silva et al. (2025); Kundu and Barbosa (2024). Crucially, no studies in this survey analyze the students' reception or use of LLM-generated scores and feedback in the context of argumentative essays. **A research agenda is needed around pedagogical validity, human–LLM interaction, and feedback quality judgments.**

4.6.3 Variability

Studies report substantial run-to-run variability in model outputs Geckin et al. (2023); Xu et al. (2025), which is undesirable. Averaging across multiple runs or using normalization and aggregation can reduce variance Kim (2025), but only a minority of studies (22%) adopt these practices. Temperature is often fixed at zero Uchida (2024) but this does not eliminate the stochasticity in the models. Future work should report **multi-run averaged results, confidence intervals, and calibration analyses rather than single deterministic outputs**.

4.6.4 Human oversight

Many studies treat AAES as a purely predictive task, neglecting classroom context, teacher workflows, and the need for continuous human oversight. Given risks of hallucination, mis-scoring, or non-compliance (*e.g.,* models refusing to output scores), **systems must incorporate interpretability, guardrails, monitoring, and alignment with educational policy guidelines**. Deployment studies should examine usability, cost, accessibility, and the impact on teachers and learners, not only performance metrics.

4.6.5 Privacy

Despite the centrality of student data protection in educational assessment Favero et al. (2025), none of the reviewed studies conducted an empirical privacy analysis or implemented privacy-preserving mechanisms, confirming a critical gap in AAES research. Because argumentative essays could contain sensitive personal, cultural, and social content, sharing these data with a proprietary LLM raises concerns about data retention and potential privacy leaks Kwapisz et al. (2024). The widespread reliance on closed GPT-family models exacerbates these risks, particularly when inference is performed through consumer platforms rather than institutionally governed environments. **Future AAES systems must therefore incorporate explicit privacy-by-design principles**, including local deployment of open models (for example, with Ollama [15]) or transparent data policies. Without such measures, large-scale deployment in educational settings remains ethically and legally precarious.

4.6.6 Environmental impact

Environmental sustainability receives almost no attention within current AAES research: only two studies Eltanbouly et al. (2025); Xiao et al. (2025) report runtime, and none quantify energy consumption or carbon cost associated with LLM inference or fine-tuning. This absence is increasingly problematic given the heavy reliance on large proprietary models and multi-call scoring pipelines (*e.g.,* few-shot, ensemble, or multi-agent systems), all of which inflate computational cost. Especially as educational systems are meant to scale, the cumulative environmental footprint of automated scoring would be clearly non-negligible, particularly in high-volume testing contexts. Future work should evaluate the trade-offs between model size, reasoning depth, accuracy, and energy consumption. A promising direction consists of adopting small, locally deployable open-weighted models. Benchmarks that explicitly report energy metrics also need to be considered. **Integrating sustainability considerations into AAES design would align the field with broader responsible-AI commitments and institutional climate goals**.

4.6.7 Implications and open research directions

The human-centric gaps identified in the studies included in this survey point to a research agenda centered on responsible AAES development rather than purely

[15] https://ollama.com

performance-oriented optimization. First, **future systems should embed FATEN-guided auditing pipelines** Oliver (2019). Second, **theoretical and psychometric foundations must be strengthened**: construct-valid rubrics, multi-rater datasets, and transparent scoring models are needed for credible deployment. Finally, the emergence of powerful small LLMs Cai et al. (2025); Eltanbouly et al. (2025) creates an opportunity to **shift towards reproducible, privacy-preserving, lower-carbon scoring ecosystems through local models**. Overall, **interdisciplinary collaboration**, bridging AI, psychometrics, and educational outcomes, is a prerequisite to designing AAES systems that are not only accurate but also equitable, trustworthy, and aligned with classroom practice.

5 Conclusion

This scoping and critical review examined 46 studies published between 2022 and 2025 on Large Language Model–based automated argumentative essay scoring (AAES), revealing a rapidly expanding yet methodologically uneven field. LLMs have catalyzed notable progress in assessing argumentative writing, particularly through trait and rubric prompting, structured scoring outputs, and reasoning capabilities. However, the evidence shows that current systems do not yet offer a psychometrically robust, equitable, or pedagogically grounded alternative to human evaluation. Across the literature, dataset fragmentation emerges as a structural barrier: the 29 corpora identified vary widely in scale, availability, prompt diversity, rater configurations, and learner demographics. Trait analyses further demonstrate that widely used benchmarks tend to privilege rhetorical and linguistic surface features while neglecting deeper argumentative constructs such as logical cogency, evidential strength, dialectical engagement, and deliberative norms. As a result, LLM-based scorers are frequently aligned with stylistic fluency rather than substantive reasoning, raising concerns about construct under-representation and the validity of inferences drawn from model-generated scores.

Methodologically, the field is dominated by proprietary LLMs, especially the GPT family, while open-source and fine-tuned models remain comparatively underexplored. Most studies rely on prompt-based scoring pipelines, with limited adoption of multi-agent reasoning systems, parameter-efficient fine-tuning, or psychometric calibration. Although several studies report substantial agreement with human raters, ablation and robustness analyses consistently highlight sensitivity to prompt design, sampling randomness, score imbalance, and English proficiency. These findings question the stability and generalizability of current AAES LLM methods, especially in multilingual contexts. The FATEN analysis underscores additional challenges. Fairness concerns persist, few studies examine the safety, environmental cost, transparency, privacy and ethical implications of extensive proprietary LLM use. Rationales or feedback generated by LLMs show promise for transparency and instructional utility, but their pedagogical alignment can be inconsistent. Overall, responsible implementations and analysis are not yet sufficient, and existing systems fall short of widely accepted standards for educational assessment.

These findings point to a set of priorities for future research. First, the field needs construct-valid, publicly available datasets with clearly defined argumentative traits that capture the full spectrum of reasoning quality. Second, progress requires theoretically grounded scoring rubrics aligned with argumentation research rather than ad hoc linguistic categories. Third, evaluation practices—particularly those related to psychometrics, generalizability, fairness auditing, and robustness testing—must be standardized. Fourth, methodological innovation should expand beyond prompt engineering to include fine-tuning, multi-agent architectures, and hybrid models that integrate explicit reasoning frameworks. Finally, researchers and practitioners must embed the FATEN principles throughout the design and deployment pipeline. As LLM capabilities continue to evolve, the central question is no longer whether these models can score essays, but whether they can do so in ways that reflect the complexity of human reasoning and the norms of responsible educational assessment. Addressing this challenge requires a shift from performance-centric experimentation toward deeper theoretical, psychometric, and ethical foundations. Only then can LLM-based AAES mature into a reliable, interpretable, and epistemically sound component of writing assessment ecosystems.

6 Funding declaration

L.F. and N.O. have been partially supported by a nominal grant received at the ELLIS Unit Alicante Foundation from the Regional Government of Valencia in Spain (Resolución de la Conselleria de Innovación, Industria, Comercio y Turismo, Dirección General de Innovación). L.F. has also been partially funded by a grant from the Banc Sabadell Foundation. G.G is supported by the ALTA Institute which is supported by Cambridge University Press & Assessment.

References

Van den Akker OR, Peters GY, Bakker C, et al (2025) Generalized systematic review registration form. MetaArXiv, https://doi.org/10.31222/osf.io/g5fj, URL https://doi.org/g5fj

Al-Zahrani AM (2024) Unveiling the shadows: Beyond the hype of AI in education. Heliyon 10(9). https://doi.org/10.1016/j.heliyon.2024.e30696, URL https://www.cell.com/heliyon/abstract/S2405-8440(24)06727-6

Albuquerque Da Silva D, Eduardo De Mello C, Garcia A (2024) Analysis of the Effectiveness of Large Language Models in Assessing Argumentative Writing and Generating Feedback:. In: Proceedings of the 16th International Conference on Agents and Artificial Intelligence. SCITEPRESS - Science and Technology Publications, pp 573–582, https://doi.org/10.5220/0012466600003636, URL https://www.scitepress.org/DigitalLibrary/Link.aspx?doi=10.5220/0012466600003636

Altamimi AB (2023) Effectiveness of ChatGPT in Essay Autograding. In: 2023 International Conference on Computing, Electronics & Communications Engineering (iCCECE). IEEE, pp 102–106, https://doi.org/10.1109/iCCECE59400.2023.10238541, URL https://ieeexplore.ieee.org/document/10238541/

Ameli S, Zhuang S, Stoica I, et al (2024) A statistical framework for ranking llm-based chatbots. arXiv preprint arXiv:241218407

Arif Cem Topuz, Mine Yıldız, Elif Taşlıbeyaz, et al (2025) Is generative AI ready to replace human raters in scoring EFL writing? Comparison of human and automated essay evaluation. Educational Technology & Society 28(3). https://doi.org/10.30191/ETS.202507_28(3).SP04

Blanchard D, Tetreault J, Higgins D, et al (2013) TOEFL11: A CORPUS OF NON-NATIVE ENGLISH. ETS Research Report Series 2013(2). https://doi.org/10.1002/j.2333-8504.2013.tb02331.x, URL https://onlinelibrary.wiley.com/doi/10.1002/j.2333-8504.2013.tb02331.x

Bouziane K, Bouziane A (2024) AI versus human effectiveness in essay evaluation. Discover Education 3(1):201. https://doi.org/10.1007/s44217-024-00320-6, URL https://link.springer.com/10.1007/s44217-024-00320-6

Braet AC (1992) Ethos, pathos and logos in aristotle's rhetoric: A re-examination. Argumentation 6(3):307–320. https://doi.org/10.1007/bf00154696

Bui NM, Barrot J (2025a) Using generative artificial intelligence as an automated essay scoring tool: A comparative study. Innovation in Language Learning and Teaching pp 1–16. https://doi.org/10.1080/17501229.2025.2521003, URL https://www.tandfonline.com/doi/full/10.1080/17501229.2025.2521003

Bui NM, Barrot JS (2025b) ChatGPT as an automated essay scoring tool in the writing classrooms: How it compares with human scoring. Education and Information Technologies 30(2):2041–2058. https://doi.org/10.1007/s10639-024-12891-w, URL https://link.springer.com/10.1007/s10639-024-12891-w

Cai Y, Liang K, Lee S, et al (2025) Rank-Then-Score: Enhancing Large Language Models for Automated Essay Scoring. https://doi.org/10.48550/arXiv.2504.05736, URL http://arxiv.org/abs/2504.05736, 2504.05736

Cambridge University Press & Assessment (2023) B2 First Handbook for teachers for exams. URL https://www.cambridgeenglish.org/images/167791-b2-first-handbook.pdf

Cardwell R, LaFlair GT, Settles B (2022) Duolingo english test: technical manual. Duolingo Reseach Report https://englishtest duolingo com/research

Chen S, Lan Y, Yuan Z (2024) A Multi-task Automated Assessment System for Essay Scoring. In: Olney AM, Chounta IA, Liu Z, et al (eds) Artificial Intelligence in Education, vol 14830. Springer Nature Switzerland, p 276–283, https://doi.org/10.1007/978-3-031-64299-9_22, URL https://link.springer.com/10.1007/978-3-031-64299-9_22

Chen X, Zhou Z, Prado M (2025) ChatGPT-3.5 as an automatic scoring system and feedback provider in IELTS exams. International Journal of Assessment Tools in Education 12(1):62–77. https://doi.org/10.21449/ijate.1496193, URL http://dergipark.org.tr/en/doi/10.21449/ijate.1496193

Chen Y, Li X (2024) Plaes: Prompt-generalized and level-aware learning framework for cross-prompt automated essay scoring. In: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pp 12775–12786

Crossley S, Tian Y, Baffour P, et al (2023) The english language learner insight, proficiency and skills evaluation (ellipse) corpus. International Journal of Learner Corpus Research 9(2):248–269

Crossley S, Tian Y, Baffour P, et al (2024) A large-scale corpus for assessing written argumentation: Persuade 2.0. Assessing Writing 61:100865

Da Silva DCA, De Mello CE, Garcia ACB (2025) Exploring the Role of Large Language Models in Evaluating Argumentative Writing in Military School Education. In: Rocha AP, Steels L, Van Den Herik J (eds) Agents and Artificial Intelligence, vol 15592. Springer Nature Switzerland, p 287–301, https://doi.org/10.1007/978-3-031-87330-0_15, URL https://link.springer.com/10.1007/978-3-031-87330-0_15

Doewes A, Pechenizkiy M (2021) On the limitations of human-computer agreement in automated essay scoring. International educational data mining society

Doewes A, Kurdhi N, Saxena A (2023) Evaluating quadratic weighted kappa as the standard performance metric for automated essay scoring. In: 16th International Conference on Educational Data Mining, EDM 2023, International Educational Data Mining Society (IEDMS), pp 103–113

Dubey A, Jauhri A, Pandey A, et al (2024) The llama 3 herd of models. arXiv e-prints pp arXiv–2407

ElMassry AM, Zaki N, AlSheikh N, et al (2025) A systematic review of pretrained models in automated essay scoring. IEEE Access

Eltanbouly S, Albatarni S, Elsayed T (2025) TRATES: Trait-Specific Rubric-Assisted Cross-Prompt Essay Scoring. https://doi.org/10.48550/arXiv.2505.14577, URL http://arxiv.org/abs/2505.14577, 2505.14577

Emirtekin E (2025) Large Language Model-Powered Automated Assessment: A Systematic Review. Applied Sciences 15(10):5683. https://doi.org/10.3390/app15105683, URL https://www.mdpi.com/2076-3417/15/10/5683

Farzi R (2024) Calibrating Generative AI for Second Language Writing Assessment: Combining Statistical Validation with Prompt Design. Assessment and Practice in Educational Sciences 2(4):1–12. URL https://www.journalapes.com/index.php/apes/article/view/91

Favero L, Pérez-Ortiz JA, Käser T, et al (2024) Enhancing critical thinking in education by means of a socratic chatbot. In: International Workshop on AI in Education and Educational Research, Springer, pp 17–32

Favero L, Pérez-Ortiz JA, Käser T, et al (2025) Do ai tutors empower or enslave learners? toward a critical use of ai in education. arXiv preprint arXiv:250706878

Gao H, Hashim H, Md Yunus M (2025) Assessing the reliability and relevance of DeepSeek in EFL writing evaluation: A generalizability theory approach. Language Testing in Asia 15(1):33. https://doi.org/10.1186/s40468-025-00369-6, URL https://languagetestingasia.springeropen.com/articles/10.1186/s40468-025-00369-6

Geckin V, Kiziltas E, Cinar C (2023) Assessing second-language academic writing: AI vs. Human raters. Journal of Educational Technology and Online Learning 6(4):1096–1108. https://doi.org/10.31681/jetol.1336599, URL http://dergipark.org.tr/en/doi/10.31681/jetol.1336599

Geng S, Cooper H, Moskal M, et al (2025) Generating structured outputs from language models: Benchmark and studies. arXiv e-prints pp arXiv–2501

Gu J, Jiang X, Shi Z, et al (2024) A Survey on LLM-as-a-Judge. https://doi.org/10.48550/arXiv.2411.15594

Habernal I, Wachsmuth H, Gurevych I, et al (2018) The argument reasoning comprehension task: Identification and reconstruction of implicit warrants. In: Walker M, Ji H, Stent A (eds) Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). Association for Computational Linguistics, New Orleans, Louisiana, pp 1930–1940, https://doi.org/10.18653/v1/N18-1175, URL https://aclanthology.org/N18-1175/

Hamner B, Morgan J, lynnvandev, et al (2012) The hewlett foundation: Automated essay scoring. https://kaggle.com/competitions/asap-aes, kaggle Competition

He J, Yu L, Li C, et al (2025) Survey of uncertainty estimation in large language models-sources, methods, applications, and challenge

Hou ZJ, Ciuba A, Li XL (2025) Improve LLM-based Automatic Essay Scoring with Linguistic Features. https://doi.org/10.48550/arXiv.2502.09497, URL http://arxiv.org/abs/2502.09497, 2502.09497

Huang Y, Palermo C, Liu R, et al (2025) An early review of generative language models in automated writing evaluation: Advancements, challenges, and future directions for automated essay scoring and feedback generation. Chinese/English Journal of Educational Measurement and Evaluation 6(2):5

Huot B (1990) Reliability, Validity, and Holistic Scoring: What We Know and What We Need to Know. College Composition and Communication 41(2):201–213. https://doi.org/10.2307/358160, URL https://www.jstor.org/stable/358160, publisher: National Council of Teachers of English

Ishikawa S (2013) The icnale and sophisticated contrastive interlanguage analysis of asian learners of english. Learner corpus studies in Asia and the world 1:91–118

Ishikawa S (2020) Aim of the icnale gra project: Global collaboration to collect ratings of asian learners' l2 english essays and speeches from an elf perspective. Learner Corpus Studies in Asia and the World 5:121–144

Jin R, Zhao M, Niu C, et al (2025) Evaluating the performance of ChatGPT and Claude in automated writing scoring: Insights from the Many-facet Rasch model. Education and Information Technologies https://doi.org/10.1007/s10639-025-13774-4, URL https://link.springer.com/10.1007/s10639-025-13774-4

Jordan J, Yin X, Fabros M, et al (2025) MAGIC: Multi-Agent Argumentation and Grammar Integrated Critiquer. https://doi.org/10.48550/arXiv.2506.13037, URL http://arxiv.org/abs/2506.13037, 2506.13037

Ke Z, Ng V (2019) Automated essay scoring: A survey of the state of the art. In: IJCAI, pp 6300–6308

Kim Y (2025) Automated Essay Scoring With GPT -4 for a Local Placement Test: Investigating Prompting Strategies, Intra-Rater Reliability, and Alignment With Human Scores. TESOL Quarterly p tesq.3405. https://doi.org/10.1002/tesq.3405, URL https://onlinelibrary.wiley.com/doi/10.1002/tesq.3405

Koroteev MV (2021) Bert: a review of applications in natural language processing and understanding. arXiv preprint arXiv:210311943

Küçük D, Can F (2020) Stance detection: A survey. ACM Comput Surv 53(1). https://doi.org/10.1145/3369026, URL https://doi.org/10.1145/3369026

Kundu A, Barbosa D (2024) Are Large Language Models Good Essay Graders? https://doi.org/10.48550/arXiv.2409.13120, URL http://arxiv.org/abs/2409.13120, 2409.13120

Kwapisz MB, Kohli A, Rajivan P (2024) Privacy Concerns of Student Data Shared with Instructors in an Online Learning Management System. In: Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, CHI '24, pp 1–16, https://doi.org/10.1145/3613904.3642914, URL https://dl.acm.org/doi/10.1145/3613904.3642914

Kyle K, McNamara D (2015) The tool for the automatic analysis of text cohesion (taaco): Automatic assessment of local, global, and text cohesion. Behavior research methods 48. https://doi.org/10.3758/s13428-015-0651-7

Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. biometrics pp 159–174

Lawrence J, Reed C (2019) Argument mining: A survey. Computational Linguistics 45(4):765–818. https://doi.org/10.1162/coli_a_00364, URL https://aclanthology.org/J19-4006/

Lee H, Phatale S, Mansoor H, et al (2023) Rlaif vs. rlhf: Scaling reinforcement learning from human feedback with ai feedback, 2024. URL https://arxiv.org/abs/2309.00267

Lee S, Cai Y, Meng D, et al (2024) Unleashing Large Language Models' Proficiency in Zero-shot Essay Scoring. https://doi.org/10.48550/arXiv.2404.04941, URL http://arxiv.org/abs/2404.04941, 2404.04941

Lin D, Pu X (2024) Effects of Prompts and Time on the Automated Scoring of English Argumentative Essays by ChatGPT 4. In: Proceedings of the 2024 International Symposium on Artificial Intelligence for Education. ACM, pp 302–312, https://doi.org/10.1145/3700297.3700350, URL https://dl.acm.org/doi/10.1145/3700297.

3700350

Liu A, Feng B, Xue B, et al (2024) Deepseek-v3 technical report. arXiv preprint arXiv:241219437

Liu Y, Lu X, Qi H (2025a) Comparing GPT-based approaches in automated writing evaluation. Assessing Writing 66:100961. https://doi.org/10.1016/j.asw.2025.100961, URL https://linkinghub.elsevier.com/retrieve/pii/S1075293525000480

Liu Y, Qi H, Lu X (2025b) Enhancing GPT-based automated essay scoring: The impact of fine-tuning and linguistic complexity measures. Computer Assisted Language Learning pp 1–20. https://doi.org/10.1080/09588221.2025.2518430, URL https://www.tandfonline.com/doi/full/10.1080/09588221.2025.2518430

Madaan A, Tandon N, Gupta P, et al (2023) Self-refine: Iterative refinement with self-feedback. Advances in Neural Information Processing Systems 36:46534–46594

Mahdi HS, Alkhateeb A (2025) Revolutionising Essay Evaluation: A Cutting-Edge Rubric for AI-Assisted Writing. International Journal of Computer-Assisted Language Learning and Teaching 15(1):1–19. https://doi.org/10.4018/IJCALLT.368226, URL https://services.igi-global.com/resolvedoi/resolve.aspx?doi=10.4018/IJCALLT.368226

Mansour W, Albatarni S, Eltanbouly S, et al (2024) Can Large Language Models Automatically Score Proficiency of Written Essays? URL http://arxiv.org/abs/2403.06149, 2403.06149

Marcondes FS, Gala A, Magalhães R, et al (2025) Using ollama. In: Natural Language Analytics with Generative Large-Language Models: A Practical Approach with Ollama and Open-Source LLMs. Springer, p 23–35

Marvin G, Hellen N, Jjingo D, et al (2023) Prompt engineering in large language models. In: International conference on data intelligence and cognitive informatics, Springer, pp 387–402

Mathias S, Bhattacharyya P (2018) Asap++: Enriching the asap automated essay grading dataset with essay attribute scores. In: Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)

Mizumoto A, Eguchi M (2023) Exploring the potential of using an AI language model for automated essay scoring. Research Methods in Applied Linguistics 2(2):100050. https://doi.org/10.1016/j.rmal.2023.100050, URL https://linkinghub.elsevier.com/retrieve/pii/S2772766123000101

Naismith B, Mulcaire P, Burstein J (2023) Automated evaluation of written discourse coherence using GPT-4. In: Proceedings of the 18th Workshop on Innovative Use of

NLP for Building Educational Applications (BEA 2023). Association for Computational Linguistics, pp 394–403, https://doi.org/10.18653/v1/2023.bea-1.32, URL https://aclanthology.org/2023.bea-1.32

Oketch K, Lalor JP, Yang Y, et al (2025) Bridging the LLM Accessibility Divide? Performance, Fairness, and Cost of Closed versus Open LLMs for Automated Essay Scoring

Oliver N (2019) Governance in the era of data-driven decision-making algorithms. Women Shaping Global Economic Governance 171

Olson CB, Woodworth K, Arshan N, et al (2020) The pathway to academic success: Scaling up a text-based analytical writing intervention for latinos and english learners in secondary school. Journal of Educational Psychology 112(4):701

OpenAI (2023) Gpt-3.5: Openai's large language model. https://platform.openai.com/docs/models/gpt-3-5, available at https://platform.openai.com

OpenAI (2024) Gpt-4 technical report. https://arxiv.org/abs/2303.08774, arXiv:2303.08774

OpenAI (2025) Gpt-5 "deep research" (deep search) variant. https://openai.com/index/introducing-gpt-5/, part of the GPT-5 model family; see "Introducing GPT-5" and system card

Ormerod CM, Kwako A (2024) Automated Text Scoring in the Age of Generative AI for the GPU-poor. https://doi.org/10.48550/arXiv.2407.01873, URL http://arxiv.org/abs/2407.01873, 2407.01873

Pack A, Barrett A, Escalante J (2024) Large language models and automated essay scoring of English language learner writing: Insights into validity and reliability. Computers and Education: Artificial Intelligence 6:100234. https://doi.org/10.1016/j.caeai.2024.100234, URL https://linkinghub.elsevier.com/retrieve/pii/S2666920X24000353

Page MJ, McKenzie JE, Bossuyt PM, et al (2021) The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. BMJ p n71. https://doi.org/10.1136/bmj.n71, URL https://www.bmj.com/lookup/doi/10.1136/bmj.n71

Rinott R, Dankin L, Alzate Perez C, et al (2015) Show me your evidence - an automatic method for context dependent evidence detection. In: Màrquez L, Callison-Burch C, Su J (eds) Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Lisbon, Portugal, pp 440–450, https://doi.org/10.18653/v1/D15-1050, URL https://aclanthology.org/D15-1050/

Romberg J, Maurer M, Wachsmuth H, et al (2025) Towards a perspectivist turn in argument quality assessment. In: Chiruzzo L, Ritter A, Wang L (eds) Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). Association for Computational Linguistics, Albuquerque, New Mexico, pp 7458–7485, https://doi.org/10.18653/v1/2025.naacl-long.382, URL https://aclanthology.org/2025.naacl-long.382/

Shermis MD (2025) Using ChatGPT to score essays and short-form constructed responses. Assessing Writing 66:100988. https://doi.org/10.1016/j.asw.2025.100988, URL https://linkinghub.elsevier.com/retrieve/pii/S1075293525000753

Shin D, Lee JH (2024) Exploratory study on the potential of ChatGPT as a rater of second language writing. Education and Information Technologies 29(18):24735–24757. https://doi.org/10.1007/s10639-024-12817-6, URL https://link.springer.com/10.1007/s10639-024-12817-6

Stab C, Gurevych I (2017) Parsing argumentation structures in persuasive essays. Computational Linguistics 43(3):619–659. https://doi.org/10.1162/COLI_a_00295, URL https://aclanthology.org/J17-3005/

Stahl M, Biermann L, Nehring A, et al (2024) Exploring LLM Prompting Strategies for Joint Essay Scoring and Feedback Generation

Sui Y, Chuang YN, Wang G, et al (2025) Stop overthinking: A survey on efficient reasoning for large language models. arXiv preprint arXiv:250316419

Sun J, Song T, Peng W, et al (2025a) A survey of automated essay scoring: Challenges, advances, and future. Neurocomputing p 130916

Sun J, Song T, Peng W, et al (2025b) A survey of automated essay scoring: Challenges, advances, and future. Neurocomputing 650:130916. https://doi.org/10.1016/j.neucom.2025.130916, URL https://linkinghub.elsevier.com/retrieve/pii/S0925231225015887

Tang X, Chen H, Lin D, et al (2024) Harnessing LLMs for multi-dimensional writing assessment: Reliability and alignment with human judgments. Heliyon 10(14):e34262. https://doi.org/10.1016/j.heliyon.2024.e34262, URL https://linkinghub.elsevier.com/retrieve/pii/S2405844024102939

Tate TP, Steiss J, Bailey D, et al (2024) Can AI provide useful holistic essay scoring? Computers and Education: Artificial Intelligence 7:100255. https://doi.org/10.1016/j.caeai.2024.100255, URL https://linkinghub.elsevier.com/retrieve/pii/S2666920X24000584

Tekin S, Aydogdu S (2024) Automated Assessment of Students' Critical Writing Skills with Chatgpt. https://doi.org/10.2139/ssrn.4826249, URL https://www.ssrn.com/

abstract=4826249

Uchida S (2024) Evaluating the Accuracy of ChatGPT in Assessing Writing and Speaking: A Verification Study Using ICNALE GRA. https://doi.org/10.24546/0100487710, URL https://doi.org/10.24546/0100487710

Uyar AC, Büyükahıska D (2025) Artificial intelligence as an automated essay scoring tool: A focus on ChatGPT. International Journal of Assessment Tools in Education 12(1):20–32. https://doi.org/10.21449/ijate.1517994, URL http://dergipark.org.tr/en/doi/10.21449/ijate.1517994

Wachsmuth H, Naderi N, Hou Y, et al (2017) Computational argumentation quality assessment in natural language. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, pp 176–187

Wang J, Wang J, Athiwaratkun B, et al (2024) Mixture-of-agents enhances large language model capabilities. arXiv preprint arXiv:240604692

Wang Y, Huang J, Du L, et al (2025a) Evaluating large language models as raters in large-scale writing assessments: A psychometric framework for reliability and validity. Computers and Education: Artificial Intelligence 9:100481. https://doi.org/10.1016/j.caeai.2025.100481, URL https://linkinghub.elsevier.com/retrieve/pii/S2666920X25001213

Wang Y, Zhang T, Yao L, et al (2025b) A scoping review of empirical studies on generative artificial intelligence in language education. Innovation in Language Learning and Teaching pp 1–28

Williamson DM, Xi X, Breyer FJ (2012) A framework for evaluation and use of automated scoring. Educational measurement: issues and practice 31(1):2–13

Xiao C, Ma W, Song Q, et al (2025) Human-AI Collaborative Essay Scoring: A Dual-Process Framework with LLMs. In: Proceedings of the 15th International Learning Analytics and Knowledge Conference. ACM, pp 293–305, https://doi.org/10.1145/3706468.3706507, URL https://dl.acm.org/doi/10.1145/3706468.3706507

Xu W, Mahmud R, Hoo WL (2024) A systematic literature review: Are automated essay scoring systems competent in real-life education scenarios? IEEE Access 12:77639–77657

Xu W, Kassim MSS, Hoo WL, et al (2025) Explainable AI for education: Enhancing essay scoring via rubric-aligned chain-of-thought prompting. International Journal of Modern Physics C p 2542013. https://doi.org/10.1142/S0129183125420136, URL https://www.worldscientific.com/doi/10.1142/S0129183125420136

Yamashita T (2024) An application of many-facet Rasch measurement to evaluate automated essay scoring: A case of ChatGPT-4.0. Research Methods in Applied Linguistics 3(3):100133. https://doi.org/10.1016/j.rmal.2024.100133, URL https://linkinghub.elsevier.com/retrieve/pii/S2772766124000399

Yang L, Zhao S (2025) Argus: A neuro-symbolic system integrating gnns and llms for actionable feedback on english argumentative writing. Systems 13(12):1079

Yannakoudakis H, Briscoe T, Medlock B (2011) A new dataset and method for automatically grading esol texts. In: Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies, pp 180–189

Yavuz F, Celik O, Yavas Celik G (2025) Utilizing large language models for EFL essay grading: An examination of reliability and validity in rubric-based assessments. British Journal of Educational Technology 56(1):150–166. https://doi.org/10.1111/bjet.13494, URL https://bera-journals.onlinelibrary.wiley.com/doi/10.1111/bjet.13494

Yeung S (2025) A comparative study of rule-based, machine learning and large language model approaches in automated writing evaluation (AWE). In: Proceedings of the 15th International Learning Analytics and Knowledge Conference. ACM, pp 984–991, https://doi.org/10.1145/3706468.3706566, URL https://dl.acm.org/doi/10.1145/3706468.3706566

Yildiz Durak H, Onan A (2025) A systematic review of ai-based feedback in educational settings. Journal of Computational Social Science 8(4):96

Yoo H, Han J, Ahn SY, et al (2025) DREsS: Dataset for Rubric-based Essay Scoring on EFL Writing. https://doi.org/10.48550/arXiv.2402.16733, URL http://arxiv.org/abs/2402.16733, 2402.16733

Yoshida L (2025) Do We Need a Detailed Rubric for Automated Essay Scoring Using Large Language Models? In: Cristea AI, Walker E, Lu Y, et al (eds) Artificial Intelligence in Education, vol 15882. Springer Nature Switzerland, p 60–67, https://doi.org/10.1007/978-3-031-98465-5_8, URL https://link.springer.com/10.1007/978-3-031-98465-5_8

Yu Y, Si X, Hu C, et al (2019) A review of recurrent neural networks: Lstm cells and network architectures. Neural computation 31(7):1235–1270

Zhao W, Hu Y, Deng Y, et al (2025) Beware of your po! measuring and mitigating ai safety risks in role-play fine-tuning of llms. arXiv preprint arXiv:250220968

Zheng L, Chiang WL, Sheng Y, et al (2023) Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. https://doi.org/10.48550/arXiv.2306.05685, URL http://arxiv.org/abs/2306.05685, arXiv:2306.05685 [cs]

A Glossary of definitions and abbreviations

Table 10 containing a comprehensive glossary of definitions and abbreviations is provided to support clarity and consistency throughout the manuscript.

B Further details the searching strategy

B.1 Search details

The literature search was conducted across multiple academic databases and repositories, accessed either programmatically through their APIs or manually via their dedicated web interfaces. The selection of databases aimed to ensure comprehensive coverage of both computer science and educational research domains relevant to automated argumentative essay scoring with large language models (LLMs). The specific access methods and search strategies are detailed below:

- arXiv – Accessed programmatically through the official API.
- ERIC (Education Resources Information Center) – Accessed through the official platform: https://eric.ed.gov/.
- PubMed – Queried through both the web interface (https://pubmed.ncbi.nlm.nih.gov/advanced/) and the API.
- SpringerLink – Accessed through the official platform: https://www.springer.com/gp.
- ACM Digital Library – Queried via the ACM Cross-API.
- Web of Science – Accessed through institutional credentials provided by the university library.
- ScienceDirect – Accessed through the official platform:https://www.sciencedirect.com/. Due to query limitations (maximum of eight Boolean expressions), the following search string was used:

  ```
  ‘‘LLM’’ OR ‘‘GPT’’ OR ‘‘language models’’)
  AND (‘‘essay scoring’’ OR ‘‘feedback’’ OR ‘‘AES’’)
  AND ‘‘argumentative’’}
  ```

- Google Scholar – Accessed through the API, retrieving the top 500 results per publication year based on relevance ranking.

Note that studies published in ACL Anthology proceedings were covered by the Google Scholar search.

B.2 Screening details

A two-stage screening process was implemented to efficiently manage the review workload while maintaining accuracy in study selection:

Table 10: Key terms and definitions used in LLM-based automated argumentative essay scoring.

1. Assessment Concepts	
AAES (Automated Argumentative Essay Scoring)	The use of computational methods to assign scores to argumentative essays.
Argumentative essay	A genre of writing that advances a claim and supports it through reasoning and evidence.
Construct	The theoretical skill or ability that a scoring system is intended to measure.
Construct validity	The extent to which a score reflects the intended construct rather than unrelated features.
L1/L2 writer	An L1 writer uses their first language; an L2 writer writes in an additional learned language.
Prompt	The task instruction or essay question that elicits a student's written response.
Rubric	A structured scoring guide that defines performance levels for each trait.
Trait	A specific dimension of writing quality (e.g., coherence, evidence use) scored independently.
2. LLM Methods and Technical Approaches	
Fine-tuning	Training an LLM on task-specific data to specialize or adapt its behavior.
Large Language Model (LLM)	A neural network trained on large text corpora to generate and analyze language.
Multi-agent approaches	Systems where multiple LLMs interact (e.g., evaluator, critic, refiner) to improve scoring or reasoning.
Reasoning-oriented LLMs	Models explicitly optimized for structured reasoning, often via multi-step or chain-of-thought processes.
Rubric-based prompting	Prompting that embeds rubric descriptors to structure and constrain evaluation.
Structured output formats	Constrained formats (e.g., JSON, numeric scales) ensuring predictable, machine-readable outputs.
Zero-shot/Few-shot prompting	Approaches using no examples (zero-shot) or a few examples (few-shot) to guide LLM responses.
3. Evaluation and Psychometrics	
Human–model agreement	The degree to which automated scores align with human rater scores.
Non-determinism	Variability in LLM outputs resulting from probabilistic sampling.
PRISMA	A standardized framework for transparent reporting of scoping and systematic reviews.
Psychometric modeling	Statistical approaches (e.g., Rasch, IRT) for validating and interpreting assessment scores.
Reliability	The consistency of scoring across raters, prompts, or model instances.
Score distribution shift	A mismatch between predicted score distributions and human-provided score distributions.
4. Ethics, Fairness and Transparency	
Fairness/Bias in AAES	Systematic score differences associated with irrelevant demographic or linguistic attributes.
Transparency / Interpretability	The extent to which system decisions and mechanisms can be inspected or explained.

1. **Title and abstract screening** Records were first screened based on titles and abstracts to exclude clearly irrelevant studies through focused relevance judgments.
2. **Full-Text screening** Remaining studies were assessed in full to confirm eligibility against predefined inclusion and exclusion criteria.

This two-step approach, standard in systematic and scoping reviews, enabled iterative filtering without prematurely discarding potentially relevant research.

Several assurance procedures were introduced to support internal validity:

- A random subset of $10 - -15\%$ of both included and excluded studies was re-screened after a two-week interval to assess intra-rater reliability.
- All screening decisions, at the full-text stage, were documented in a centralized log for transparency and auditability. Ambiguous cases were retained for full-text review rather than excluded at earlier stages.

B.3 Extraction details

To ensure a comprehensive and reproducible synthesis of the literature, the following entities will be systematically extracted from each included source using a structured coding framework. The aim is to enable both quantitative and qualitative meta-synthesis across diverse study designs, methodologies, and reporting standards:

1. **Study metadata**
   - Title
   - Abstract
   - Author(s)
   - Year of publication
   - Institutional affiliation(s)
   - Country of origin
   - Venue (journal or conference)

   These metadata support bibliometric profiling, risk of bias assessment, and exploration of publication patterns over time.
2. **Study context and objectives**
   - Educational level (e.g., secondary, undergraduate, graduate)
   - Language of instruction and writing
   - Target assessment context
   - Study aims and research questions

   Capturing contextual variables facilitates an understanding of the scope, generalizability, and educational relevance of the findings.
3. **Methodological characteristics**
   - Study design (e.g., experimental, quasi-experimental, benchmarking study, case study)
   - Data sources (e.g., public datasets, proprietary student essays)

   This information enables quality appraisal and supports subgroup analysis across study types.

4. **LLM specific technical information**
   - Type of model used
   - Fine-tuning, prompt engineering approach, or other techniques
   - Training or inference settings (e.g., temperature, max tokens, model size)
   - Model access modality (e.g., API, open, proprietary)

   These technical descriptors are essential to understanding system behavior, performance variation, and replicability.
5. **Target tasks and scoring dimensions**
   - Nature of assessment (e.g., holistic scoring, argument structure detection, claim–premise classification, stance detection)
   - Scoring scale or rubric used (e.g., 6-point rubric, argument quality dimensions)
   - Within-prompt or cross-prompting
   - Feedback generation (e.g., formative, summative, content-based, structure-based)
   - Use of external rubrics or benchmarks (e.g., ETS, AWE standards)

   This categorization supports thematic clustering of systems by pedagogical intent and task complexity.
6. **Outcome measures**
   - Evaluation metrics (e.g., accuracy, macro-F1, quadratic weighted kappa, BLEU, ROUGE)
   - Human–LLM agreement statistics
   - Reported effect sizes (e.g., Cohen's d, Pearson's r)
   - Qualitative findings (e.g., student perceptions, educator feedback, thematic analysis)
   - System limitations or error types reported

   Extracting both statistical and qualitative outcome data ensures multi-dimensional synthesis of system performance and educational impact.
7. **Risk of bias and trustworthiness indicators**
   - Validation approach (e.g., cross-validation, holdout set)
   - Rater calibration and inter-rater reliability
   - Transparency and explainability of the system
   - Bias mitigation strategies (e.g., demographic fairness audits)
   - Alignment with pedagogical or ethical frameworks

   This dimension supports the critical appraisal of each study's contribution to responsible AI deployment in education.

To assess internal consistency and reduce potential bias, a self-reliability check was implemented: after a two-week interval, the extractor re-reviewed a random 15% subset of the extracted studies. Agreement between the two rounds was assessed qualitatively and verified through consistency in categorical coding and numerical entries. Any discrepancies were resolved through cross-checking with the original sources and updated extraction notes.

C Trait taxonomy

Table 11: Taxonomy of Argument Quality (AQ). Source: Romberg et al. (2025, Table 5).

Category	Description
Logical cogency	An argument is cogent if it has acceptable premises that are relevant to its conclusion and that are sufficient to draw the conclusion.
Local acceptability	A premise of an argument is acceptable if it is rationally worthy of being believed to be true.
Local relevance	A premise of an argument is relevant if it contributes to the acceptance or rejection of the argument's conclusion.
Local sufficiency	An argument's premises are sufficient if, together, they give enough support to make it rational to draw its conclusion.
Rhetorical effectiveness	Argumentation is effective if it persuades the target audience of (or corroborates agreement with) the author's stance on the issue.
Clarity	Argumentation has a clear style if it uses correct and widely unambiguous language as well as if it avoids unnecessary complexity and deviation from the issue.
Credibility	Argumentation creates credibility if it conveys arguments and similar in a way that makes the author worthy of credence.
Appropriateness	Argumentation has an appropriate style if the used language supports the creation of credibility and emotions as well as if it is proportional to the issue.
Emotional appeal	Argumentation makes a successful emotional appeal if it creates emotions in a way that makes the target audience more open to the author's arguments.
Arrangement	Argumentation is arranged properly if it presents the issue, the arguments, and its conclusion in the right order.
Dialectical reasonableness	Argumentation is reasonable if it contributes to the issue's resolution in a sufficient way that is acceptable to the target audience.
Global acceptability	Argumentation is acceptable if the target audience accepts both the consideration of the stated arguments for the issue and the way they are stated.
Global relevance	Argumentation is relevant if it contributes to the issue's resolution, *i.e.*, if it states arguments or other information that help to arrive at an ultimate conclusion.
Global sufficiency	Argumentation is sufficient if it adequately rebuts those counterarguments to it that can be anticipated.
Deliberative norms	Argumentation adheres to deliberative norms if it promotes a respectful and inclusive exchange of rational or alternative forms of argument, with the aim of reaching mutual understanding.
Rationality	Deliberation is rational if it is centered on arguments that are supported by solid evidence (either through facts that can be verified or through a shared understanding of moral or normative behavior), arguments and further information that are put forward in the discourse are relevant to the topic, and an informed ground for discussion is built (*e.g.*, through providing an information base in the beginning of the discussion, or information requests by participants to make the discourse more informed). With respect to the dimensions of argumentation quality, the focus is on normatively well-reasoned arguments and not on how good these are perceived by the target audience.
Interactivity	Deliberation is interactive if the participants actively engage with each other by exchanging arguments in a way where they listen to the other participants, understand their perspective, and relate to it in a substantive way (*e.g.*, by valuing, critiquing, or countering other's arguments, or question asking).
Equality	Deliberation is equal if all participants (irrespective of their background) have the same opportunity to participate by putting forward their own arguments and responding to other's claims. This dimension of deliberation quality tackles inclusiveness and accessibility.
Civility	Deliberation is civil if the participants show respect to the other participants by recognizing them as equal actors in the discourse and acknowledging the value of opposing claims. Respectful interaction is regarded as a prerequisite for participants to be convincable by other opinions and to reach a consensus decision in the sense of deliberation.

Category	Description
Common good reference	Deliberation is oriented towards the common good if arguments are justified by promoting the well-being of the community as a whole rather than serving the interests of narrow interest groups. What exactly is considered the common good can include different goals, such as achieving the best outcome for the greatest number of people or prioritizing the needs of the most disadvantaged members of society. The joint focus on a common good is regarded as a basis for participants with diverse interests to be able to convince each other.
Constructiveness	Deliberation is constructive if it contributes to finding a consensus decision for the issue of discussion through actions like proposing new solutions, searching for common ground, appeals for mobilisation, or questions addressed to the community.
Alternative forms of communication	In scenarios in which not all participants are able to adhere to the rigid concept of rational argumentation based on verifiable facts, other forms of communication can provide a valuable resource for good deliberation. These include storytelling, testimonies, narratives, emotional talk, casual talk, humor, or even gossip.
Overall quality	An overarching measure of the quality of arguments.

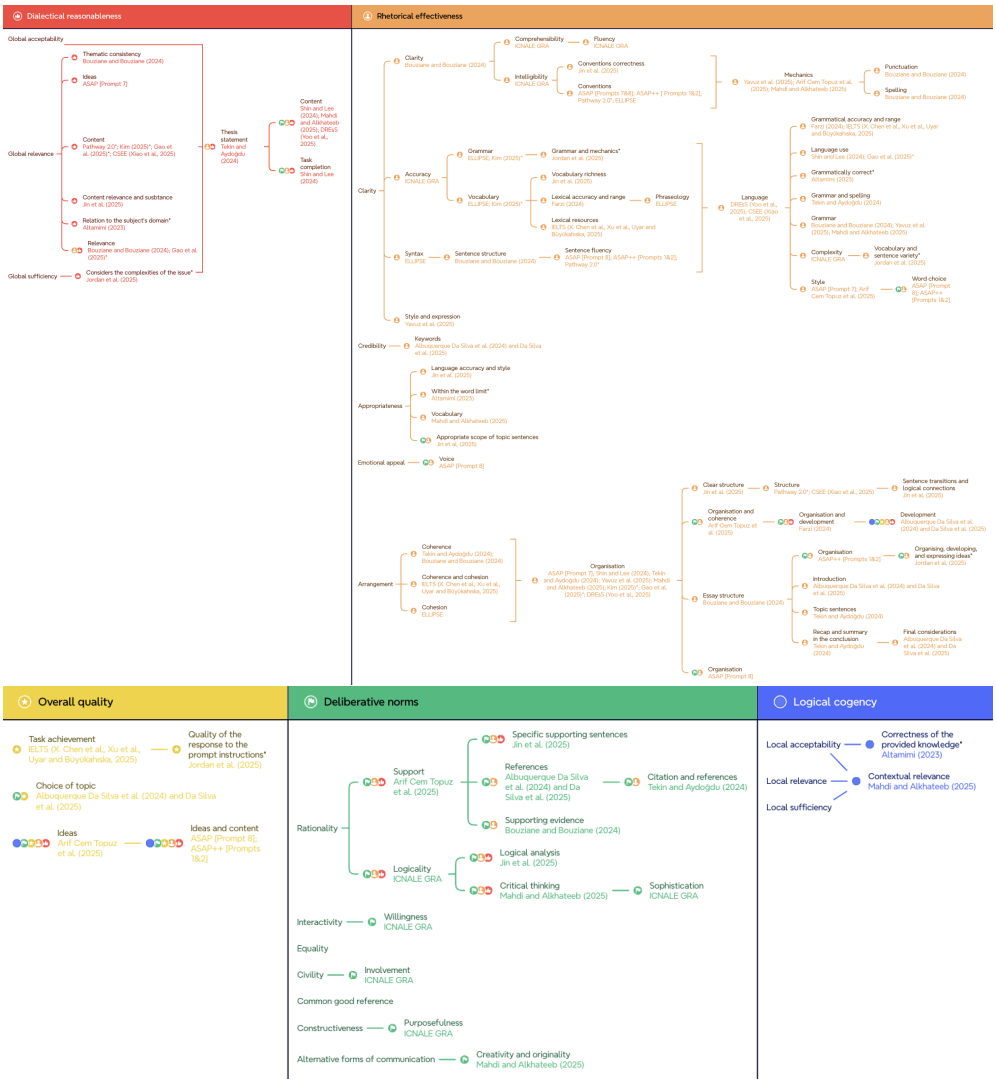

Fig. 8: Hierarchical mapping of the essay trait of the surveyed datasets in Tables 3 and 4 to the five Argument Quality (AQ) categories (dialectical reasonableness, rhetorical effectiveness, logical cogency, deliberative norms, and overall quality) introduced in Table 11. We use an asterisk symbol (*) to denote the trait names or datasets for which we did not have definitions; for these in particular, we inferred the trait meanings from the names only.