Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 00 (2026) 000–000

Procedia Computer Science

www.elsevier.com/locate/procedia

The 17th International Conference on Ambient Systems, Networks and Technologies (ANT)
April 14-16, 2026, Istanbul, Türkiye

Socratic Pitches: Agent-Led Entrepreneur–Investor Debates for Entrepreneurship Literacy

Afshin Khadangi[a,*], Amir Sartipi[a], Muriel-Larissa Frank[a], Igor Tchappi[a], Gilbert Fridgen[a]

[a] *SnT - Interdisciplinary Center for Security, Reliability and Trust, University of Luxembourg, Luxembourg, Luxembourg*

Abstract

Entrepreneurship education often treats pitching and due diligence as endpoints for assessment rather than core learning processes. To address this gap, we present **Debate-to-Learn Entrepreneurship** (D2LE), an agent-led instructional approach in which two multimodal roles—an **Entrepreneur** and an **Investor**—engage in structured debate to develop entrepreneurship literacy. The Entrepreneur agent synthesizes opportunity hypotheses into concrete artifacts. The Investor agent conducts multimodal due diligence by challenging assumptions, requesting evidence, and critiquing artifacts. A rubric-guided moderator manages turn-taking, enforces retrieval and citation when claims are made, and aligns debate moves with learning objectives including opportunity recognition, market sizing, go-to-market strategy, unit economics, risk analysis, and ethics. To reduce hallucination and promote reliability, D2LE integrates cross-agent verification, retrieval-augmented checking, and explicit citation. To support equity and inclusion, the system adapts language level, domain context, and risk tolerance to learner background and locale. We specify design choices and outline an evaluation roadmap with planned ablations to isolate the effects of retrieval and cross-examination.

Keywords: Large Language Models;LLM-based multi-agent;Entrepreneurship Education

1. Introduction

Entrepreneurship literacy—the applied capacity to recognize and evaluate opportunities, analyze markets, mobilize go-to-market strategies, reason with unit-economics under uncertainty and risk, and make ethical and sustainable decisions—maps to validated competence models [36]. Yet a persistent *teachability dilemma* holds that much of entrepreneurship involves experiential, procedural "knowing how" that traditional knowledge-transmission methods struggle to develop, contributing to mixed and context-dependent Entrepreneurship Education (EE) outcomes [5, 31]. Consistent with learning-science evidence that interactive, argumentative activity drives deeper learning than passive exposure [33], we therefore advocate approaches that shift learners from *receiving* content to *constructing and de-*

* Corresponding author.
E-mail address: afshin.khadanki@uni.lu

fending evidence-backed claims [12, 4]. At the same time, Large Language Model (LLM) have been successfully applied across a wide range of domains, including conversation between agents [42], classification [44] and art [25]. Moreover, advances in multimodal foundation models have made it practical to deploy conversational agents capable of reasoning with external tools and information sources. Multi-agent debate improves factuality and reasoning [24, 15], while retrieval and verification pipelines reduce hallucination and support citation [28, 14, 3]. Multimodal composition further enables agents to discuss and critique artifacts beyond text [45]. Recent frameworks show that role-playing and conversation programming can structure such agent interactions toward specific goals [42, 29].

We operationalize the entrepreneur–investor dialectic as a learnable routine rather than a performance. In *Debate-to-Learn Entrepreneurship* (D2LE), an **Entrepreneur** agent advances opportunity hypotheses into concrete artifacts (market-sizing tables, unit-economics models, Go-To-Market (GTM) drafts), while an **Investor** agent conducts due diligence by probing assumptions, requesting sources, and proposing counter-calculations. A rubric-guided moderator coordinates turns and enforces a *retrieve-before-assert* rule so that debate moves map directly onto targeted literacy outcomes (opportunity recognition, market sizing, go-to-market, unit economics, risk, ethics). A rubric-guided moderator manages turn-taking, enforces retrieval-and-citation when claims are made, and aligns debate moves with learning objectives: opportunity recognition [2, 6], market sizing, go-to-market, unit economics, risk analysis, and ethics.

D2LE is *Socratic* by design, i.e., learners observe competing explanations interrogate one another in public view, which the Interactive, Constructive, Active, Passive (engagement framework) (ICAP) framework predicts will encourage interactive and constructive engagement [12]. To promote reliability, we integrate three mechanisms: (i) cross-agent verification and counter-argument; (ii) retrieval-augmented checking with explicit source citation; and (iii) post-hoc, chain-of-verification–style self-checks [28, 14, 3]. To support equity and inclusion, the moderator adapts language level, prior-knowledge scaffolds, domain exemplars, and risk tolerance to learner background and locale, drawing on Universal Design for Learning and culturally responsive pedagogy [9, 26].

Authentic practice in entrepreneurship involves iterated pitching, cross-examination, and diligence around evidence, assumptions, and feasibility. The entrepreneur–investor dyad operationalizes these norms as *structured* argumentation: founders propose, investors probe, and both surface gaps that trigger retrieval and computation. This mirrors cognitive accounts of opportunity recognition as a process of pattern detection and hypothesis refinement under uncertainty [2, 6]. This paper makes three contributions:

1. **Instructional pattern.** We formalize *Socratic Pitches* as an agent-led debate pattern for EE, mapping debate moves to specific literacy objectives and artifacts.
2. **Design for reliability.** We specify mechanisms for retrieval-augmented citation, cross-agent verification, and chain-of-verification to mitigate hallucinations during instruction [28, 14, 3].
3. **Evaluation roadmap.** We outline ablations that isolate the effects of retrieval, cross-examination, and rubric-guided moderation on learning outcomes, alongside analyses of inclusivity [12, 31].

By aligning debate structure with EE learning goals and modern LLM capabilities, D2LE aims to transform pitching from a performance into a learning activity grounded in evidence, critique, and inclusion.

2. Related Work

Meta-analyses and reviews report mixed and context-dependent effects of EE on intentions and performance, with calls for stronger designs that emphasize active, evidence-based practice [31, 37]. Within pedagogy, work has argued for moving beyond static business plans toward contingency-aware, iterative approaches and practice-oriented curricula [23, 38, 32]. However, most implementations still rely on cases, projects, and pitch days that foreground presentation rather than *structured* argumentation and adversarial testing of claims.

Research in the learning sciences shows that argumentation can deepen conceptual understanding when activities are well-scaffolded [4, 12, 1]. Computer-supported argumentation systems review decades of tools for making claims, evidence, and warrants explicit, yet they often require substantial instructor orchestration and rarely tie debate moves to domain-specific rubrics [39]. Our work adopts debate as the core learning activity but operationalizes it with agents that enforce retrieval, citation, and rubric alignment automatically.

Dialog-oriented Intelligent Tutoring System (ITS) and human tutoring can deliver large learning gains relative to lecture and practice-only conditions [41, 20]. These systems typically embody a single tutor role that guides a learner through scripted or adaptive prompts. In contrast, our approach leverages *multi-agent* interaction (Entrepreneur vs. Investor) to surface and resolve disagreements, with a moderator mapping debate moves to EE learning objectives.

Multi-agent debate improves factuality and reasoning in general Question Answering (QA) and reasoning tasks [24, 15], while frameworks for role-play and conversation programming make it easier to specify agent roles and protocols [42, 29]. Reliability techniques—retrieval-augmented generation and post-hoc verification—reduce hallucinations and support explicit citation [28, 3, 14]. The aforementioned approaches, however, optimizes for task accuracy or truthfulness rather than for instructional alignment to entrepreneurship literacy, and typically operates on text-only problems rather than multimodal venture artifacts.

Empirical studies show that investors interrogate teams, markets, technology, and unit economics through structured diligence processes [19]. Educational experiences simulate these practices informally (e.g., pitch competitions), but there is little work that *systematizes* the entrepreneur–investor dialectic into a reusable, rubric-aligned instructional pattern with automated retrieval and citation.

To our knowledge, there is no EE system that (i) instantiates a *entrepreneur–investor* debate with explicit, rubric-guided mapping to entrepreneurship literacy outcomes; (ii) requires retrieval and citation whenever claims are made; (iii) integrates cross-agent verification and post hoc chain-of-verification; and (iv) operates over multimodal venture artifacts (e.g., market-sizing sheets, unit-economics calculators, mockups). D2LE addresses this gap by combining multi-agent debate with reliability mechanisms and rubric-aligned moderation to turn pitching from a performance event into a repeatable learning activity.

3. System Overview

3.1. Roles, Artifacts, and Debate Protocol

Figure 1 represents the D2LE architecture. D2LE instantiates three roles: an **Entrepreneur** agent, an **Investor** agent, and a **Moderator**. The Entrepreneur synthesizes opportunity hypotheses into artifacts (e.g., problem statement, Jobs To Be Done (JTBD)/persona, market-sizing worksheet with Total Addressable Market (TAM)/Serviceable Available Market (SAM)/Serviceable Obtainable Market (SOM), unit-economics calculator, go-to-market plan, lean experiment design, and risk/ethics memo). The Investor performs due diligence by interrogating assumptions, requesting evidence, and critiquing artifacts. The Moderator orchestrates turns, enforces retrieval-and-citation for claims, and maps debate moves to entrepreneurship literacy outcomes.

Each learning episode proceeds in R rounds (typically 3–5):

1. **Entrepreneur move (proposal).** The Entrepreneur advances a concrete claim and/or artifact revision (e.g., a top-down TAM estimate, a Customer Acquisition Cost (CAC) assumption, or a channel hypothesis) with cited evidence when making knowledge claims.
2. **Investor move (cross-examination).** The Investor targets rubric-aligned facets—opportunity viability, market sizing, go-to-market fit, unit economics, risks, and ethics—by: (a) requesting sources; (b) challenging assumptions; (c) asking for alternative computations (e.g., bottom-up sizing); and (d) proposing counter-examples.
3. **Moderator interventions.** The Moderator (i) checks that claims are supported by retrieval with explicit citation; (ii) issues targeted prompts to close rubric gaps; and (iii) schedules verification checks (Section 3.3).

Rounds terminate when either (a) the target rubric cells reach mastery thresholds or (b) a budget on turns, tools, or uncertainty is met. The episode yields: (i) a versioned artifact bundle with embedded citations; (ii) a debate log linked to rubric evidence; and (iii) a learner-facing reflection with next steps.

3.2. Rubric Alignment and Enforcement

We operationalize entrepreneurship literacy as a rubric with cells for opportunity recognition, market sizing, go-to-market, unit economics, risk analysis, and ethics. Each debate move is tagged to one or more cells, and the Moderator

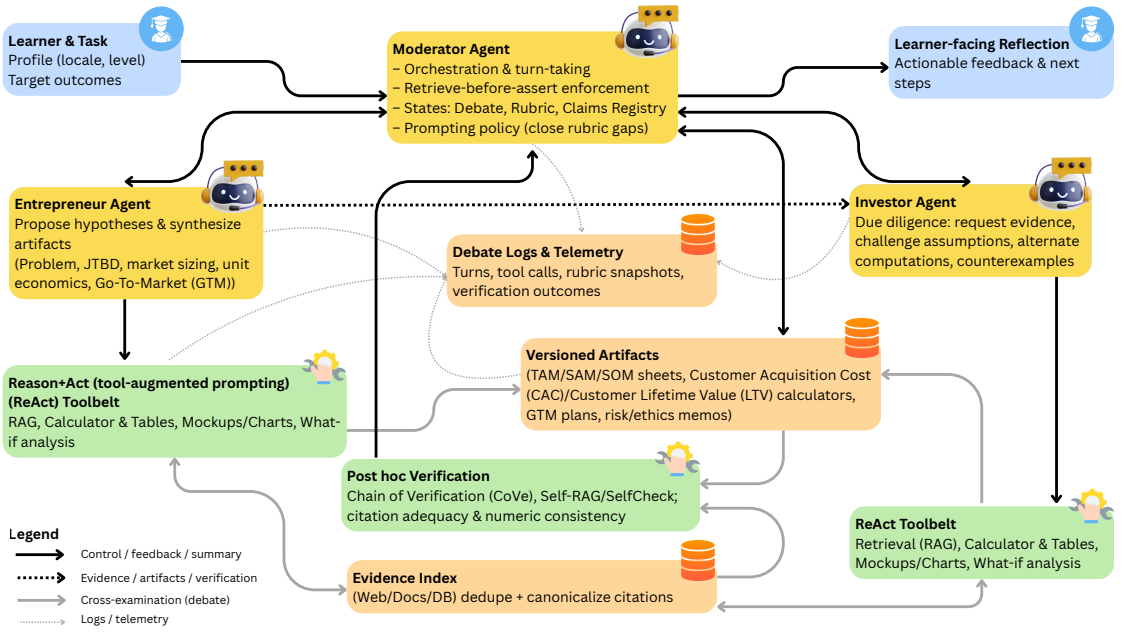

Fig. 1. **D2LE architecture.** Line types clarify flows: solid black (control/feedback), solid gray (data/evidence), dashed (cross-examination), dotted light gray (logs). The Moderator orchestrates Entrepreneur–Investor debate, enforces retrieve-before-assert, tracks rubric/claims states, and triggers verification. Agents use ReAct toolbelts for retrieval, computation, and artifact creation; post hoc verification audits citations and numerical consistency.

maintains a coverage/maturity state. When the Investor requests a claim, the Moderator enforces a *retrieve-before-assert* rule; uncited claims trigger repair prompts. Formative feedback is provided at the cell level (criteria, evidence observed, and actionable revision suggestions), leveraging best practices for rubric-guided learning and assessment [35]. The Moderator prioritizes prompts that maximize instructional utility, e.g., requesting a bottom-up sizing after a weak top-down estimate or a sensitivity analysis after brittle unit-economics assumptions.

3.3. Reliability Mechanisms

To reduce hallucinations and improve factual grounding, D2LE integrates complementary mechanisms:

- **Retrieval-augmented checking.** Claims invoking external knowledge use a Retrieval-Augmented Generation (RAG) pipeline with explicit citations to sources [28]. Agents interleave reasoning and tool use (ReAct) to plan evidence seeking and computations [43].
- **Cross-agent verification.** The Moderator samples salient claims and requests the counterpart to independently verify or refute them; disagreements trigger focused retrieval and, if needed, a forced replication of calculations or counterfactual analysis.
- **Post hoc verification.** After each round, a verification pass audits (a) citation adequacy and entailment, (b) numerical consistency in artifacts (e.g., CAC, Customer Lifetime Value (LTV), contribution margin), and (c) mismatch between cited sources and stated claims, using chain-of-verification/self-check style procedures [14, 3].

3.4. Adaptivity and Inclusion

The Moderator adapts language level, exemplars, and risk tolerance to learner background and locale (e.g., local regulations, purchasing power, and channel availability), consistent with Universal Design for Learning (UDL) and culturally responsive pedagogy [9, 26]. Readability is controlled via target ranges estimated by standard formulas to

keep explanations accessible without oversimplifying domain content [30]. Domain context (e.g., markets, competitors) and data sources are geo-tailored to the learner's region.

3.5. Implementation Notes

We have publicly released the source code on GitHub[1]. D2LE is implemented as a programmatic multi-agent scaffold: roles are specified via conversation programming with tool-use policies for web retrieval, tabular computation, and basic visualization. The Moderator maintains (i) a debate state (round, open issues, evidence ledger), (ii) a rubric state, and (iii) a *claims registry* that links each assertion to its supporting citations and verification status. We use an index that combines lexical and embedding retrieval for evidence, with source de-duplication and citation canonicalization. Agents use a lightweight action space for artifact operations (create, revise, compute, cite) and due-diligence queries.

4. Evaluation Methods

We will evaluate whether D2LE improves entrepreneurship literacy and why. **RQ1 (Learning efficacy).** Does D2LE yield higher post-test entrepreneurship literacy than strong baselines? **H1.** Participants using D2LE outperform baselines on rubric-aligned tasks after controlling for pre-test scores.

RQ2 (Component contributions). Which components drive learning? **H2.** Retrieval and explicit citation (*RAG*) and cross-examination (Entrepreneur–Investor debate) each contribute additively; post hoc verification adds incremental reliability benefits.

RQ3 (Reliability). Does D2LE reduce hallucinations and increase source support? **H3.** D2LE reduces unsupported/contradicted claims and raises citation adequacy relative to baselines.

RQ4 (Transfer). Do gains transfer to novel opportunity domains and data sources not seen during practice? **H4.** D2LE shows higher accuracy and argument quality on held-out scenarios.

RQ5 (Equity and inclusion). Are effects robust across language proficiency, prior entrepreneurship exposure, and locale? **H5.** D2LE maintains benefits across subgroups, with moderated effects modeled as interactions.

4.1. Experimental Conditions

Between-subjects randomized design with four arms: **Case+Tutor (baseline).** Single tutor-style agent providing guidance without adversarial debate or enforced citation (dialog ITS analogue). **Debate (no-RAG).** Entrepreneur–Investor debate without retrieval enforcement (tests debate-only effect). **Debate+RAG.** Debate with retrieval-and-citation required for claims. **D2LE (full).** Debate+RAG plus cross-agent verification and post hoc chain-of-verification. Arms are matched on time-on-task and content coverage. Randomization is blocked by site/course.

4.2. Participants and Setting

Undergraduate and graduate entrepreneurship courses, incubator bootcamps, and an online open cohort. We will collect demographics (age, gender, locale), prior entrepreneurship exposure, and English proficiency (self-report). Assignment is individual; collaboration is disallowed during assessment.

4.3. Tasks and Materials

Participants will complete two practice episodes (guided) and one graded *capstone episode* on a novel domain. All arms will receive identical prompts, datasets templates (e.g., market tables), and time budgets. The capstone requires: (i) opportunity definition, (ii) market sizing (top-down and bottom-up), (iii) unit economics (CAC, LTV, contribution margin), (iv) go-to-market rationale, and (v) risk/ethics analysis. Outputs include versioned artifacts and debate logs.

[1] https://github.com/akhadangi/D2LE

4.4. Measures

Primary learning outcomes. **Rubric composite (human-scored).** Blinded raters score opportunity recognition, market sizing, go-to-market, unit economics, risk, and ethics; interrater reliability via Krippendorff's α [22]. Analysis uses post-test with pre-test covariate (Analysis of Covariance (ANCOVA)) [40]. **Objective accuracy.** (a) Market sizing error: Mean Absolute Percentage Error (MAPE) on TAM/SAM/SOM against hidden reference calculations; (b) Unit-economics error: absolute deviation on CAC/LTV and margin given provided cost/revenue tables.

Secondary outcomes. **Argument/evidence quality.** Evidence density (citations per 1k words), support coverage (fraction of claims with adequate sources), and contradiction rate (claims contradicted by cited sources). **Transfer score.** Performance on a held-out domain and data source; items calibrated with item response theory for difficulty/severity [16]. **Entrepreneurial self-efficacy.** Short Entrepreneurial Self-Efficacy (ESE) scale pre/post [11]. **Process/workload/usability.** NASA Task Load Index (NASA–TLX) workload [21], System Usability Scale (SUS) usability [8]. **Effectuation/causation orientation.** Brief C&E scale to detect shifts in reasoning style [10].

Reliability and safety metrics. **Hallucination rate.** Fraction of factual claims flagged unsupported by verification passes and confirmed by human audit (claim-level adjudication). **Citation adequacy.** Precision/recall of citations relative to human gold labels (is the cited source sufficient and entailing?). **Numerical consistency.** Automated checks for internal consistency (e.g., LTV components sum to quoted value).

4.5. Analysis Plan

Primary analysis. Mixed-effects ANCOVA on rubric composite and objective accuracy with fixed effect for condition, covariate for pre-test, and random intercepts for site/instructor [18, 40]. Report *Hedges' g* and 95% CIs [13]. Control False Discovery Rate (FDR) across families via Benjamini–Hochberg [7]. **Component ablations.** Planned contrasts: (Debate no-RAG vs. Case+Tutor), (Debate+RAG vs. Debate), (D2LE vs. Debate+RAG). **Transfer.** Same model on held-out scenario scores; item difficulties accounted for via Item Response Theory (IRT)-based expected a posteriori scores [16]. **Reliability.** Between-arm comparisons on hallucination rate, citation adequacy, and numerical consistency via generalized linear mixed models. **Equity.** Interaction terms (Condition $\times$ Subgroup); where no harms are expected, test non-inferiority/equivalence to rule out practically meaningful gaps using Two One-Sided Tests (equivalence) (TOST) [27]. **Power.** Target small-to-moderate effects ($d \approx 0.30$–0.40). With four arms and pre-test covariate, simulations and standard power tools (e.g., G*Power) guide sample sizing [17]. **Pre-registration and materials.** We will pre-register hypotheses, analyses, and stopping rules; de-identified data and rubrics will be released post-study [34].

4.6. Threats to Validity and Mitigation

Internal validity. Random assignment with blocked randomization; identical prompts/time budgets across arms; proctoring for assessments; blinding raters to condition; manipulation checks (evidence density, enforcement logs). **Construct validity.** Multi-method outcomes (human rubric + objective accuracy + IRT-transfer); reliability checks (Krippendorff's α); validated self-report instruments. **External validity.** Multi-site deployment (courses, bootcamps, online cohort); report context and learner characteristics to support transportability. **Validity of the statistical conclusion.** A priori power analysis; FDR control; model diagnostics; robustness checks (change scores vs. ANCOVA, non-parametric tests). **Ethics.** Institutional Review Board (IRB) review, informed consent, privacy-preserving logs, and opt-out options.

5. Conclusion

We introduced *Socratic Pitches*—the D2LE framework in which an **Entrepreneur** and **Investor** agent debate under a rubric-guided **Moderator** to teach entrepreneurship literacy. The system converts pitching from a one-way performance into a structured learning activity: claims must be retrieved and cited, artifacts are iteratively produced

and critiqued (market sizing, unit economics, go-to-market, risk/ethics), and cross-agent plus post hoc verification reduce unsupported assertions. Adaptivity in language level, exemplars, and risk tolerance targets equity and inclusion.

We framed a study plan that makes the mechanism testable: ablations separate the effects of debate, retrieval with citation, and verification; outcomes mix human-scored rubrics with objective accuracy on market and unit-economics calculations, transfer to novel domains, reliability metrics (hallucination and citation adequacy), and subgroup analyses. All debate logs, artifact versions, and enforcement events are captured to support analysis and replication.

Limitations: D2LE depends on the quality and coverage of external evidence; retrieval errors or noisy sources can still slip through despite verification. Numerical checks catch consistency, but not all modeling mistakes. Debate dynamics may occasionally converge prematurely or amplify weak priors without targeted moderation. Compute and latency budgets can restrict classroom use, and multilingual or low-resource contexts can produce uneven evidence availability.

Future Work: We will (i) run multi-site evaluations with pre-registration and public releases of code, prompts, rubrics, and anonymized traces; (ii) explore hybrid modes where instructors inject challenges or override rulings; (iii) extend to multilingual and locale-specific data sources, with fairness audits on accessibility and differential workload; (iv) add stronger entailment checks and citation deduplication tuned to business/market domains; and (v) study long-term retention and downstream behaviors. Beyond entrepreneurship, we see the entrepreneur–investor dialectic as a template for other domains that benefit from adversarial yet evidence-grounded practice (policy analysis, product management, bioethics).

Acknowledgments

This research was conducted as part of the "LLM Cognition Toolbox for SMEs" project in partnership with the crowdfunding analytics platform Wixdom (https://www.wixdom.io/). It was also partially funded by the Luxembourg National Research Fund (FNR) under grant reference NCER22/IS/16570468/NCER-FT and by the PayPal-FNR PEARL grant under reference number 13342933/Gilbert Fridgen. For open access purposes, the authors have applied a CC BY 4.0 license to any Author Accepted Manuscript arising from this submission.

References

[1] Andriessen, J., Baker, M., Suthers, D.D., 2013. Arguing to learn: Confronting cognitions in computer-supported collaborative learning environments. volume 1. Springer Science & Business Media.

[2] Ardichvili, A., Cardozo, R., Ray, S., 2003. A theory of entrepreneurial opportunity identification and development. Journal of Business venturing 18, 105–123.

[3] Asai, A., Wu, Z., Wang, Y., Sil, A., Hajishirzi, H., 2024. Self-rag: Learning to retrieve, generate, and critique through self-reflection .

[4] Asterhan, C.S., Schwarz, B.B., 2016. Argumentation for learning: Well-trodden paths and unexplored territories. Educational Psychologist 51, 164–187.

[5] Bae, T.J., Qian, S., Miao, C., Fiet, J.O., 2014. The relationship between entrepreneurship education and entrepreneurial intentions: A meta–analytic review. Entrepreneurship theory and practice 38, 217–254.

[6] Baron, R.A., 2006. Opportunity recognition as pattern recognition: How entrepreneurs "connect the dots" to identify new business opportunities. Academy of management perspectives 20, 104–119.

[7] Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal statistical society: series B (Methodological) 57, 289–300.

[8] Brooke, J., 1996. Sus: A quick and dirty usability scale, in: Usability Evaluation in Industry, Taylor & Francis.

[9] CAST, 2024. Cast universal design for learning guidelines version 3.0. CAST Guidelines. URL: `https://udlguidelines.cast.org/`.

[10] Chandler, G.N., DeTienne, D.R., McKelvie, A., Mumford, T.V., 2011. Causation and effectuation processes: A validation study. Journal of business venturing 26, 375–390.

[11] Chen, C.C., Greene, P.G., Crick, A., 1998. Does entrepreneurial self-efficacy distinguish entrepreneurs from managers? Journal of business venturing 13, 295–316.

[12] Chi, M.T., Wylie, R., 2014. The icap framework: Linking cognitive engagement to active learning outcomes. Educational psychologist 49, 219–243.

[13] Cohen, J., 2013. Statistical power analysis for the behavioral sciences. routledge.

[14] Dhuliawala, S., Komeili, M., Xu, J., Raileanu, R., Li, X., Celikyilmaz, A., Weston, J., 2024. Chain-of-verification reduces hallucination in large language models, in: Findings of the association for computational linguistics: ACL 2024, pp. 3563–3578.

[15] Du, Y., Li, S., Torralba, A., Tenenbaum, J.B., Mordatch, I., 2023. Improving factuality and reasoning in language models through multiagent debate, in: Forty-first International Conference on Machine Learning.

[16] Embretson, S.E., Reise, S.P., 2013. Item response theory for psychologists. Psychology Press.
[17] Faul, F., Erdfelder, E., Lang, A.G., Buchner, A., 2007. G* power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behavior research methods 39, 175–191.
[18] Gelman, A., Hill, J., 2007. Data analysis using regression and multilevel/hierarchical models. Cambridge university press.
[19] Gompers, P.A., Gornall, W., Kaplan, S.N., Strebulaev, I.A., 2020. How do venture capitalists make decisions? Journal of Financial Economics 135, 169–190.
[20] Graesser, A.C., Lu, S., Jackson, G.T., Mitchell, H.H., Ventura, M., Olney, A., Louwerse, M.M., 2004. Autotutor: A tutor with dialogue in natural language. Behavior Research Methods, Instruments, & Computers 36, 180–192.
[21] Hart, S.G., Staveland, L.E., 1988. Development of nasa-tlx (task load index): Results of empirical and theoretical research, in: Advances in psychology. Elsevier. volume 52, pp. 139–183.
[22] Hayes, A.F., Krippendorff, K., 2007. Answering the call for a standard reliability measure for coding data. Communication methods and measures 1, 77–89.
[23] Honig, B., 2004. Entrepreneurship education: Toward a model of contingency-based business planning. Academy of management learning & education 3, 258–273.
[24] Irving, G., Christiano, P., Amodei, D., 2018. Ai safety via debate. arXiv preprint arXiv:1805.00899 .
[25] Khadangi, A., Sartipi, A., Tchappi, I., Fridgen, G., 2025. Cognartive: Large language models for automating art analysis and decoding aesthetic elements. arXiv preprint arXiv:2502.04353 .
[26] Ladson-Billings, G., 1995. Toward a theory of culturally relevant pedagogy. American educational research journal 32, 465–491.
[27] Lakens, D., 2017. Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. Social psychological and personality science 8, 355–362.
[28] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.t., Rocktäschel, T., et al., 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in neural information processing systems 33, 9459–9474.
[29] Li, G., Hammoud, H., Itani, H., Khizbullin, D., Ghanem, B., 2023. Camel: Communicative agents for" mind" exploration of large language model society. Advances in Neural Information Processing Systems 36, 51991–52008.
[30] Mc Laughlin, G.H., 1969. Smog grading-a new readability formula. Journal of reading 12, 639–646.
[31] Nabi, G., Liñán, F., Fayolle, A., Krueger, N., Walmsley, A., 2017. The impact of entrepreneurship education in higher education: A systematic review and research agenda. Academy of management learning & education 16, 277–299.
[32] Neck, H.M., Corbett, A.C., 2018. The scholarship of teaching and learning entrepreneurship. Entrepreneurship Education and Pedagogy 1, 8–41.
[33] Noroozi, O., Weinberger, A., Biemans, H.J., Mulder, M., Chizari, M., 2012. Argumentation-based computer supported collaborative learning (abcscl): A synthesis of 15 years of research. Educational Research Review 7, 79–106. URL: https://www.sciencedirect.com/science/article/pii/S1747938X11000522, doi:https://doi.org/10.1016/j.edurev.2011.11.006.
[34] Nosek, B.A., Ebersole, C.R., DeHaven, A.C., Mellor, D.T., 2018. The preregistration revolution. Proceedings of the National Academy of Sciences 115, 2600–2606.
[35] Panadero, E., Jonsson, A., 2013. The use of scoring rubrics for formative assessment purposes revisited: A review. Educational research review 9, 129–144.
[36] Rațiu, A., Maniu, I., Pop, E.L., 2023. Entrecomp framework: A bibliometric review and research trends. Sustainability 15, 1285.
[37] Rideout, E.C., Gray, D.O., 2013. Does entrepreneurship education really work? a review and methodological critique of the empirical literature on the effects of university-based entrepreneurship education. Journal of small business management 51, 329–351.
[38] Sarasvathy, S.D., 2001. Causation and effectuation: Toward a theoretical shift from economic inevitability to entrepreneurial contingency. Academy of management Review 26, 243–263.
[39] Scheuer, O., Loll, F., Pinkwart, N., McLaren, B.M., 2010. Computer-supported argumentation: A review of the state of the art. International Journal of Computer-supported collaborative learning 5, 43–102.
[40] Van Breukelen, G., 2006. Ancova versus change from baseline: more power inrandomized studies, more bias in nonrandomized studies. Journal of ClinicalEpidemiology , 59–920.
[41] VanLehn, K., 2011. The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. Educational psychologist 46, 197–221.
[42] Wu, Q., Bansal, G., Zhang, J., Wu, Y., Li, B., Zhu, E., Jiang, L., Zhang, X., Zhang, S., Liu, J., et al., 2024. Autogen: Enabling next-gen llm applications via multi-agent conversations, in: First Conference on Language Modeling.
[43] Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K.R., Cao, Y., 2022. React: Synergizing reasoning and acting in language models, in: The eleventh international conference on learning representations.
[44] Zang, Q., Zgrzendek, C., Tchappi, I., Khadangi, A., Sedlmeir, J., 2025. Kg-htc: Integrating knowledge graphs into llms for effective zero-shot hierarchical text classification. arXiv preprint arXiv:2505.05583 .
[45] Zeng, A., Attarian, M., Ichter, B., Choromanski, K., Wong, A., Welker, S., Tombari, F., Purohit, A., Ryoo, M., Sindhwani, V., et al., 2022. Socratic models: Composing zero-shot multimodal reasoning with language. arXiv preprint arXiv:2204.00598 .