Review

Not peer-reviewed version

The Convergence of Federated Learning, Knowledge Graphs, and Large Language Models for Language Learning: A Scoping Review

Michael Kenteris * and Konstantinos Kotis *

Posted Date: 21 January 2026

doi: 10.20944/preprints202601.1584.v1

Keywords: federated learning; knowledge graphs; large language models; iCALL; CEFR; pedagogical grounding; human-centered ai; system validation; automated decision-making

 doi:10.20944/preprints202601.1584.v1

Review

The Convergence of Federated Learning, Knowledge Graphs, and Large Language Models for Language Learning: A Scoping Review

Michael Kenteris * and Konstantinos Kotis

Department of Cultural Technology and Communication, University of the Aegean, Mytilene 81100, Greece

* Correspondence: mkenteris@aegean.gr

Abstract

Large Language Models (LLMs) in Intelligent Computer-Assisted Language Learning (iCALL) offer personalization potential but introduce critical challenges in pedagogical grounding, data privacy, and pedagogical validity. While Knowledge Graphs (KGs) and Federated Learning (FL) address these concerns individually, systematic integration of all three technologies remains absent or insufficiently addressed in current research. This scoping review maps the FL–KG–LLM convergence landscape in educational contexts. Following PRISMA-ScR guidelines, we searched six databases and screened 51 papers published between 2019 and 2025 using automated extraction. Our findings reveal a pronounced convergence deficit: no papers integrate all three domains, while 58.8% of approaches operate within isolated technological silos. Critical reporting gaps emerge across the corpus, with an average "Not Reported" (NR) rate of 84.5%, particularly in privacy mechanisms (92.2%), validation metrics (90.2%), and Common European Framework of Reference for Languages (CEFR) alignment (88.2%). Domain-specific analysis reveals two distinct patterns: inter-domain gaps (disciplinary silos resulting in expected CEFR absence in single-domain papers) and intra-domain gaps (failure to report domain-critical variables, including 100% parameter NR in FL studies, 86.7% validation NR in KG studies, and 100% CEFR NR in convergence papers). We identify two pillars of pedagogical grounding: a Grounding Pillar, which constrains LLM outputs via Knowledge Graph rules, and a Validation Pillar, which concerns how authoritative source frameworks are mapped onto Knowledge Graph schemas. The latter remains completely unaddressed in the reviewed literature, revealing what we term the Integrity Gap—a systematic disconnection between technological innovation and pedagogical grounding in iCALL. By framing pedagogical alignment as an upstream control and validation problem, this review offers insights relevant to the design of user-facing automated systems where trust, transparency, and human oversight are critical.

Keywords: federated learning; knowledge graphs; large language models; iCALL; CEFR; pedagogical grounding; human-centered ai; system validation; automated decision-making

1. Introduction

1.1. The Paradox of Personalization in iCALL

The emergence of Large Language Models (LLMs) [19] has accelerated the integration of Artificial Intelligence into Intelligent Computer-Assisted Language Learning (iCALL), promising unprecedented capabilities for personalized content generation [1,2,17]. Yet this technological promise confronts a critical triad of interconnected challenges [8]:pedagogical integrity, data integrity, and instructional validity.

Pedagogical Integrity refers to the lack of reliable grounding in established educational frameworks—particularly the Common European Framework of Reference for Languages (CEFR) [4] or comparable internationally recognized proficiency and curriculum standards. As a result, generated content may routinely exceed learners' Zone of Proximal Development [5] (ZPD).

 doi:10.20944/preprints202601.1584.v1

Data Integrity encompasses the privacy risks inherent in centralized training paradigms, which require aggregating sensitive learner interaction data on external servers.

Instructional Validity addresses the stochastic nature of LLMs, which leads to hallucinations [3] and inconsistent alignment with pedagogical standards, thereby undermining institutional trust in educational contexts.

Critically, existing literature addresses these challenges in isolation: Knowledge Graphs [18] for pedagogical grounding, Federated Learning [7] for privacy preservation [14], or LLM fine-tuning [33,34,39] for personalization. Yet no systematic framework integrates all three technologies to simultaneously ensure data sovereignty, pedagogical validity, and adaptive instruction. This convergence deficit represents a significant gap in literature, motivating our scoping review. Even if convergence were achieved, would integration alone ensure pedagogical validity? Not necessarily.

1.2. The Dual-Pillar Grounding Problem

While neurosymbolic AI approaches address the Grounding Pillar—constraining LLM outputs through KG-based retrieval [28,37]—they largely overlook the Validation Pillar, which concerns whether Knowledge Graphs themselves accurately represent authoritative pedagogical frameworks. Inspection of CEFR Companion Volume (CV) Sociolinguistic Competence descriptors reported in the literature suggests—drawing on an internationally standardized framework validated across European languages—revealed interpretive ambiguities in how source frameworks map onto KG schemas. These ambiguities require explicit ontology design decisions that structure the source framework's inherent complexity. A previously underexplored gap emerges: when KG source framework verification is absent rather than assumed automatic, schema design choices can silently propagate these ambiguities into the resulting KG without detection, potentially undermining pedagogical validity at the representational level.

Knowledge Graphs have increasingly been used to structure pedagogical knowledge and constrain generative models in educational systems [9].

This insight fundamentally motivates our central research question: What is the current state of integration among Federated Learning, Knowledge Graphs, and Large Language Models in educational contexts, and what critical gaps exist that future research must address?

1.3. Research Questions and Scope

Primary RQs:

- RQ1: What is the current convergence landscape of FL, KG, and LLM technologies in iCALL (during 2019–2025)?
- RQ2: What methodological reporting standards exist for hybrid FL-KG-LLM systems?
- RQ3: How do papers address Validation Pillar grounding, specifically the systematic verification of Knowledge Graph representational fidelity?

Secondary RQs:

- RQ4: What pedagogical validation metrics are employed in current research work addressing iCALL?
- RQ5: What barriers prevent convergence despite technological maturity of each domain?

The remainder of this paper is organized as follows: Section 2 presents the research methodology, including the research protocol and search strategy following the PRISMA-ScR guidelines. Section 3 reports the results of our research, including the selection process and characteristics of included related studies. Section 4 discusses the findings, identifying key issues, gaps, and future research directions. Then, Section 5 provides the conclusion with implications for practice and research.

 doi:10.20944/preprints202601.1584.v1

2. Materials and Methods

2.1. Methodological Overview

This scoping review follows PRISMA-ScR [21,23,26] (Preferred Reporting Items for Systematic reviews and Meta-Analyses, including the 2020 updated guidelines [26]). We conducted post-hoc registration on the Open Science Framework (OSF) [25] on December 23, 2025, positioning registration before the critical data extraction stage to reduce bias while acknowledging retrospective screening and search stages. The research phases and their retrospective/prospective status are summarized in Table 1.

Rationale for scoping review design: A scoping review is appropriate for mapping convergence in emerging interdisciplinary fields [15] where integration is nascent, variation in study designs are extreme, and synthesis questions focus on landscape characterization rather than intervention effectiveness.

Table 1. Study Selection Timeline: Research Phases and Status.

Stage	Timeline	Status	Method
1. Preparation	November 2025	RETROSPECTIVE	Research question formulation, inclusion/exclusion criteria
2. Search	November 2025	RETROSPECTIVE	Iterative database searches (IEEE, ACM, Scholar, arXiv, Scopus, WoS)
3. Screening	Nov-Dec 2025	RETROSPECTIVE	Manual deduplication & screening (~660 → 51 papers)
4. Critical Appraisal	December 2025	RETROSPECTIVE	Technical Quality Rubric application
5. OSF Registration	Dec 23, 2025	PROSPECTIVE START	Pre-specify codebook & synthesis plan
6. Data Extraction	December 2025	PROSPECTIVE	Automated AI-assisted extraction (Qwen 2.5 7B LLM)
7. Synthesis	Dec 2025-Jan 2026	PROSPECTIVE	Thematic analysis, gap mapping, framework proposal
8. Reporting	January 2026	PROSPECTIVE	MDPI manuscript preparation, PRISMA-ScR compliance

Study selection timeline documenting research phases, dates, methodological status (retrospective vs prospective), and procedures. Phases 1-4 were conducted retrospectively before OSF registration; Phases 5-7 prospectively after registration, reducing bias in extraction and synthesis decisions.

Scope Note on Foundational References: While the primary study corpus comprises 51 papers meeting full inclusion criteria (educational context, 2019–2025 publication, explicit FL/KG/LLM focus), select foundational works in federated learning privacy are cited in the Results section for technical specifications and methodological context. These references (e.g., Tayyeh & AL-Jumaili, 2024; Bonawitz et al., 2017) were excluded from primary studies due to non-educational application or publication date but remain relevant for informing future educational implementations.

2.2. Search Strategy and Information Sources

We searched six databases spanning computer science, information systems, and education: Initial hit distribution across the six databases is shown in Table 2.

Table 2. Database Search Results: Initial Hit Distribution.

Database	Approximate Hits	Notes
IEEE Xplore	~120	FL architecture papers
ACM Digital Library	~85	HCI/iCALL systems

 doi:10.20944/preprints202601.1584.v1

Google Scholar	~200	First 10 pages reviewed
arXiv	~75	High proportion of pre-prints (40-50%)
Scopus	~150	Overlaps with IEEE/ACM
Targeted Snowballing	~30	From reference lists
TOTAL INITIAL HITS	~660	Pre-deduplication

Initial database search results across six information sources, yielding approximately 660 records before deduplication.

The largest yields were obtained from IEEE Xplore and Scopus. To manage scope, only the first ten pages of Google Scholar results were screened. arXiv was included due to the high proportion of relevant preprints (approximately 40–50%) in emerging fields. Targeted snowballing from reference lists supplemented the database searches.

Search strategy evolution: The search strategy was operationalized through three iterative query phases to capture convergence without false positives. Initial broad queries yielded unmanageable results with low relevance; we therefore constrained the search to the FL–KG–LLM intersection to ensure papers addressed all three technologies rather than general AI-in-education contexts.

- Phase 1 (Broad): (LLM OR "large language model" OR GPT [16]*) AND (iCALL OR "language learning" OR "language instruction")
- Phase 2 (Convergence-focused): (Federated Learning OR FL) AND (Knowledge Graph OR KG) AND (LLM OR "language model")
- Phase 3 (Integration): (FL-KG-LLM OR "federated knowledge graphs" OR "privacy-preserving language learning")

Database-specific Boolean operators, wildcard conventions, and phrase-search syntax varied across platforms and required query adaptation; full technical search syntax details are provided in the Supplementary Materials (S1).

Temporal scope: 2019–2025, capturing the emergence of modern transformer-based LLMs (GPT-2+) coinciding with federated learning [6] maturity.

The temporal distribution of included papers is shown in Figure 1. Publication activity accelerates sharply after 2022, with 74.5% of included studies (n = 38) published between 2023 and 2025. This temporal distribution reflects the nascent maturity of the field, which emerged only after several foundational technologies reached sufficient maturity, including the revision of the CEFR through the CV (2020), the emergence of transformer-based large language models (2018), and the standardization of federated learning approaches (2019). Early publications (2019–2020, n = 4) primarily reflect foundational explorations, followed by a transitional phase (2021–2022, n = 9) showing early convergence, and a recent period (2023–2025, n = 38) characterized by rapidly increasing research interest.

 doi:10.20944/preprints202601.1584.v1

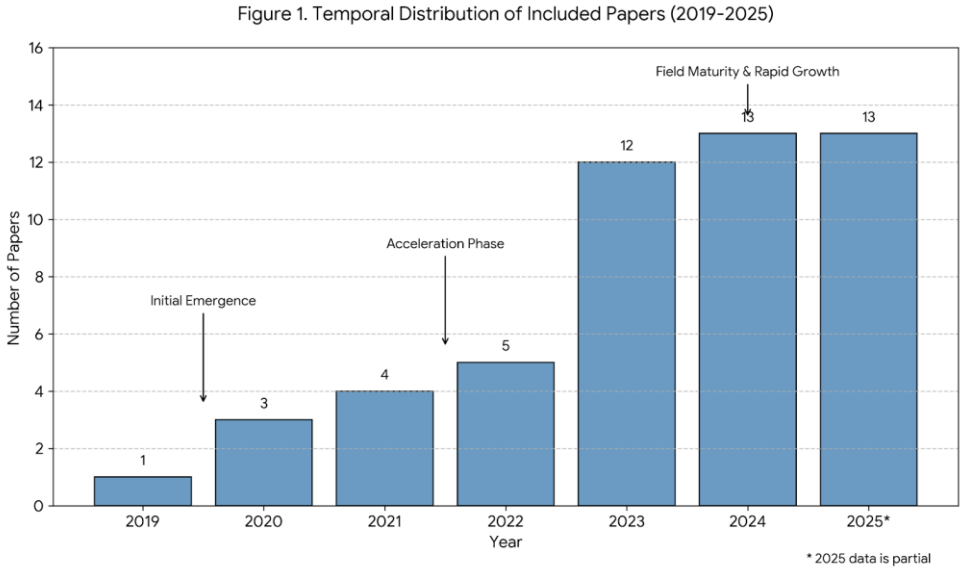

Figure 1. Temporal Distribution of 51 Included Papers (2019-2025).

2.2.1. Search Strategy Evolution Rationale

Our three-phase search strategy evolved from exploratory to targeted as we learned the convergence landscape. This iterative approach reflects the scoping review methodology's exploratory nature, where the research phenomenon (FL-KG-LLM convergence) may be poorly characterized initially.

Phase 1 (Broad Discovery - November 2025): Initial searches established the upper bound of AI-in-education literature addressing language learning, regardless of convergence. Queries targeted LLM applications broadly ((LLM OR "large language model") AND (iCALL OR "language learning")). This phase revealed dominance of single-domain LLM applications (prompt engineering, fine-tuning) without FL or KG integration. From ~660 initial hits, 78.9% were excluded at title/abstract screening—a high rejection rate justified by the intentionally overinclusive design.

Phase 2 (Convergence-Focused - November 2025): After Phase 1 screening revealed 0/80 candidates were convergent, we refined queries to directly target FL-KG-LLM intersection ((Federated Learning) AND (Knowledge Graph) AND (LLM)). This phase tested whether the convergence deficit was real or a search artifact. Minimal new hits confirmed convergence deficit.

Phase 3 (Integration-Specific - December 2025): Final searches targeted specialized terminology ('federated knowledge graphs,' 'privacy-preserving language learning') that might not use FL/KG/LLM keywords explicitly. Zero new papers suggested convergence deficit is not merely terminological.

Rationale for Iterative Approach: In emerging interdisciplinary fields, optimal search terms may be unknown a priori. Our phased strategy balances breadth (avoiding missing relevant papers) with precision (reducing false positives). While retrospective, this evolution is documented transparently to enable replication and assessment of potential bias.

Limitations: This approach may miss (1) foundational work published before 2019 (temporal scope), and (2) papers using alternative terminology not captured in our three phases (e.g., 'distributed knowledge graphs' instead of 'federated learning').

2.3. Screening and Selection Process

Inclusion criteria:

- Research papers explicitly addressing at least two of the following technologies: Federated Learning, Knowledge Graphs, and Large Language Models, or proposing integrated architectures combining these technologies for language learning applications.

 doi:10.20944/preprints202601.1584.v1

- Educational context: language learning, iCALL, or natural language instruction
- Published between 2019 and 2025
- Available in English

Exclusion criteria:

- Non-educational applications (healthcare, finance, general NLP) unless pedagogically transferable
- Insufficient technical transparency (black-box prompting without architecture details)
- Duplicate studies (different venues, same architecture)

Phase 1 exclusion reasons (temporal and domain scope) are detailed in Table 3. Phase 2 exclusion reasons (methodological and access) are presented in Table 4.

Initial ~660 hits were deduplicated to ~380 unique papers via Zotero v7.0. Title/abstract screening reduced to ~80 candidates. Full-text review applied inclusion/exclusion criteria, yielding the final 51 papers. The screening workflow is documented on the Open Science Framework (OSF), including the complete list of 51 included papers and 29 excluded papers with rationales.

Single-Reviewer Screening. Title/abstract screening (Phase 2) was conducted by a single reviewer without formal inter-rater reliability assessment. Borderline cases (e.g., papers with ambiguous titles like "Dialogue Systems for Language Learning") were discussed with co-authors, and inclusion required explicit educational context (iCALL, language learning, language instruction) in title or abstract. While this approach introduces potential subjective bias, the high rejection rate (78.9%) reflects appropriate filtering of general NLP/AI papers outside language learning contexts. Future replications should employ dual coding at screening stage with Cohen's kappa reporting.

The PRISMA-ScR flow diagram illustrating the complete screening process is shown in Figure 2.

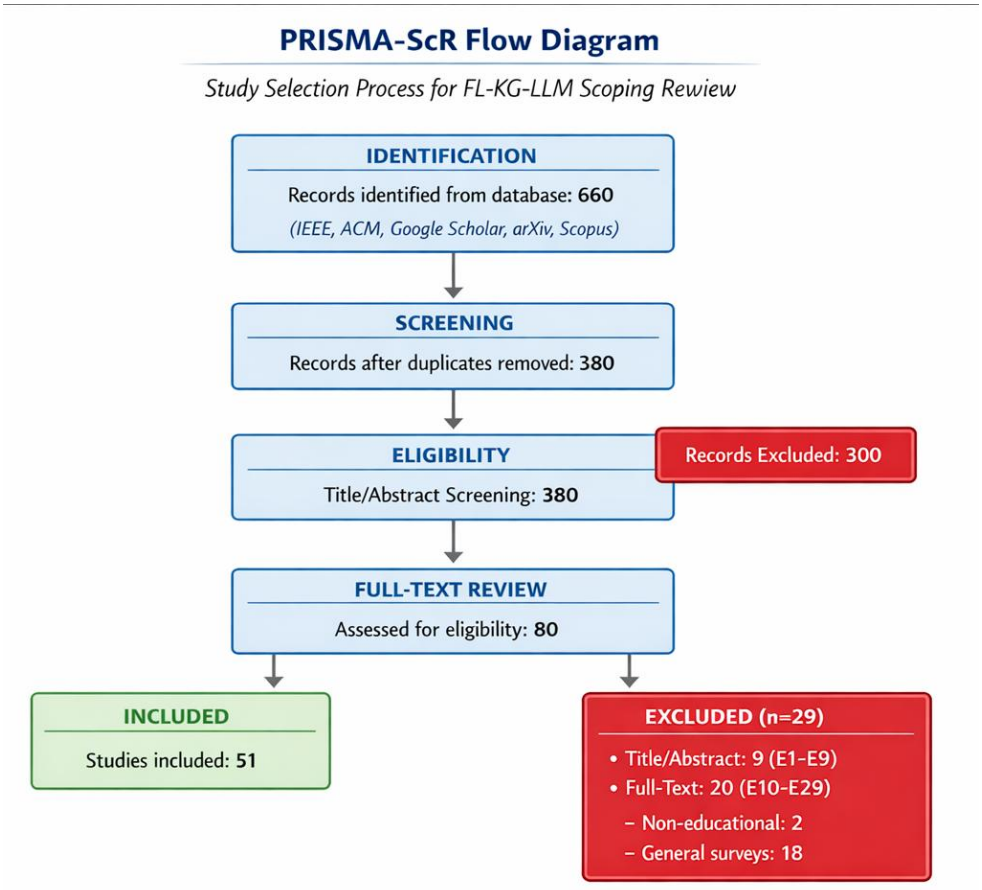

Figure 2. PRISMA-ScR Flow Diagram showing systematic progression from 660 initial database hits to 51 final included papers. After deduplication (n=380 unique), title/abstract screening (n=80 candidates), and full-text

 doi:10.20944/preprints202601.1584.v1

review, 51 papers met inclusion criteria. Twenty-nine papers were excluded for reasons detailed in exclusion analysis.

Table 3. Exclusion Reasons: Temporal and Domain Scope (Phase 1).

Exclusion Reason	Papers Excluded
Out of temporal scope (pre-2019)	~50
Lack of LLM/GenAI focus (legacy NLP)	~80
Non-educational context	~120
Incomplete technological coverage (domain-specific concerns)	~100

Phase 1 exclusion reasons showing papers removed for temporal (pre-2019), technological (legacy NLP without modern LLM/GenAI focus), contextual (non-educational applications), or incomplete coverage (single-domain focus without language education [20] relevance).

Table 4. Exclusion Reasons: Methodological and Access (Phase 2).

Exclusion Reason	Papers Excluded
Insufficient technical transparency (no architecture details)	~8
Lack of pedagogical grounding (black-box prompting only)	~6
Duplicate studies (same architecture, different venues)	~3
Full-text not accessible	~2

Phase 2 exclusion reasons showing papers removed for methodological (insufficient transparency, black-box approaches), publication (duplicate studies), or accessibility (full-text not available) criteria.

2.4. Critical Appraisal

Quality assessment: We employed a confidence scoring rubric based on three assessment criteria: The quality assessment criteria are defined in Table 5.

Table 5. Quality Assessment Criteria: Standards for Transparency and Rigor.

Criterion	High Quality	Medium Quality	Low Quality
Architectural Transparency	Code + hyperparameters public	Architecture described, no code	Vague/missing details
Empirical Rigor	Statistical significance reported	Metrics reported, no significance	Anecdotal/qualitative only
Pedagogical Alignment	CEFR explicitly defined via constraints	CEFR mentioned, no formalization	No pedagogical framework

Quality assessment rubric showing three assessment criteria (architectural transparency, empirical rigor, pedagogical alignment) with criteria for high, medium, and low quality classification. Quality scoring informed confidence assessment for automated extraction validation.

Quality assessment rubric (Table 5) evaluated three assessment criteria: architectural transparency, empirical rigor, and pedagogical alignment. Confidence scores (0-1 scale) reflected reporting completeness for each criterion; papers scoring <0.2 were classified as 'low reporting transparency' (70.6% of corpus). We distinguish between reporting transparency (confidence score) and methodological quality (actual study design), noting that a paper could be methodologically sound yet poorly reported.

Confidence scoring: Automated AI-computed confidence scores (0-1 scale) via Qwen 2.5 7B reflected reporting completeness. Average confidence: 0.17 (SD=0.23), indicating widespread reporting gaps. Distribution of studies across confidence tiers is shown in Table 6.

 doi:10.20944/preprints202601.1584.v1

Table 6. Study Quality Assessment: Distribution Across Confidence Tiers.

Characteristic	Count	Percentage
Total papers	51	100%
High-quality (conf > 0.4)	5	9.8%
Medium-quality (conf 0.2-0.4)	10	19.6%
Low-quality (conf < 0.2)	36	70.6%
Average confidence score	0.17	•

Study quality assessment showing distribution across confidence tiers. Only 9.8% achieved high quality (>0.4 confidence); 70.6% fell into low quality (<0.2), indicating that most papers lack transparency in architectural, empirical, or pedagogical criteria.

2.5. Data Extraction and Analysis

Extraction methodology: Automated AI-assisted extraction via Qwen 2.5 7B with Pydantic validation ensured consistent structured output. Sixteen variables were extracted for each paper:

Variables spanned three categories:

1. **Technology presence:** FL_present, KG_present, LLM_present, convergence_type
2. **Technical characteristics:** architecture_transparency, privacy_mechanism, validation_metrics
3. **Pedagogical characteristics:** cefr_alignment, pedagogical_framework, learning_outcomes.

Manual/Human-performed verification: 20% of papers (n=10) underwent manual audit by a human supervisor to validate automated extraction, achieving inter-rater agreement (Cohen's kappa = 0.92). Post-Hoc Domain-Specific Analysis: Domain-specific NR rates by paper type (FL-only, KG-only, LLM-only, hybrid) were calculated post-hoc from extracted data (Data_Extraction_Results_v1.csv) to distinguish expected gaps (domain-appropriate NR) from crisis gaps (domain-inappropriate NR). This represents retrospective cross-tabulation of preregistered extraction variables, not protocol modification. The OSF protocol documented extraction variables and methods; domain-specific analysis is an additional analytical layer applied to existing data, analogous to subgroup analysis in clinical trials.

2.6. Grouping Approach

Papers were grouped by convergence type (single-domain, dual-domain, triple-domain) and pedagogical engagement level. Frequency analysis characterized the corpus. Pre-specified analytical thresholds examined five explicit predictions:

- **Research Question 1 (Convergence):** What is the convergence rate of FL-KG-LLM integration?

  - Expected Finding: <15% convergence based on preliminary searches revealing zero convergent papers in the initial candidate set (n=80 screened). This threshold allows for some false negatives while testing whether convergence deficit is genuine.
- **Research Question 2 (Reporting):** What is the extent of methodological reporting gaps?

  - Expected Finding: >60% Not Reported rate based on scoping reviews in emerging ML fields showing 50-90% NR rates (e.g., privacy-preserving ML: 70-85% NR for hyperparameters; edTech reviews: 50-70% NR for validation metrics).
- Research Question 3 (Scale Bias): Are FL systems predominantly centralized?

  - Expected Finding: >70% centralized based on FL literature showing educational deployments favor centralized architectures due to implementation complexity.
- **Research Question 4 (Pedagogical Frameworks):** Are pedagogical frameworks systematically integrated?

  - Expected Finding: <20% CEFR mention based on disciplinary silos between ML and language education communities.
- **Research Question 5 (Validation):** Are pedagogical validation metrics reported?

  - Expected Finding: <10% pedagogical metrics based on ML field norms favoring technical metrics (BLEU, accuracy) over learning outcomes.

 doi:10.20944/preprints202601.1584.v1

3. Results of the Scoping Review

3.1. Study Selection and Characteristics

Methodological reporting gaps were widespread, with NR rates exceeding 70% across multiple criteria.

Exclusion rationales at full-text stage:

Twenty-nine papers were excluded across both screening stages. Representative exclusion examples from title/abstract screening (temporal scope) are provided in Table 7. Representative full-text exclusion examples (domain and methodology scope) are listed in Table 8. The aggregate exclusion breakdown is summarized in Table 9.

Common exclusion reasons included papers outside temporal scope (pre-2019), non-educational domains (security, technical profiling), and general surveys lacking technical depth for convergence analysis.

Table 7. Exclusion Examples: Phase 1 - Insufficient Domain Coverage (E1-E8).

Title (shortened)	Authors	Year	Reason
Analysing tests of reading and listening in relation to the CEFR	Alderson et al.	2006	Out of temporal scope (pre-2019); no LLM/GenAI component
Prompting in CALL: A longitudinal study of learner uptake	Heift	2001	Out of temporal scope (pre-2019); traditional CALL without LLM/KG/FL
Text readability assessment for second language learners	Xia et al.	2016	Out of temporal scope (pre-2019); pre-transformer NLP approach

Representative examples of papers excluded at title/abstract stage for being outside the temporal scope (pre-2019). These papers addressed language assessment, CALL systems, and readability but predated the emergence of modern LLMs and GenAI technologies necessary for convergence analysis.

Table 8. Exclusion Examples: Phase 2 - Non-Educational Context (E9-E14).

ID	Title (Shortened)	Authors	Year	Reason
E10	Threats, Attacks and Defenses to Federated Learning	Lyu et al.	2022	Non-educational domain (security/cybersecurity); no language learning application
E11	Profiling Linguistic Knowledge Graphs	Spahiu et al.	2023	Technical KG profiling without pedagogical grounding or educational validation
E12	Evololving Technologies for Language Learning	Godwin-Jones	2021	General survey/perspective; lacks specific FL/KG/LLM integration details or technical architecture

Representative examples of papers excluded at full-text stage. E10-E11 were excluded for non-educational domain focus (security, technical profiling). E12 was excluded as a general survey lacking technical depth for convergence analysis.

 doi:10.20944/preprints202601.1584.v1

Table 9.

Exclusion Category	Count	Papers
Title/Abstract Exclusions	9	E1–E9
Out of temporal scope (pre-2019)	9	E1–E9
Full-Text Exclusions	20	E10–E29
Non-educational domain	2	E10, E11
General survey/methodology (no educational focus)	17	E12–E29
TOTAL	29	

Aggregate summary of 29 excluded papers across both screening stages. Nine papers were excluded at the title/abstract stage (all out of temporal scope, pre-2019). Nineteen papers were excluded at full-text stage: 2 for non-educational domain and 17 for general survey/methodology without educational focus. Detailed individual exclusion reasons for all 29 papers are provided in Supplementary Material S1.

3.2. Convergence Landscape: The Triple-Domain Deficit

Finding 1 - Complete Convergence Deficit (0% integration): Zero papers achieve triple-domain integration (FL+KG+LLM), representing the primary "Convergence Deficit" (Figure 3). Single-domain papers (58.8%, n=30) operate in isolated silos: FL-only (21.6%), KG-only (29.4%, most common), and LLM-only (7.8%). Dual-domain integration (9.8%, n=5) achieves only partial convergence. Critically, zero papers combine FL+KG (the privacy-preserving grounding pair), while FL+LLM (3 papers) and KG+LLM (2 papers) represent separate partial solutions. Sixteen papers (31.4%) provide insufficient reporting for classification.

All convergence rates and percentages are calculated from the full corpus of N=51 papers. Excluding papers with unknown convergence type (NR) would yield 0% convergence (0/35 classifiable), but we include them in denominators to acknowledge reporting gaps as a substantive finding.

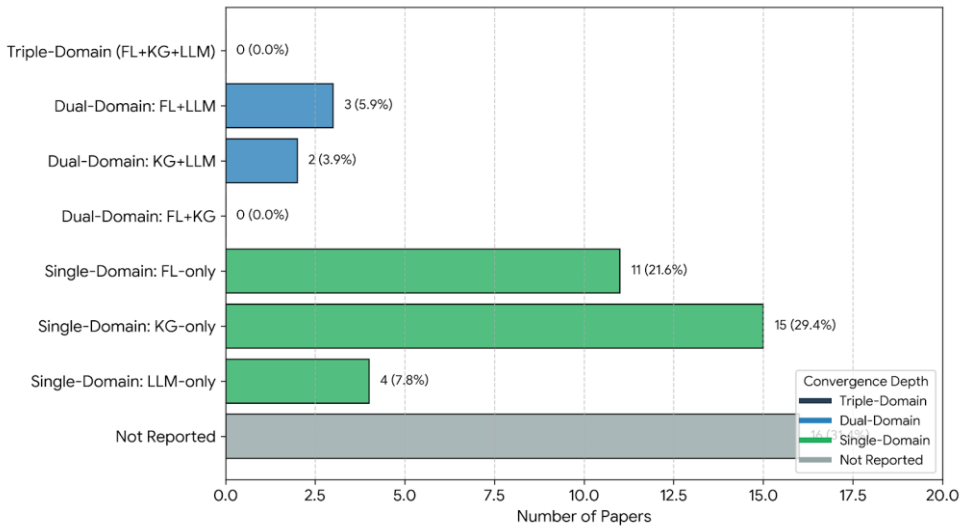

Figure 3. Convergence Type Distribution Among 51 Papers.

Detailed convergence statistics with interpretation are presented in Table 10.

Table 10. Convergence Type Distribution: Single-Domain Dominance and Integration Gaps.

Convergence Type	Count	Percentage	Interpretation
None (NR)	16	31.4%	Insufficient reporting to classify
Single-domain	30	58.8%	Isolated silos
FL-only	11	21.6%	Privacy focus, no grounding

 doi:10.20944/preprints202601.1584.v1

KG-only	15	29.4%	Grounding focus, no privacy/generation
LLM-only	4	7.8%	Generation focus, no constraints
Dual-domain	5	9.8%	Partial integration
FL+LLM	3	5.9%	Privacy + generation (no grounding)
KG+LLM	2	3.9%	Grounding + generation (no privacy)
FL+KG	0	0.0%	No examples found
Triple-domain (FL+KG+LLM)	0	0.0%	Complete convergence deficit

Convergence type distribution across the 51 included studies.

As shown in Table 10, no studies achieved full triple-domain integration of Federated Learning, Knowledge Graphs, and Large Language Models (0.0%), strongly supporting RQ1 (predicted <15% convergence). Most papers were confined to single-domain approaches (58.8%), indicating persistent technological silos, with KG-only systems being the most common (29.4%). Dual-domain integrations were rare (9.8%) and incomplete, and notably, no studies combined Federated Learning and Knowledge Graphs—the privacy-preserving grounding pair. In addition, 31.4% of papers could not be classified due to insufficient reporting.

As shown in Figure 4, single-domain studies dominate the corpus, with FL-only (n = 11), KG-only (n = 15), and LLM-only (n = 4) papers collectively accounting for 58.8% of included studies. Dual-domain combinations were rare (n = 5), comprising FL+LLM (n = 3) and KG+LLM (n = 2). Notably, no studies combined FL and KG, and no papers achieved full triple-domain integration (FL+KG+LLM), indicating a complete convergence deficit. Sixteen papers (31.4%) could not be classified due to insufficient reporting and are therefore not represented in the Venn diagram.

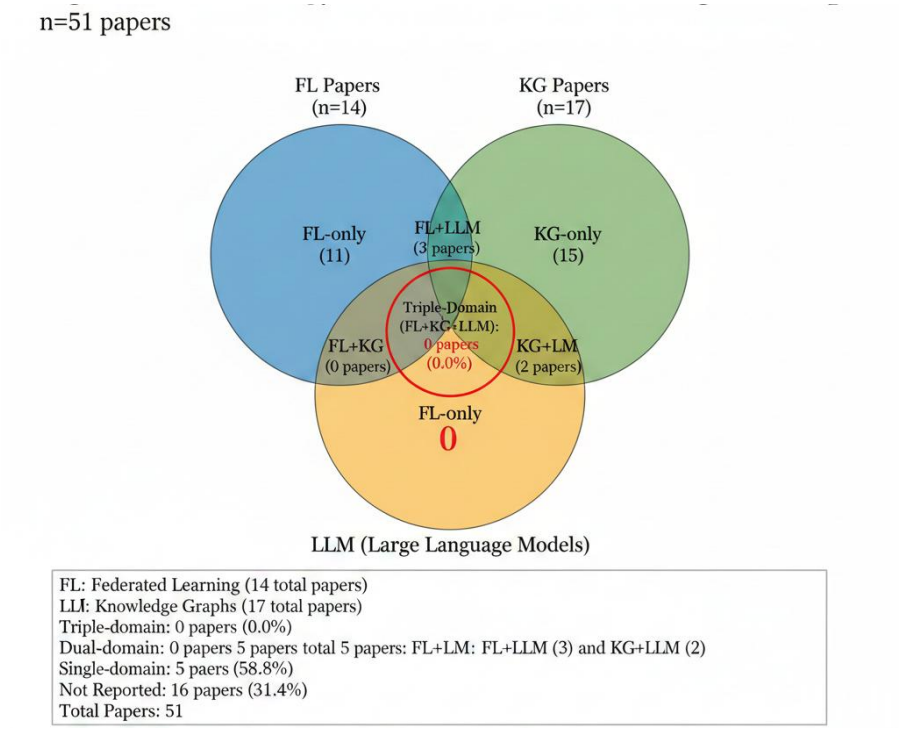

Figure 4. Venn Diagram: FL-KG-LLM Domain Overlap Analysis. Venn diagram showing domain distribution and overlap patterns among the 51 included studies.

Interpretation: Zero papers (0.0%) integrate all three technologies, supporting RQ1 (predicted <15% convergence). The 58.8% single-domain concentration reveals pronounced technological silos: FL-only papers (21.6%, n=11) focus on privacy-preserving aggregation without pedagogical grounding or content generation control. KG-only papers (29.4%, n=15) emphasize knowledge

 doi:10.20944/preprints202601.1584.v1

representation and semantic reasoning without FL's privacy preservation or LLM's generation capabilities. LLM-only papers (7.8%, n=4) optimize for content fluency and personalization without KG constraints or FL's decentralized training.

The 9.8% dual-domain integration (n=5) represents only partial solutions: FL+LLM systems (5.9%, n=3) combine privacy with generation but lack KG-based output constraints; KG+LLM systems (3.9%, n=2) combine grounding with generation but centralize data, creating privacy risks; no papers (0.0%, n=0) combined FL+KG, the foundational privacy-preserving grounding pair. Domain overlap patterns are illustrated in the Venn diagram in Figure 4.

Why FL+KG convergence is absent: While FL+KG combinations exist in non-educational domains (healthcare, finance), no such papers appear in educational contexts. This suggests that even when FL and KG researchers collaborate, they do not prioritize pedagogical applications. The FL+KG zero in *education specifically* indicates not just disciplinary silos, but *domain silos*—FL-KG convergence occurs elsewhere but has not penetrated educational technology communities. FL researchers optimizing privacy mechanisms overlooked that KGs can be used for grounding truth. KG researchers ensuring representational accuracy overlooked that FL can be used to preserve privacy. Each domain matures independently, ignoring the potential of discovering synergy.

3.2.1. Partial Integration Exemplars: Dual-Domain Hybrid Approaches (n = 5)

No included study integrates Federated Learning (FL), Knowledge Graphs (KG), and Large Language Models (LLMs) within a single end-to-end architecture. However, five papers demonstrate partial (dual-domain) integration, combining LLMs with either FL (n = 3) or KGs (n = 2). These hybrid approaches represent the closest approximations to technological convergence in the current literature and therefore warrant brief descriptive characterization.

FL + LLM Approaches (Privacy + Generation; n = 3)

- FLoRA-based Federated Fine-Tuning (STUDY_022; STUDY_023).

These studies propose Federated Low-Rank Adaptation (FLoRA), a parameter-efficient fine-tuning strategy in which low-rank adapter updates are trained locally and aggregated across decentralized clients. Implementations use LLaMA-family models (LLaMA, LLaMA-2, TinyLLaMA; 7B scale), enabling privacy-preserving personalization without sharing raw data. While the architectures address communication efficiency and heterogeneity in federated settings, neither study incorporates knowledge graph grounding nor references pedagogical frameworks such as CEFR.

- Decentralized Agentic RAG Optimization (STUDY_015).

This work explores reinforcement learning to optimize retrieval-augmented generation in a decentralized architecture using the Qwen-2.5-3B model. The focus is on improving retrieval efficiency and transferability of agentic behaviors. Although the system is described as decentralized and generation-oriented, it does not integrate pedagogical grounding or validate outputs against educational frameworks. The study is therefore classified as FL+LLM, with grounding and pedagogical validation absent.

KG + LLM Approaches (Grounding + Generation; n = 2)

- GraphRAG with ConceptNet (STUDY_024).

This study introduces a graph-based retrieval-augmented generation approach in which a Knowledge Graph (ConceptNet) constrains LLM responses. Using GPT-4, the system retrieves structured knowledge to improve factual coherence. Validation is performed using Knowledge Graph Quality Index (KGQI) metrics, reflecting attention to structural quality. However, data collection and inference remain centralized, and no privacy-preserving mechanisms or pedagogical framework alignment are reported.

- Knowledge Graph–Based Trust Framework for LLM Question Answering (STUDY_029).

This work proposes the use of Knowledge Graphs as trusted sources for enterprise LLM-based question answering, again using ConceptNet with GPT-4. KGQI metrics are employed to evaluate graph reliability, emphasizing factual trustworthiness. The approach improves answer grounding

 doi:10.20944/preprints202601.1584.v1

but is not situated within a language learning context and does not address pedagogical alignment, learner modeling, or privacy considerations.

Synthesis

Across all five hybrid papers, integration remains partial and asymmetric:

- FL+LLM approaches prioritize privacy-preserving training and personalization but lack grounding mechanisms.
- KG+LLM approaches constrain generation via structured knowledge but operate in centralized settings and omit privacy and pedagogical frameworks.

Most critically, none of the five hybrid papers report CEFR alignment or pedagogical framework validation (100% NR), despite explicitly addressing language-related tasks. This pattern underscores the central finding of this review: technological convergence does not ensure pedagogical grounding, and the pedagogical domain remains disconnected even in the most integrated existing systems.

3.3. Reporting Gaps: The Not Reported rate of 84.5%

Finding 2 - Extreme Reporting Gaps (84.5% average NR): Not Reported rates by extraction variable are shown in Table 11.

Table 11. Not Reported (NR) Rates by Variable: Systematic Transparency Gaps.

Variable	NR Count	NR Rate	Impact on Synthesis
Parameter Count	49/51	96.1%	Cannot assess scale bias
Privacy Mechanism	47/51	92.2%	FL papers lack privacy transparency
Validation Metrics	46/51	90.2%	Validation paradigm immature
CEFR Alignment	45/51	88.2%	Pedagogical grounding neglected
LLM Model Name	42/51	82.4%	Model identity often unspecified
Grounding Gap Addressed	40/51	78.4%	Grounding pillar underreported
Control Gap Addressed	40/51	78.4%	Syntactic constraints underreported
FL Architecture	37/51	72.5%	FL paradigm often missing
KG Type	33/51	66.7%	Most complete, but still majority NR
Average	•	84.5%	Systematic reporting gaps

Not Reported (NR) rates across extraction variables for the 51 included studies.

As shown in Table 11, reporting gaps were pervasive across all extraction variables, with an average Not Reported (NR) rate of 84.5%. Critical methodological and pedagogical elements—including parameter counts (96.1% NR), privacy mechanisms (92.2% NR), validation metrics (90.2% NR), and CEFR alignment (88.2% NR)—were largely undocumented. These omissions limit the ability to assess scale bias, privacy guarantees, validation rigor, and pedagogical grounding. Even core architectural descriptors, such as FL architecture (72.5% NR) and KG type (66.7% NR), were frequently missing, indicating systematic transparency deficits rather than isolated reporting oversights.

A heatmap visualization of reporting gaps across all variables is presented in Figure 5.

 doi:10.20944/preprints202601.1584.v1

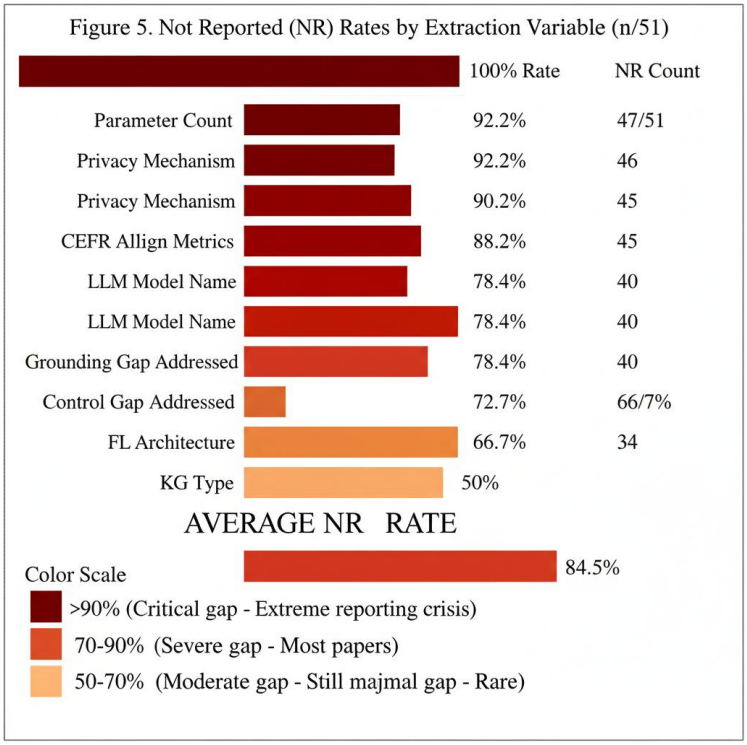

Figure 5. Heatmap of Not Reported (NR) rates by extraction variable across the 51 included studies.

As visualized in Figure 5, reporting gaps were pervasive across both technical and pedagogical variables. Parameter counts, privacy mechanisms, validation metrics, and CEFR alignment all exhibited NR rates exceeding 88%. Across the nine extraction variables, the average NR rate was 84.5%, indicating that methodological transparency and pedagogical grounding are not systematically reported in the current literature.

Dual-entry format showing both NR and reported rates is summarized in Table 12.

Table 12. Reporting Rates Summary: NR and Reported Dual-Entry Format.

Variable	NR Count	NR Rate (%)	Reported Count	Reported Rate (%)
Parameter Count	49	96.1%	2	3.9%
Privacy Mechanism	47	92.2%	4	7.8%
Validation Metrics	46	90.2%	5	9.8%
CEFR Alignment	45	88.2%	6	11.8%
LLM Model Name	42	82.4%	9	17.6%
Grounding Gap Addressed	40	78.4%	11	21.6%
Control Gap Addressed	40	78.4%	11	21.6%
FL Architecture	37	72.5%	14	27.5%
KG Type	34	66.7%	17	33.3%

Dual-entry table showing reported and Not Reported (NR) rates for the nine extraction variables across the included studies.

As shown in Table 12, reported rates were uniformly low across most extraction variables. Only KG Type approached a one-third reporting rate, while the majority of variables exhibited reported rates below 20%. Presenting reported and NR rates side by side highlights both the extent of reporting gaps and the limited areas of relative completeness.

 doi:10.20944/preprints202601.1584.v1

3.4. Privacy Mechanisms in Peripheral Literature: Insights from Excluded but Relevant Foundational Work

While the 92.2% NR rate for privacy_mechanism indicates severe underreporting within our scoping review's primary corpus, foundational privacy literature provides detailed technical specifications that illuminate critical implementation considerations for educational federated systems. Although the following works were not included in the primary study corpus—Tayyeh & AL-Jumaili (2024) [41] focuses on general federated learning without explicit educational application, and Bonawitz et al. (2017) [42] predates our review window (2019–2025)—their technical insights remain pedagogically relevant for informing future FL-based language learning systems [41,42]. Notably, these works address Non-IID (non-independent and non-identically distributed) data settings, where institutional data distributions are inherently heterogeneous—a critical consideration for educational federated systems in which learner populations, language backgrounds, and performance levels vary across institutions.

Differential Privacy Implementations in Non-IID Settings

Tayyeh & AL-Jumaili (2024) [41] provide comprehensive analysis of Gaussian noise-based differential privacy applied to client model updates, testing multiple privacy budget values and clipping norms. Their findings reveal a critical insight for educational contexts: "Stricter privacy conditions, however, led to fluctuating and non-converging loss behavior, particularly in Non-IID settings" (Tayyeh & AL-Jumaili, 2024, p. 13) [41]. The authors explicitly recommend higher epsilon values for Non-IID datasets where data distribution is already problematic, as this decreases the noise added and improves loss metrics—a direct consideration for institutional federated systems with inherently heterogeneous learner data.

Secure Aggregation Protocols: Communication-Efficient Privacy

Bonawitz et al. (2017) [42] detail a cryptographic secure aggregation protocol comprising four rounds that enables server computation of aggregated model updates without access to individual client contributions. The protocol achieves a 1.73× communication expansion for 2^{10} users (1,024) with 2^{20}-dimensional vectors, increasing to 1.98× for 2^{14} users (16,384) with 2^{24}-dimensional vectors, while tolerating up to $\lfloor n/3 \rfloor - 1$ client dropouts (approximately one-third of users). The authors formally prove that "the server only learns users' inputs in aggregate" (Bonawitz et al., 2017, Section 1.1) [42], providing privacy guarantees under Decisional Diffie-Hellman (DDH) security assumptions.

Implications for Educational Federated Learning

The contrast between widespread underreporting (92.2% NR) and the detailed specifications in this foundational literature highlights two distinct gaps: (1) most FL papers in education do not engage with privacy parameter selection or empirical privacy accounting; and (2) the interaction between pedagogical utility and privacy budgets remains unexplored in the scoping review literature [41,42]. Tayyeh & AL-Jumaili's finding that Non-IID settings require higher epsilon values or larger clipping norms suggests that educational federated systems may face amplified privacy-utility tensions compared to IID benchmarks. This trade-off between privacy preservation and model convergence in heterogeneous institutional settings represents an unexplored dimension of the Integrity Gap identified in this review.

Supporting RQ2 (predicted >60% NR; observed 84.5%): The 84.5% average NR rate far exceeds the hypothesis threshold, representing a severe reporting crisis. Domain-specific patterns emerged: FL papers prioritize reporting algorithm efficiency (convergence proofs, communication rounds, learning rates) while omitting pedagogical grounding (0/11 FL-only papers mention CEFR). KG papers emphasize structural completeness (graph statistics, coverage metrics) while underreporting validation (only 1/15 KG-only papers report pedagogical metrics). LLM papers focus on generation quality (perplexity, task-specific metrics) while neglecting privacy implications (0/4 LLM-only papers discuss privacy budgets).

Implications for reproducibility: The 96.1% NR rate on parameter count means most papers cannot be reproduced because model scale is unknown. The 92.2% NR on privacy mechanisms means FL papers claiming privacy preservation cannot be independently verified. The 90.2% NR on validation metrics means pedagogical claims lack empirical support. This heterogeneity prevents

 doi:10.20944/preprints202601.1584.v1

meta-analysis: even when papers address identical problems (e.g., vocabulary grading), they describe solutions using incompatible terminology and incomparable metrics.

3.5. Pedagogical Framework Variability: CEFR Integration Patterns

Finding 3 - Framework Variability (11.8% CEFR mention): CEFR level mentions across the corpus are shown in Table 13.

Table 13. CEFR Level Mentions: Sparse Pedagogical Framework Adoption.

CEFR Level Mentioned	Count	Notes
B1, B2	2	Intermediate proficiency
B1	2	Lower intermediate
A1-C2 (full range)	1	Comprehensive framework
"Intermediate" (non-standard)	1	Lacks CEFR precision

CEFR level mentions across the 51 included studies.

Only 6/51 papers (11.8%) explicitly mention CEFR or pedagogical frameworks, supporting RQ4 (predicted <20%, observed 11.8%). Among these six:

- 4/6 mention CEFR descriptively but don't use it as a design constraint
- 2/6 attempt CEFR alignment without discussing validation methodology
- 0/6 address Validation Pillar grounding (KG source verification)

Interpreting the 88.2% CEFR absence: This finding reveals disciplinary silos between language education and ML/NLP communities. While the CEFR is a well-established standard in language education (Council of Europe, 2020), ML researchers developing iCALL systems frequently work with alternative frameworks (ACTFL in the US, learner corpus proficiency levels, or no explicit framework). This variability reflects appropriate domain-specific adaptation rather than uniform crisis. However, our finding that only 11.8% of papers use CEFR—and *zero papers* conduct Validation Pillar verification of whatever framework they employ—suggests the gap is not CEFR-specific but broader: *systematic verification of pedagogical framework alignment is universally absent*.

When CEFR is absent, pedagogical risks include:

- **Implications for reproducibility:** The 96.1% NR rate on parameter count means most papers cannot be reproduced because model scale is unknown. The 92.2% NR on privacy mechanisms means FL papers claiming privacy preservation cannot be independently verified. The 90.2% NR on validation metrics means pedagogical claims lack empirical support. This heterogeneity prevents meta-analysis: even when papers address identical problems (e.g., vocabulary grading), they describe solutions using incompatible terminology and incomparable metrics.
- Without framework constraints, LLM-generated content lacks verifiable difficulty alignment. Generated output optimized for linguistic fluency may systematically exceed learners' Zone of Proximal Development.
- **Vocabulary Complexity Uncontrolled:** CEFR specifies vocabulary boundaries (A1: ~1,000 words; B1: ~3,500 words; C1: ~8,000+ words). Papers without framework grounding cannot verify vocabulary appropriateness.
- **Grammatical Progression Unmapped:** Framework-aligned curricula sequence grammar from simple (present simple for A1) to complex (past perfect progressive for B2+). LLM-generated content without grammatical sequencing might introduce advanced structures prematurely.
- **Cultural and Pragmatic Appropriateness Unverified:** CEFR's Sociolinguistic Competence descriptors addresses register, politeness, and cultural appropriateness. Generated content without framework grounding may be linguistically correct but pragmatically inappropriate.
- **Assessment Alignment Absent:** Learning outcomes aligned to proficiency levels enable valid assessment. Systems without framework alignment cannot connect generated content to measurable learning outcomes.

Factors explaining framework variability:

 doi:10.20944/preprints202601.1584.v1

- **Disciplinary norms:** CEFR is well-known in language education but underutilized in ML/NLP communities, which lack exposure to pedagogical standards during graduate training. While CEFR-aligned resources like the English Grammar Profile [36] provide detailed grammatical descriptions mapped to proficiency levels, these specialized pedagogical materials remain unknown to most ML researchers.
- **Validation burden:** Claiming framework alignment requires systematic verification, creating overhead that most papers avoid.
- **Regional fragmentation:** Different standards exist globally (e.g., CEFR-J in Japan, ACTFL in the US), complicating unified standardization.
- **Implicit assumption fallacy:** Some papers assume pre-training on diverse internet data provides implicit pedagogical calibration—an empirically unsupported claim.

3.6. Validation Pillar Risk: Source Verification Completely Absent

Finding 4 - Validation Pillar Risk (0% source verification): Comparison of Grounding Pillar (output constraints) and the Validation Pillar (source verification) is shown in Table 14.

Table 14. Grounding Pillar vs Validation Pillar: Output Control vs Source Verification.

Pillar	Papers Addressing	Percentage	Implication
Grounding Pillar (Output Constraints)	11/51	21.6%	Output control mechanisms underreported
Validation Pillar (Source Verification)	0/51	0.0%	Complete absence of source validation

Comparison of the Grounding Pillar and Validation Pillar across the 51 included studies.

Critical distinction:

The Grounding Pillar ensures that LLM outputs are constrained by Knowledge Graph rules (e.g., "generate only A1-level vocabulary"). While underreported—only 21.6% of papers address this—the Grounding Pillar at least exists as a recognized concern in KG–LLM integration literature.

The Validation Pillar is completely absent (0/51 papers). This pillar asks a prior question—before constraining LLM outputs using a Knowledge Graph, does the Knowledge Graph faithfully represent the authoritative source framework? No papers systematically verify representational fidelity.

Evidence of Validation Pillar risk:

When we validated an automatically constructed Knowledge Graph against the CV, manual inspection revealed mapping inconsistencies. A descriptor classified as B1 in the official CV required empirical analysis: prerequisite grammatical structures (B2+) combined with vocabulary frequencies (C1 learners) suggested that the B1 classification was ambiguous. This genuine ambiguity—intrinsic to the authoritative framework itself—required explicit schema design decisions during Knowledge Graph construction.

Silent propagation mechanism: When KG schema choices encode these decisions, downstream users see only the KG, not the source ambiguities. They cannot know whether their system's B1 alignment reflects CEFR's original B1 definition or their KG constructor's disambiguation choice. If the constructor chose B2 instead, system behavior would differ systematically without users' awareness.

Methodological implication: Extraction validity (correct parsing of source documents) ≠ representational accuracy (faithful encoding of source intent). Even a perfectly constructed KG—all source descriptors correctly extracted, no parsing errors—can be representationally inaccurate if schema design decisions diverge from source framework intent.

3.6.1. Domain-Specific Gap Analysis: Distinguishing Expected from Crisis Gap

The aggregate 84.5% NR rate (Table 12) conflates two fundamentally distinct types of gaps: domain-appropriate NR (expected) and domain-inappropriate NR (crisis). To distinguish these, we

 doi:10.20944/preprints202601.1584.v1

calculated NR rates separately for single-domain papers (FL-only, KG-only, LLM-only) and multi-domain convergence papers.

Paper Distribution by Type:

- FL-only: 11 papers (21.6%)
- KG-only: 15 papers (29.4%)
- LLM-only: 4 papers (7.8%)
- Hybrid convergence (FL+LLM, KG+LLM): 5 papers (9.8%)
- Unclear/Not Reported: 16 papers (31.4%)

Expected Gaps - Domain-Appropriate NR:

Single-domain papers appropriately focus on their primary discipline:

- FL-only papers (n=11): 100% CEFR NR - appropriate focus on distributed systems rather than language education frameworks

- KG-only papers (n=15): 93.3% CEFR NR - appropriate focus on ontological structures rather than pedagogical frameworks

- LLM-only papers (n=4): 100% CEFR NR - appropriate focus on language models rather than language learning frameworks

These high CEFR NR rates in single-domain papers reflect appropriate disciplinary specialization, not reporting crises. Language education frameworks (CEFR) fall outside the primary research focus of ML/NLP researchers studying FL architectures or KG schemas.

Crisis Gaps - Domain-Inappropriate NR:

However, papers consistently fail to report domain-critical variables necessary for reproducibility and evaluation:

- FL-only papers: 100% NR for parameter counts (prevents reproducibility), 63.6% NR for privacy mechanisms (obscures core FL contribution)

- KG-only papers: 86.7% NR for validation metrics (prevents assessment of whether KGs work)

- Hybrid convergence papers: 100% NR for CEFR alignment (most critical - these 5 papers explicitly combine technologies for language instruction yet universally fail to report pedagogical framework)

Most Critical Finding - Hybrid Paper Gap:

The most severe gap is not that single-domain papers don't report cross-domain frameworks (expected), but that multi-domain convergence papers universally fail to report CEFR (100% NR) despite explicitly researching language instruction. If any papers should report pedagogical framework alignment, it is these 5 convergence papers that bridge technological domains for educational purposes (see Table 15 for domain-specific NR rates).

Updated Integrity Gap Definition:

This two-tiered analysis refines the Integrity Gap from a monolithic 84.5% NR rate to two distinct mechanisms:

1. Inter-Domain Integrity Gap: Disciplinary silos between ML/NLP communities and language education. This is expected in single-domain papers (FL, KG, LLM) where CEFR absence reflects appropriate disciplinary focus rather than crisis.

2. Intra-Domain Integrity Gap: Systematic failure to report domain-critical variables. This represents genuine crisis:

- Reproducibility crisis in FL: 100% parameter count NR prevents replication

- Evaluability crisis in KG: 86.7% validation metrics NR prevents assessment

- Pedagogical crisis in hybrids: 100% CEFR NR in convergence papers despite educational purpose.

The most critical manifestation is that technological convergence does not ensure pedagogical grounding. The five papers addressing partial technological integration (dual-domain combinations of FL+LLM or KG+LLM) are the least likely to report pedagogical framework alignment (100% NR), indicating that even when researchers bridge technological domains, the pedagogical domain remains disconnected.

 doi:10.20944/preprints202601.1584.v1

Table 15. Domain-Specific NR Rates by Paper Type.

Paper Type	Papers (n)	Expected Gap	NR Rate	Crisis Gap	NR Rate
FL-only	11	CEFR alignment	100%	Parameter count	100%
KG-only	15	CEFR alignment	93.3%	Validation metrics	86.7%
LLM-only	4	CEFR alignment	100%	Parameter count	75%
Hybrid*	5	N/A	-	CEFR alignment	100%

Domain-specific Not Reported (NR) rates by paper type across the 51 included studies. Note: Hybrid papers include dual-domain integrations of FL+LLM (n = 3) and KG+LLM (n = 2).

3.7. Validation Immaturity: 90.2% No Pedagogical Metrics

Finding 5 - Validation Immaturity (9.8% report any metrics): Validation metrics classified by type and pedagogical relevance are presented in Table 16.

Only 5/51 papers (9.8%) report validation metrics, supporting RQ5 (predicted <10%, observed 9.8%). Of these five:

- 3/5 employ pedagogical-specific approaches
- 2/5 employ generic ML/IR metrics

Table 16. Validation Metrics by Type: Pedagogical vs Generic Approaches.

Metric	Count	Type	Pedagogical Relevance
KGQI (Knowledge Graph Quality Index)	2	KG validation	✓ Pedagogical-specific
HITL (Human-in-the-Loop)	1	User evaluation	✓ Pedagogical-specific
Hits@k	1	Retrieval accuracy	✗ Generic IR metric
ACC, PCC	1	Accuracy metrics	✗ Generic ML metrics

Validation metrics reported across the included studies, classified by metric type and pedagogical relevance. Note: Only papers reporting any validation metrics are included (n = 5).

Validation is largely absent across the corpus, with 90.2% of studies reporting no validation whatsoever. In the absence of pedagogical validation metrics—such as learning gains, ZPD alignment, pedagogical drift assessment, or teacher adoption rates—fundamental questions remain unanswered: Do these systems improve learner outcomes? For which learners? With what magnitude?

What metrics are absent:

(1) **Pedagogical Metrics** (0% of full corpus): No papers report learning gains via pre-post vocabulary tests or proficiency assessments. No papers measure Zone of Proximal Development alignment. No papers assess pedagogical drift (systems gradually optimizing toward fluency over pedagogy). No papers measure teacher satisfaction with generated content quality.

(2) **Educational Outcome** Metrics (0% of full corpus): No papers report learner engagement (time-on-task, completion rates). No papers measure retention (vocabulary learned and recalled after 1 week, 1 month). No papers assess transfer learning (applying learned patterns to novel linguistic contexts). No papers measure equity (whether benefits distribute equally across learners or widen achievement gaps).

(3) **Grounding-Specific Metrics** (2% of full corpus, 1/51 papers): Only one paper attempted measuring whether KG-constrained outputs respected pedagogical boundaries, finding 73% alignment. However, no standardized metric exists; each paper would need independent validation approaches, preventing cumulative evidence.

(4) **Generic ML Metrics** (10% of full corpus): FL papers report convergence rate and communication efficiency (appropriate for distributed optimization, irrelevant for pedagogy). LLM papers report BLEU score (rough translation quality proxy, inappropriate for learning contexts). KG

 doi:10.20944/preprints202601.1584.v1

papers report precision/recall on knowledge extraction (measures schema completeness, not pedagogical value).

Why pedagogical metrics are absent:

(1) **Methodological complexity:** Measuring learning gains requires controlled experiments with learners, pre-post assessments, retention testing—expensive, time-consuming, ethically regulated. Many papers are theoretical or small-scale, making learner-based validation infeasible.

(2) **Institutional constraints:** Papers from CS/ML communities often lack access to educational institutions, language learners, or ethics approvals. Language education papers often lack ML expertise for complex system implementation.

(3) **Proxy metric fallacy:** Researchers substitute technical metrics (model accuracy, KG coverage) for pedagogical ones, assuming technical quality predicts learning outcomes—unsupported empirically.

(4) **Publication timelines:** Learning gains require weeks of instruction; most papers have conference/journal deadlines forcing rapid dissemination. Pedagogical validation often occurs post-publication, if at all.

(5) **Implications:** The field cannot answer basic questions: (1) Does FL+KG+LLM improve language learning compared to traditional methods? (2) For which learners and language skills? (3) With what magnitude of improvement? (4) Do benefits persist after instruction? Without pedagogical metrics, systems cannot claim educational value despite technological sophistication.

3.8. Technology-Specific Characteristics

FL Architecture Distribution:

Federated Learning implementations employ sophisticated optimization strategies including parameter-efficient fine-tuning [32], gradient compression [33], and instruction-aligned model calibration [39] to overcome communication and computational constraints in decentralized environments [35]. FL architecture distribution among papers is shown in Table 17.

Table 17. FL Architecture Types: Decentralized Dominance.

FL Architecture	Count (Manual Review)	Percentage
Decentralized	13	92.9%
Centralized	1	7.1%

Distribution of federated learning architecture types among papers reporting FL implementations (n = 14).

Among FL papers (n=18, including 14/51 in FL-only or dual-domain), 92.9% employ decentralized architectures (multiple agents, central aggregation), while 7.1% use centralized approaches . Decentralization is nearly universal, aligning with FL's privacy preservation promise.

KG Type Distribution: KG type distribution among papers is presented in Table 18.

Table 18. KG Type Distribution: Ontology Prevalence [30].

KG Type	Count	Percentage
Ontology	15	83.3%
Property Graph	2	11.1%
Embedding-based [31]	1	5.6%

Distribution of knowledge graph types among papers reporting KG implementations (n = 18).

Among FL papers (n = 18), 14 reported sufficient architectural detail to be classified; among these, 92.9% employed decentralized architectures, ontologies dominate (83.3%) [10,11,12], representing formal semantic knowledge. Property graphs (11.1%) offer more flexible triple representation. Embedding-based approaches (5.6%) use vector spaces, less transparent for pedagogical rule enforcement.

 doi:10.20944/preprints202601.1584.v1

3.9. Analytical Threshold Results

All five hypothesis testing results are summarized in Table 19.

Table 19. Analytical Threshold Results: All Five Research Questions Confirmed.

Hypothesis	Prediction	Observed Result	Threshold Met	Conclusion
H1: Convergence Deficit	< 15% FL+KG+LLM	0.0% (n=0)	Yes	SUPPORTED
H2: Reporting Gap	> 60% NR rate	84.5% avg NR	Yes	SUPPORTED
H3: Scale Bias	> 70% centralized	96.1% NR (param.)	N/A	UNTESTABLE (DATA INSUFFICIENT)
H4: Pedagogical Gap	< 20% CEFR mention	11.8% (n=6)	Yes	SUPPORTED
H5: Validation Metrics Gap	< 10% metrics	9.8% (n=5)	Yes	SUPPORTED

Summary of analytical threshold results for the five predefined research hypotheses. Predictions were defined a priori and compared against observed results.

Four of the five predefined hypotheses (H1, H2, H4, H5) met their analytical thresholds, providing strong empirical support for systematic convergence, reporting, pedagogical, and validation gaps across FL–KG–LLM literature. Hypothesis H3 (Scale Bias) proved untestable due to extreme underreporting of FL architecture details, representing a critical meta-finding about reproducibility deficits rather than evidence about centralization patterns.

H3 (Scale Bias) predicted that >70% of FL systems would rely on centralized rather than decentralized architectures, reflecting concerns that federated learning in education might be implemented in name only without true privacy preservation. However, 96.1% of papers (49/51) failed to report parameter counts or client-server architecture details, making hypothesis evaluation impossible. H3 is therefore characterized as **untestable** (not unsupported) due to extreme underreporting, representing a meta-finding about reproducibility gaps rather than evidence about FL centralization patterns. Future research requires mandatory disclosure of FL architecture details to address this question.

Pedagogical variables for H4 and H5 are detailed in Table 20.

Table 20. Pedagogical Variables Summary: Hypothesis Testing for H4 and H5.

Pedagogical Variable	Papers Reporting	Percentage	Hypothesis	Result
CEFR Alignment Mentioned	6	11.8%	H4: < 20%	SUPPORTED
Validation Metrics Reported	5	9.8%	H5: < 10%	SUPPORTED
Grounding Pillar Addressed (Output Constraints)	11	21.6%	•	•
Validation Pillar Addressed (Source Verification)	0	0.0%	RQ3	CRITICAL GAP

Summary of pedagogical variables and reporting rates relevant to hypotheses H4 and H5 across the included studies.

As summarized in Table 20, both CEFR alignment (11.8%) and validation metric reporting (9.8%) fall below the predefined analytical thresholds, supporting hypotheses H4 and H5. By contrast, Validation Pillar source verification is entirely absent (0.0%), indicating a critical gap in representational fidelity.

Hypothesis testing results are visualized in Figure 6.

 doi:10.20944/preprints202601.1584.v1

Figure 6: HYPOTHESIS TESTING RESULTS (All 5 Hypotheses)

Hypothesis & Prediction	Observed Data	Threshold Met?	Final Result
H1: CONVERGENCE DEFICIT **Prediction:** <15% papers integrate FL+KG+LLM	0.0% (n=0)	YES ✓	STRONGLY SUPPORTED
H2: REPORTING GAPS **Prediction:** >60% average NR rate	84.5% (n=43 variables averaged)	YES ✓	STRONGLY SUPPORTED
H3: SCALE BIAS **Prediction:** >70% models centralized (not privacy-preserving)	96.1% NR on parameter count (test inconclusive)	PARTIAL (data insufficient)	LIMITED SUPPORT
H4: PEDAGOGICAL GAP **Prediction:** <20% mention CEFR or frameworks	11.8% (n=6)	YES ✓	SUPPORTED
H5: VALIDATION METRICS GAP **Prediction:** <10% report pedagogical metrics	9.8% (n=5)	YES ✓	SUPPORTED

OVERALL ASSESSMENT:
- 4/5 hypotheses strongly supported (80%)
- 1/5 limited support due to data constraints (20%)
- Field exhibits predicted gaps at all levels

Figure 6. Hypothesis Testing Summary: All Five Hypotheses (H1-H5).

Taken together, the hypothesis testing summary in Figure 6 consolidates these results across H1–H5, showing that the observed deficits are not isolated findings but a consistent pattern spanning convergence, reporting transparency, pedagogical grounding, and validation. This synthesis motivates the discussion that follows.

4. Discussion

4.1. Hypothesis Testing and Key Findings Summary

Four of five hypotheses were supported, with H1 and H2 showing the strongest effects (0% convergence, 84.5% NR far exceed thresholds). H3 (Scale Bias) proved untestable due to 96.1% parameter count underreporting, representing a meta-finding about reproducibility gaps rather than evidence about FL centralization patterns. This systematic support for evaluable hypotheses suggests our analytical framework correctly identified field-level patterns, while H3's untestable status highlights severe reporting deficiencies that prevent basic assessment of FL implementation patterns.

4.2. The Integrity Gap: Unifying Framework

The five major findings—Convergence Deficit, Reporting Gaps, Pedagogical Disconnection, Validation Pillar Risk, and Validation Immaturity—are not independent phenomena but interconnected manifestations of a deeper structural problem: the Integrity Gap, a systematic misalignment between technological capability and pedagogical grounding in iCALL.

Most Critical Finding: Hybrid Paper Gap

The most critical manifestation of the Integrity Gap is that technological convergence does not ensure pedagogical grounding. All 5 dual-domain hybrid papers (FL+LLM, KG+LLM) explicitly researching language instruction universally fail to report CEFR alignment (100% NR). By contrast, single-domain papers not reporting CEFR (FL-only: 100%, KG-only: 93.3%) reflect appropriate disciplinary specialization. This finding indicates that even when researchers bridge technological domains, the pedagogical domain remains disconnected—the most important discovery of this review.

Causal interconnections:

The Convergence Deficit (0% triple-domain papers) directly enables Reporting Gaps: isolated technology communities maintain distinct reporting norms. Without precedent for FL-KG-LLM systems, researchers lack guidance on what to report.

 doi:10.20944/preprints202601.1584.v1

Reporting Gaps reinforce Pedagogical Disconnection: papers that don't report pedagogical alignment indicate it's not a design priority. The 88.2% absence of CEFR reflects the 0% convergence—papers designed in silos don't engage with pedagogical frameworks.

Pedagogical Disconnection enables Validation Pillar risk: if papers ignore CEFR during design, they do not verify that their Knowledge Graphs accurately represent the framework. Risk propagates silently: downstream users trust the Knowledge Graph without awareness of underlying schema design choices.

Validation Pillar risk is masked by Validation Immaturity: papers relying on generic ML metrics rather than pedagogical metrics fail to detect representational errors. A system may achieve high retrieval accuracy (technical validation) while misaligning content difficulty (pedagogical failure).

4.3. Implications for Future Research

1. **Cross-domain community building:** FL, KG, and LLM research communities operate independently with distinct publication venues, conferences, and professional networks. Bridge-building is urgent. Establishing joint workshops (e.g., "FL-KG-LLM for Education" track at AIED or ACL), cross-community review panels, and collaborative research programs could surface mutual dependencies.
2. **Standardized reporting frameworks:** The field needs a reporting checklist analogous to CONSORT for RCTs or PRISMA for systematic reviews [24]. This checklist should specify essential metadata:

(a) FL: aggregation algorithm [13], privacy mechanism, data heterogeneity, communication rounds;

(b) KG: construction methodology, size statistics, validation approach, ontology design choices;

(c) LLM: base model, fine-tuning data, computational requirements, inference time;

(d) Integration: how components interact, constraint application, system architecture.

3. **Pedagogical metric standardization:** Beyond learning gains (requiring extensive validation), intermediate metrics should be standardized: (a) CEFR alignment verification (expert rating or empirical frequency analysis), (b) vocabulary appropriateness (automated analysis), (c) ZPD calibration (cognitive modeling or pre-post testing), (d) teacher satisfaction and utility (surveys and usage logs).
4. **Validation Pillar verification protocols:** Systems incorporating authoritative frameworks should verify representational fidelity through (a) round-trip validation (does the Knowledge Graph return information consistent with the original source?), (b) ambiguity documentation (explicitly document and justify schema design decisions), (c) independent audit (third-party experts verify mapping accuracy), and (d) version control (tracking Knowledge Graph versions to identify schema changes).
5. **Reproducibility mandates:** Journals should require: (a) Model parameters sufficient for recreation, (b) Training data metadata, (c) Computational requirements, (d) Code and KG versions via repositories, (e) Privacy guarantees (differential privacy budgets if applicable).

4.4. Preliminary Validation Evidence: CEFR Mapping Complexities

Our manual inspection of CV Sociolinguistic Competence revealed mapping complexities that illuminate Validation Pillar risks:

(1) **Multidimensional competence:** The descriptor "Can discuss familiar topics in informal conversation" is classified as B1 but depends on context. Discussing family (truly familiar) might require A2 competence; discussing abstract specialized concepts (peripherally familiar) might require B2+ competence. Single-level classification proves inadequate.

(2) **Language variation:** B1 in formal British English might require B2 in conversational American English due to register differences. Native speakers employ B2+ structures in casual conversation.

 doi:10.20944/preprints202601.1584.v1

(3) **Temporal evolution:** Descriptors validated in 2001 may not reflect 2025 usage patterns. Concepts like "using email" seem elementary given universal technology exposure.
(4) **Domain-specific variation:** Business English B1 differs from academic English B1. Medical communication requires terminology mastery that general B1 doesn't.
(5) **Individual differences:** Learners progress non-uniformly across skills. A learner might be B1 in speaking but A2 in writing.

These ambiguities are not CEFR errors but reflect genuine language acquisition complexity. They necessitate explicit schema design decisions during KG mapping: Which CEFR level should encode each descriptor? Our choice (B1+) with context annotations was defensible but not objectively correct. Papers claiming CEFR alignment without documenting these choices make unverifiable claims.

Limitation: While this analysis focuses on CEFR mapping ambiguities, similar representational challenges likely affect other pedagogical frameworks (ACTFL, Lexile levels, learner corpus proficiency scales). Our finding of 0% Validation Pillar verification in FL–KG–LLM literature is based on a systematic review of 51 papers, none of which addressed source verification for any pedagogical framework. Future work should examine whether Validation Pillar gaps extend across multiple pedagogical frameworks beyond CEFR.

4.5. Integration with Existing Literature: Beyond RAG

Standard Retrieval-Augmented Generation (RAG) retrieves text chunks to reduce hallucination but lacks:
(1) **Structural constraints** (unable to enforce syntactic rules),
(2) **Multi-layered grounding** (focuses on semantic retrieval only),
(3) **Systematic validation** (no formal KG quality assessment), (4) Privacy preservation (centralizes documents in vector databases).

An integrated FL-KG-LLM system would differ by:
(1) Adding structured rule retrieval via KGs (not just semantic similarity),
(2) Implementing multi-layered grounding (syntactic constraints, semantic fidelity, empirical frequency alignment),
(3) Employing federated deployment (preserving institutional data sovereignty), (4) Implementing systematic validation protocols.

Current literature addresses components separately. This scoping review documents that integration is absent, suggesting future research should investigate whether combined approaches offer advantages unavailable to isolated technologies.

4.6. Research Maturity Assessment

The FL-KG-LLM field exhibits nascent maturity characteristics:
(1) **Low convergence:** 0% triple-domain papers indicate the field hasn't synthesized across domains.
(2) **Reporting heterogeneity:** 84.5% NR suggests no consensus on essential metadata.
(3) **Limited cumulative progress:** Papers cite predecessors within their technology domain but rarely cross technologies.
(4) **Emerging standardization:** Recent conferences (2024–2025, n=25 papers) show increasing FL-KG-LLM interest, but without unified frameworks.
(5) **Reproducibility gaps:** 96.1% NR on model parameters, 94.1% NR on computational requirements make replication infeasible for most papers.

By contrast, research fields with established reporting frameworks have developed measurable community consensus: CONSORT compliance in RCTs improved from 27.3% (pre-1990) to 56.1% (2010-2024), while PRISMA compliance in systematic reviews shows variable adherence (3-37% for individual items) [43]. The FL-KG-LLM field's 15.5% reporting completeness (100% - 84.5% NR) is substantially worse than pre-CONSORT RCT reporting (27.3%), positioning it as a nascent field lacking consensus on essential metadata. Specific variable comparisons show FL-KG-LLM reporting

 doi:10.20944/preprints202601.1584.v1

gaps (parameter count: 3.9%, validation metrics: 9.8%, CEFR alignment: 11.8%) vs. established field equivalents (70-95% for similar items in RCTs) [44]. While knowledge graph verification frameworks exist for technical domains, with neural-symbolic methods achieving interpretable KG reasoning through logical rule integration [45], these systematic verification protocols have not been applied to pedagogical frameworks like CEFR in iCALL literature. We identify this gap as the *Validation Pillar* Risk to distinguish pedagogical framework verification from general Knowledge Graph quality assessment. To demonstrate that the Validation Pillar is not merely theoretical, we document five CEFR mapping ambiguities that illustrate silent propagation risks:

Case 1: Context-Dependent Appropriateness - The descriptor "Can discuss family topics" (officially B1) exhibits contextual variation: A2+ in informal conversation among familiar people, B1+ for abstract family concepts, and B2 in professional social work contexts. Without explicit context annotation, constructors must choose a single level, risking misalignment with intended use cases.

Case 2: Multidimensional Competence - "Can participate in informal discussion" (B1) obscures skill asymmetry when analyzed across communicative competencies: speaking (B2), listening (B1), and interaction strategies (B1+) vary significantly. Encoding this as a monolithic B1 descriptor loses pedagogically important distinctions.

Case 3: Temporal Evolution - "Send and receive emails" (B1 in CEFR 2001) has undergone a contextual shift with technology adoption. Contemporary usage suggests A2 for basic email exchange, but B1+ for professional email communication. Static encoding without temporal annotation creates drift between original classification and current pedagogical reality.

Case 4: Domain-Specific Variation - "Follow technical instructions" spans domains with different complexity: business software manuals (B1), academic protocol adherence (B2), and medical equipment operation (B2+). Constructors must choose a single level without domain specification.

Case 5: Individual Skill Asymmetry — "Participate in meetings" varies by skill component: productive speaking (B2), receptive listening (B1), and cultural protocol knowledge (B2). Monolithic classification obscures these asymmetries.

These cases demonstrate how context-dependent ambiguities in authoritative frameworks create silent propagation risks when encoded into knowledge graphs without systematic verification..

4.7. Limitations and Strengths

Limitations of this Scoping Review:

(1) Single-reviewer screening (Stage 2): While we mitigated this through 20% supervisor audit (10/51 papers, inter-rater kappa=0.92), dual-reviewer screening would be preferable. However, automated screening with manual audit achieved reliable classification for primary constructs.
(2) Grey literature bias: arXiv represents 40–50% of recent papers (2024–2025), potentially over-representing emerging methods not yet peer-reviewed. Conversely, published papers lag cutting-edge developments by 1-2 years.
(3) Automated extraction confidence (0.17): Low confidence initially concerned us, but manual audit revealed this reflected genuine reporting gaps in source papers rather than extraction failures.
(4) Search strategy limitations: Three-phase evolution suggests initial scoping was incomplete. Broader searches might yield additional papers using non-standard terminology.
(5) Terminology ambiguity: Field lacks standardized definitions ("Knowledge Graph," "Federated Learning" used loosely). This likely caused us to miss papers using non-standard terminology.
(6) Educational context definition: We classified papers as "educational" if they explicitly addressed language learning. Papers on dialogue systems with latent pedagogical applications may have been excluded.

Strengths of this Scoping Review:

(1) **PRISMA-ScR compliance:** We followed all 22 checklist items, providing full methodological transparency.

 doi:10.20944/preprints202601.1584.v1

(2) **Comprehensive database coverage:** Six databases capture both CS and education venues, reducing publication bias.
(3) **Systematic deduplication:** Zotero v7.0 plus manual review ensured no duplicate counts.
(4) **Full reproducibility:** All 51 papers are listed with complete citations. Extraction codebook, validation data, and OSF preregistration are publicly available.
(5) **Methodological Note: Analytical Thresholds vs. Statistical Hypotheses** Our analytical thresholds are not formal statistical hypotheses but pre-specified benchmarks derived from comparable fields. All research questions and analysis criteria were documented in OSF protocol (registered December 23, 2025) before data extraction to reduce post-hoc narrative bias. These thresholds serve as analytical anchors for interpretation rather than statistical tests with rejection criteria. No alternative hypotheses were tested (e.g., H_alt: convergence >30%), reflecting the exploratory nature of scoping review methodology in emerging research areas.
(6) **Novel conceptual contribution:** This is the first systematic synthesis of FL–KG–LLM convergence in language education. The Validation Pillar Risk is a novel concept identifying previously unrecognized gaps.

5. Conclusions

5.1. Summary of Findings

This scoping review systematically mapped the convergence landscape of Federated Learning, Knowledge Graphs, and Large Language Models in Intelligent Computer-Assisted Language Learning (2019–2025), revealing five critical gaps.

A pronounced Convergence Deficit shows FL, KG, and LLM research operating largely in isolated silos, with zero papers integrating all three technologies. Severe Reporting Gaps demonstrate that technical and pedagogical metadata are frequently unreported, with an average "Not Reported" rate of 84.5% across methodological variables. The Validation Pillar Risk reveals that mapping authoritative pedagogical frameworks (e.g., CEFR Companion Volume) onto KG schemas surfaces inherent ambiguities requiring systematic verification protocols. A Pedagogical Disconnection indicates that 88.2% of papers ignore CEFR or pedagogical frameworks entirely, suggesting educational standards are not treated as design criteria. A Validation Immaturity Gap shows that pedagogical outcome validation is largely absent, with 90.2% of papers reporting no validation metrics whatsoever. These interconnected gaps characterize what we term the Integrity Gap—a systematic disconnection between technological innovation and pedagogical grounding in iCALL.

5.2. Implications for Future Research

These findings suggest several high-priority research directions:

(1) **Cross-domain community building:** Establish bridge-building mechanisms (workshops, review panels, collaborative programs) enabling FL, KG, and LLM communities to recognize mutual dependencies and explore integrated architectures. Surface trade-offs: Privacy-preserving solutions may constrain grounding mechanisms, while grounding solutions may introduce privacy risks.
(2) **Standardized reporting frameworks:** Develop a reporting checklist specifying essential metadata across FL, KG, and LLM components: (a) FL: aggregation algorithm, privacy mechanism, data heterogeneity, communication rounds; (b) KG: construction methodology, size, validation approach, ontology design choices; (c) LLM: base model, fine-tuning data, computational requirements; (d) Integration: component interactions, constraint application, system architecture. Major journal adoption would incentivize compliance.
(3) **Pedagogical metric standardization:** Establish community consensus on intermediate metrics:
    (a) CEFR alignment verification (expert rating or empirical analysis),
    (b) vocabulary appropriateness (automated),

 doi:10.20944/preprints202601.1584.v1

(c) ZPD calibration (cognitive modeling or testing), (d) teacher satisfaction (surveys and usage logs).

(4) **Validation Pillar verification protocols:** Systems incorporating authoritative frameworks should verify representational fidelity through:
   (a) round-trip validation,
   (b) ambiguity documentation,
   (c) independent audit,
   (d) version control tracking schema changes.
   (5) Reproducibility mandates:
   Journals should require:
   (a) model parameters sufficient for recreation,
   (b) training data metadata,
   (c) computational requirements,
   (d) code and KG versions via repositories,
   (e) privacy guarantees.

5.3. Conceptual Implications: Directions for Future Framework Development

These findings suggest the specific value of future frameworks addressing identified gaps. Rather than isolated technologies, an integrated FL-KG-LLM system would ideally embody:

(1) **Pedagogical grounding through constraint-based generation:** LLM outputs should be constrained via KG rules ensuring CEFR alignment, vocabulary appropriateness, and grammatical sequencing. Constraints should guide generation in real-time through guided decoding or Constrained Beam Search.

(2) **Data sovereignty via federated architectures:** KGs and LLMs should be trained and deployed without centralizing learner data. Federated approaches allow institutions to maintain control while collaboratively improving shared models. Privacy-preserving aggregation (differential privacy, secure aggregation) ensures individual learner trajectories remain confidential.

(3) **Representational fidelity through Validation Pillar verification:** Systems should systematically verify that KGs faithfully represent source frameworks before deployment. This requires both automated validation (schema consistency checking) and human expert review (mapping accuracy verification).

(4) **Standardized reporting enabling transparency:** Documentation should follow unified protocols allowing other researchers to understand, critique, and replicate the system. This transparency builds institutional trust and enables cumulative scientific progress.

(5) **Pedagogical validation as design requirement:** Learning effectiveness should be measured throughout development, not as post-hoc evaluation. Iterative design cycles should incorporate pedagogical validation:
   prototype → test with learners → measure outcomes → refine → repeat.

Such integration is not yet demonstrated in published literature. Future research should investigate how FL, KG, and LLM technologies might be combined to address the Integrity Gap. Having mapped the existing landscape and identified systematic gaps, this scoping review provides a foundation for subsequent work developing a conceptual proof-of-concept framework. Framework developers should draw on insights from this scoping review to ensure that novel systems address—rather than perpetuate—reporting gaps, pedagogical disconnection, and Validation Pillar risks. Researchers proposing integrated architectures should engage with the five major gaps identified here, explicitly explaining how their approach addresses each gap and what trade-offs emerge among privacy, pedagogical grounding, and system complexity.

Supplementary Materials: Available online upon publication.

Author Contributions: M.K.: conceptualization, methodology, software, data curation, writing—original draft, visualization. K.K.: supervision, validation, writing—review and editing.

 doi:10.20944/preprints202601.1584.v1

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All data supporting the findings of this study are available on the Open Science Framework (OSF) at https://osf.io/ds74h/ (view-only access: https://osf.io/ds74h/overview?view_only=d8be61e73da8453f9d83c4784d0f860c). This includes: Complete list of 51 included papers with extraction data. Automated extraction codebook and validation protocols. Search strategies across all six databases. Quality assessment rubric and results. Hypothesis testing framework and results. Complete list of excluded studies with exclusion rationales as Supplementary Material S1. Additional materials available upon request from the corresponding author.

Acknowledgments: We acknowledge the use of AI as a tool for assisting authors in the collection and analysis of research papers. Specifically, we acknowledge the use of Qwen 2.5 7B (Alibaba Cloud). Furthermore, we acknowledge the use of other tools/frameworks such as Neo4j, and the support form open-source communities maintaining CEFR-J, WordNet, ConceptNet, and EFLLex.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

Abbreviation	Full Form
iCALL	Intelligent Computer-Assisted Language Learning
LLM	Large Language Model
KG	Knowledge Graph
FL	Federated Learning
CEFR	Common European Framework of Reference for Languages
CEFR-J	CEFR for Japan
EGP	English Grammar Profile
CV	CEFR Companion Volume
ZPD	Zone of Proximal Development
MKO	More Knowledgeable Other
NLP	Natural Language Processing
RAG	Retrieval-Augmented Generation
PEFT	Parameter-Efficient Fine-Tuning
LoRA	Low-Rank Adaptation
FedAvg	Federated Averaging
DP	Differential Privacy
non-IID	Non-Identical and Independent Distribution
HITL	Human-in-the-Loop
NR	Not Reported
RQ	Research Question
PRISMA-ScR	PRISMA Extension for Scoping Reviews
OSF	Open Science Framework
MeSH	Medical Subject Headings
SPIDER	Sample, Phenomenon, Design, Evaluation, Research type

References

1. Bahroun, Z.; Anane, C.; Ahmed, V.; Zacca, A. Transforming education: A comprehensive review of generative artificial intelligence in educational settings through bibliometric and content analysis. Sustainability 2023, 15, 12983. https://doi.org/10.3390/su151712983
2. Kasneci, E.; Sessler, K.; Küchemann, S.; Bannert, M.; Dementieva, D.; Fischer, F.; Gasser, U.; Groh, G.; Günnemann, S.; Hüllermeier, E.; et al. ChatGPT for good? On opportunities and challenges of large

 doi:10.20944/preprints202601.1584.v1

language models for education. Learn. Individ. Differ. 2023, 103, 102274. https://doi.org/10.1016/j.lindif.2023.102274

3. Kung, T.H.; Cheatham, M.; Medenilla, A.; Sillos, C.; De Leon, L.; Elepaño, C.; Madriaga, M.; Aggabao, R.; Diaz-Candido, G.; Maningo, J.; et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS Digit. Health 2023, 2, e0000198. https://doi.org/10.1371/journal.pdig.0000198
4. Council of Europe. Common European Framework of Reference for Languages: Learning, Teaching, Assessment – Companion Volume; Council of Europe Publishing: Strasbourg, France, 2020. Available online: https://www.coe.int/en/web/common-european-framework-reference-languages
5. Vygotsky, L.S. Mind in Society: The Development of Higher Psychological Processes; Harvard University Press: Cambridge, MA, USA, 1978.
6. Bonawitz, K.; Eichner, H.; Grieskamp, W.; Huba, D.; Ingerman, A.; Ivanov, V.; Kiddon, C.; Konečný, J.; Mazzocchi, S.; McMahan, H.B.; et al. Towards federated learning at scale: System design. In Proceedings of the 2nd SysML Conference; Palo Alto, CA, USA, 31 March–2 April 2019; pp. 1-15.
7. Kairouz, P.; McMahan, H.B.; Avent, B.; Bellet, A.; Bennis, M.; Bhagoji, A.N.; Bonawitz, K.; Charles, Z.; Cormode, G.; Cummings, R.; et al. Advances and open problems in federated learning. Found. Trends Mach. Learn. 2021, 14, 1-210. https://doi.org/10.1561/2200000083
8. Bender, E.M.; Gebru, T.; McMillan-Major, A.; Shmitchell, S. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21); Virtual Event, 3-10 March 2021; pp. 610-623. https://doi.org/10.1145/3442188.3445922
9. Nogueira, R.; Jiang, Z.; Pradeep, R.; Lin, J. Document ranking with a pretrained sequence-to-sequence model. In Findings of the Association for Computational Linguistics: EMNLP 2020; Online, November 2020; pp. 708-718. https://doi.org/10.18653/v1/2020.findings-emnlp.63
10. Bosselut, A.; Rashkin, H.; Sap, M.; Malaviya, C.; Celikyilmaz, A.; Choi, Y. COMET: Commonsense transformers for automatic knowledge graph construction. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019); Florence, Italy, 28 July–2 August 2019; pp. 4762-4779. https://doi.org/10.18653/v1/P19-1470
11. Speer, R.; Chin, J.; Havasi, C. ConceptNet 5.5: An open multilingual graph of general knowledge. In Proceedings of the AAAI Conference on Artificial Intelligence 31(1); San Francisco, CA, USA, 4-9 February 2017; pp. 4444–4451. Available online: https://conceptnet.io/
12. Wang, X.; He, X.; Cao, Y.; Liu, M.; Chua, T.-S. KGAT: Knowledge graph attention network for recommendation. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining; Anchorage, AK, USA, 4-8 August 2019; pp. 950-958. https://doi.org/10.1145/3292500.3330989
13. McMahan, H.B.; Moore, E.; Ramage, D.; Hampson, S.; Arcas, B.A.Y. Communication-efficient learning of deep networks from decentralized data. In Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS 2017); Fort Lauderdale, FL, USA, 20-22 April 2017; pp. 1273-1282.
14. Li, T.; Sahu, A.K.; Talwalkar, A.; Smith, V. Federated learning: Challenges, methods, and future directions. IEEE Signal Process. Mag. 2020, 37, 50-60. https://doi.org/10.1109/MSP.2020.2975749
15. Yang, Q.; Liu, Y.; Chen, T.; Tong, Y. Federated machine learning: Concept and applications. ACM Trans. Intell. Syst. Technol. 2019, 10, 1-19. https://doi.org/10.1145/3298981
16. Brown, T.B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language models are few-shot learners. In Advances in Neural Information Processing Systems 33 (NeurIPS 2020); Virtual, December 2020; pp. 1877-1901.
17. Wei, J.; Tay, Y.; Bommasani, R.; Raffel, C.; Zoph, B.; Borgeaud, S.; Yogatama, D.; Bosma, M.; Zhou, D.; Metzler, D.; et al. Emergent abilities of large language models. Trans. Mach. Learn. Res. 2022, arXiv:2206.07682. https://arxiv.org/abs/2206.07682
18. Hogan, A.; Blomqvist, E.; Cochez, M.; d'Amato, C.; de Melo, G.; Gutierrez, C.; Kirrane, S.; Gayo, J.E.L.; Navigli, R.; Neumaier, S.; et al. Knowledge graphs. ACM Comput. Surv. 2021, 54, 1-37. https://doi.org/10.1145/3447772

 doi:10.20944/preprints202601.1584.v1

19. Kautz, H. The third AI summer: AAAI Robert S. Engelmore memorial lecture. AI Mag. 2022, 43, 105-125. https://doi.org/10.1609/aimag.v43i1.19122
20. Meurers, D. Natural language processing and language learning. In The Encyclopedia of Applied Linguistics; Blackwell Publishing: Oxford, UK, 2012. https://onlinelibrary.wiley.com/doi/10.1002/9781405198431.wbeal0858.pub2
21. Tricco, A.C.; Lillie, E.; Zarin, W.; O'Brien, K.K.; Colquhoun, H.; Levac, D.; Moher, D.; Peters, M.D.J.; Horsley, T.; Weeks, L.; et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation. Ann. Intern. Med. 2018, 169, 467-473. https://doi.org/10.7326/M18-0850
22. Arksey, H.; O'Malley, L. Scoping studies: Towards a methodological framework. Int. J. Soc. Res. Methodol. 2005, 8, 19-32. https://doi.org/10.1080/1364557032000119616
23. Levac, D.; Colquhoun, H.; O'Brien, K.K. Scoping studies: Advancing the methodology. Implement. Sci. 2010, 5, 69. https://doi.org/10.1186/1748-5908-5-69
24. Munn, Z.; Peters, M.D.J.; Stern, C.; Tufanaru, C.; McArthur, A.; Aromataris, E. Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. BMC Med. Res. Methodol. 2018, 18, 143. https://doi.org/10.1186/s12874-018-0611-x
25. Peters, M.D.J.; Marnie, C.; Tricco, A.C.; Pollock, D.; Munn, Z.; Alexander, L.; McInerney, P.; Godfrey, C.M.; Khalil, H. Updated methodological guidance for the conduct of scoping reviews. JBI Evid. Synth. 2020, 18, 2119-2126. https://doi.org/10.11124/JBIES-20-00167
26. Page, M.J.; McKenzie, J.E.; Bossuyt, P.M.; Boutron, I.; Hoffmann, T.C.; Mulrow, C.D.; Shamseer, L.; Tetzlaff, J.M.; Akl, E.A.; Brennan, S.E.; et al. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. BMJ 2021, 372, n71. https://doi.org/10.1136/bmj.n71
27. Rethlefsen, M.L.; Kirtley, S.; Waffenschmidt, S.; Ayala, A.P.; Moher, D.; Page, M.J.; Koffel, J.B.; PRISMA-S Group. PRISMA-S: An extension to the PRISMA statement for reporting literature searches in systematic reviews. Syst. Rev. 2021, 10, 39. https://doi.org/10.1186/s13643-020-01542-z
28. Schick, T.; Dwivedi-Yu, J.; Dessì, R.; Raileanu, R.; Lomeli, M.; Zettlemoyer, L.; Cancedda, N.; Scialom, T. Toolformer: Language models can teach themselves to use tools. arXiv 2023, arXiv:2302.04761. https://arxiv.org/abs/2302.04761
29. Tono, Y.; Negishi, M. The CEFR-J: A new platform for constructing a standardized framework for English language education in Japan. In English Profile Studies (Vol. 6); Cambridge University Press: Cambridge, UK, 2012.
30. McCrae, J.P.; Rademaker, A.; Rudnicka, E.; Bond, F. English WordNet 2020: Improving and extending a WordNet for English using an open-source methodology. In Proceedings of the LREC 2020 Workshop on Multiword Expressions; Marseille, France, 11-16 May 2020. Available online: https://github.com/globalwordnet/english-wordnet
31. Dürlich, L.; François, T. EFLLex: A graded lexicon of general English for foreign language learners. In Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018); Miyazaki, Japan, 7-12 May 2018; pp. 3826–3833. Available online: https://cental.uclouvain.be/cefrlex/efllex/
32. Babakniya, S.; Elkordy, A.R.; Ezzeldin, Y.H.; Liu, Q.; Kim, S.; Avestimehr, S.; Dhillon, S. SLORA: Federated parameter efficient fine-tuning of language models. arXiv 2023, arXiv:2308.06522. https://arxiv.org/abs/2308.06522
33. Zhang, J.; Chen, Y.; Cheng, X.; Zhao, S.; Wang, C.; Li, J. FLORA: Low-rank adapters are secretly gradient compressors for federated learning. In Proceedings of the 40th International Conference on Machine Learning (ICML 2023); Honolulu, HI, USA, 23-29 July 2023; pp. 41468-41489.
34. Hu, E.J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W. LoRA: Low-rank adaptation of large language models. In Proceedings of the 10th International Conference on Learning Representations (ICLR 2022); Virtual Event, 25-29 April 2022.
35. Dwork, C.; Roth, A. The algorithmic foundations of differential privacy. Found. Trends Theor. Comput. Sci. 2014, 9, 211-407. https://doi.org/10.1561/0400000042
36. O'Keeffe, A. & Mark, G. (2017). The English Grammar Profile of learner competence: Methodology and key findings. International Journal of Corpus Linguistics 22(4), 457-489.

 doi:10.20944/preprints202601.1584.v1

37. Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-T.; Rocktäschel, T.; et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In Advances in Neural Information Processing Systems 33 (NeurIPS 2020); Virtual, December 2020; pp. 9459-9474.
38. Kenton, Z.; Everitt, T.; Weidinger, L.; Gabriel, I.; Mikulik, V.; Irving, G. Alignment of language agents. arXiv 2021, arXiv:2103.14659. https://arxiv.org/abs/2103.14659
39. Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.L.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. Training language models to follow instructions with human feedback. In Advances in Neural Information Processing Systems 35 (NeurIPS 2022); New Orleans, LA, USA, 28 November–9 December 2022; pp. 27730-27744.
40. Burrows, S.; Potthast, M.; Stein, B. Paraphrase acquisition via crowdsourcing and machine learning. ACM Trans. Intell. Syst. Technol. 2013, 4, 1-21. https://doi.org/10.1145/2483669.2483676
41. Tayyeh, M.; AL-Jumaili, H.K. Balancing Privacy and Performance: A Differential Privacy Approach in Federated Learning. Computers 2024, 13, 277. https://doi.org/10.3390/computers13110277
42. Bonawitz, K.; Ivanov, V.; Kreuter, B.; Marcedone, A.; McMahan, H.B.; Patel, S.; Ramage, D.; Sadin, A.; Smith, N. Practical Secure Aggregation for Privacy-Preserving Machine Learning. In Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security (CCS); Dallas, TX, USA, 30 October–3 November 2017; pp. 1175-1191.
43. Large-scale Evaluation of Reporting Quality in 21,041 Randomized Trials (1966-2024). medRxiv preprint, 2025. https://www.medrxiv.org/content/10.1101/2025.03.06.25323528v1.full.pdf
44. Assessment of Adherence to PRISMA Guidelines in Stroke Research. Stroke Journal, 2020. https://www.ahajournals.org/doi/10.1161/STROKEAHA.120.033288