IDEAS

Journal on Language Teaching and Learning, Linguistics and Literature

ISSN 2338-4778 (Print)
ISSN 2548-4192 (Online)

Volume 14, Number 2, December 2026
pp. 6624 - 6651

Issued by English study program of UIN Palopo

Google NotebookLM in EFL Instruction: A Narrative Review of Multimodal Affordances for AI-Integrated Language Learning

Jusak Patty
Pendidikan Bahasa Inggris, Universitas Pattimura, Ambon

Article Info

Received: 2026-03-29
Revised: 2026-06-18
Accepted: 2026-08-19

Keywords:
EFL instruction,
generative AI,
Google NotebookLM,
multimodal affordances

DOI:
10.24256/ideas. v14i1. 10014

Corresponding Author:
Jusak Patty
jusak.patty@gmail.com
Pendidikan Bahasa Inggris, Universitas Pattimura, Ambon

Abstract

Google NotebookLM has attracted growing interest as a multimodal generative AI tool with potential for English as a Foreign Language instruction, yet its features have not been evaluated against established EFL pedagogical principles. This article presents a narrative literature review (NLR) of 50 articles to assess the extent to which NotebookLM's multimodal affordances correspond to those principles, and to identify the conditions under which its affordances weaken or fail. Five evaluative criteria were derived from three theoretical traditions: social semiotics and multiliteracies theory, the cognitive theory of multimedia learning, and CALL and task-based language teaching research. The analysis finds that the tool's source-grounded architecture meaningfully supports four affordances: audio transduction, cognitive load management under default conditions, pacing alignment, and authentic task affordance. Three further affordances remain unmet under the tool's current design: proficiency differentiation, full modal diversity, and contextual accessibility under constrained conditions. The review identifies three critical boundaries. Source-grounded generation carries architectural constraints and accuracy risk; frictionless compression suppresses Critical Framing; and native-speakerist audio defaults risk reinforcing hegemonic language norms. Across all findings, teacher mediation emerges as the constitutive condition for the actualisation of affordances. The findings suggest that NotebookLM functions most effectively as a teacher-dependent instrument for multimodal content engagement, and that its responsible deployment in EFL contexts requires deliberate source curation, scaffolded task design, and explicit critical AI literacy instruction.

1. Introduction

Language learning has never been a purely linguistic process. Meaning-making draws simultaneously on written, auditory, visual, gestural, and spatial resources (Cope & Kalantzis, 2000, 2009; The New London Group, 1996) and this multimodal reality carries direct implications for how EFL instruction must be designed (Hafner et al., 2015; Rahmanu & Molnár, 2024). The theoretical case for this repositioning is well established: contemporary communicative diversity demands instruction that moves beyond single, print-based literacy and that positions learners as active designers of meaning across the full range of semiotic modes available to them (Kress, 2009; The New London Group, 1996). This shift has gradually reshaped how EFL researchers and educators conceptualise language engagement, moving beyond linguistic acquisition toward a richer, mode-diverse communicative practice.

A substantial empirical base supports the theoretical turn toward multimodality. Multimodal digital environments produce measurable gains in vocabulary development, comprehension, and communicative competence in EFL contexts, and the integration of multiple modes of input is consistently associated with meaningful learning outcomes (Hafner et al., 2015; Lai & Li, 2011; W. Li et al., 2022; Mihaylova et al., 2022; Rahmanu & Molnár, 2024). The emergence of generative artificial intelligence since 2022 has extended this landscape considerably: tools built on large language models have been adopted rapidly across EFL writing (Lo et al., 2024; Rashid et al., 2026; Zare et al., 2025), speaking (Guan et al., 2025; Huang et al., 2024), reading (Fan, 2025; Labibah et al., 2025), and vocabulary domains (Moulieswaran, 2025).

Most of this research, however, has concentrated on text-based AI interactions, examining how learners read, write to, or receive textual feedback from generative AI systems. The multimodal dimension of these tools, specifically their capacity to generate and transform content across text, audio, and visual modes within a single environment, remains comparatively underexplored in the language education literature. This gap has grown more consequential as tool capabilities have expanded.

Google NotebookLM represents a particularly compelling case within this emerging landscape. Launched in mid-2023 and built on Google's Gemini foundation model, it departs from dominant large language model architecture by operating on a principle of source grounding: rather than generating outputs from an open-domain training corpus, the tool generates all responses exclusively from the sources the user uploads (Martin & Johnson, 2023). This architectural distinction has direct pedagogical implications for EFL contexts, where learner control over the knowledge base, source accountability, and multimodal engagement with the same content are all theoretically consequential (Alisoy, 2025; Reyna, 2025). The tool accepts a range of source types on the input side.

It generates written summaries, study guides, interactive question-and-answer exchanges, visual mind maps, and AI-generated audio discussions in a conversational podcast format on the output side, with a 2025 feature expansion further broadening the range of modalities available within a single platform (Martin & Johnson, 2023; Paul & Dhami, 2025). Despite this architecturally distinctive and pedagogically promising profile, at the time of writing, no peer-reviewed study retrieved within the temporal scope of this review has examined NotebookLM within the theoretical framework of multimodal EFL pedagogy, and no review has assessed whether its features correspond to, or diverge from, what the accumulated literature identifies as effective for multimodal language learning.

This article addresses that gap through a narrative literature review guided by two research questions. The first asks: to what extent do NotebookLM's multimodal features correspond to the principles of multimodal EFL pedagogy established in the existing literature? The second asks: where do the boundaries of that correspondence lie, and under

ISSN 2338-4778 (Print)
ISSN 2548-4192 (Online)

what conditions do its affordances weaken or fail? To answer these questions, the review derives five evaluative criteria from three interconnected theoretical traditions: social semiotics and multiliteracies theory, the cognitive theory of multimedia learning, and CALL and task-based language teaching research. These criteria organise the entire analysis across a corpus of 50 articles. The five-criterion evaluative framework is the primary conceptual contribution of this review: it provides a replicable, theoretically grounded instrument for assessing AI multimodal tools in EFL pedagogy that extends beyond NotebookLM to any tool whose affordances need to be mapped against established pedagogical principles.

The article proceeds as follows. The remainder of this Introduction synthesises three theoretical traditions and derives the five evaluative criteria that anchor the analysis. The Method section describes the NLR design, search strategy, and inclusion criteria. The Results section maps NotebookLM's affordances against the five criteria. The Discussion addresses the three critical boundaries and broader implications. The article concludes with a synthesis of contributions and directions for future research.

Theoretical Framework

Three interconnected theoretical traditions collectively define what any pedagogically serious multimodal EFL tool must accomplish. Their synthesis, rather than any single tradition read in isolation, produces the five evaluative criteria that organise the analysis that follows. The tensions between these traditions are as analytically important as their shared premises, and they are explicitly addressed at the close of this section.

Social Semiotics and Multiliteracies Theory

All communication draws on multiple semiotic modes simultaneously, and the selection of one mode over another is itself a meaningful act rather than a neutral technical decision (Kress & Van Leeuwen, 2001, 2020). If linguistic, visual, auditory, gestural, and spatial modes each carry meaning that no other mode can fully replicate, then instruction confined to a single mode is structurally impoverished regardless of its quality within that mode (Cope & Kalantzis, 2009; Kress & Van Leeuwen, 2001). The programmatic response to this insight is the pedagogy of multiliteracies, which repositions learners as active designers who select, combine, and transform semiotic resources across the full range of modes available in contemporary communication (Cope & Kalantzis, 2000, 2009; The New London Group, 1996).

In digital environments, this process has been theorised with increasing precision through transpositional grammar and the concept of transduction, which map how meaning is inevitably reshaped rather than merely represented as it crosses from one mode to another, a vocabulary directly relevant to evaluating AI tools capable of converting the same content between textual, auditory, and visual formats (Cope & Kalantzis, 2020; Jewitt et al., 2016).

The criterion that follows from this tradition is not simply that a tool must offer multiple modes, but that the transitions between them must do genuine meaning-making work. **Criterion 1: Modal Diversity and Transduction Capacity** therefore has two operationally distinct components: the tool must generate output in more than one semiotic mode from a single source, and those modal transitions must create conditions for genuine transduction in which meaning is reshaped across modal boundaries rather than merely restated in a different surface format. This distinction between transduction and repetition is the primary standard against which NotebookLM's cross-modal output is assessed in the Results.

Cognitive Theory of Multimedia Learning and SLA Applications

What the social semiotic tradition does not address is how the learner's cognitive architecture mediates meaning-making. Mayer's Cognitive Theory of Multimedia Learning (CTML) establishes that verbal and visual information are processed through separate channels of limited capacity, and that meaningful learning requires active integration across both (R. E. Mayer & Fiorella, 2021). The practical implications for instructional design are well documented: audio narration generally outperforms on-screen text when combined with visual content; eliminating extraneous material improves learning; adding printed text to spoken narration can impair rather than assist processing; and allowing learners to control pacing improves outcomes.

These principles carry additional weight in L2 contexts because processing content in a foreign language consumes working memory capacity that native-speaker models treat as available for multimodal integration (Plass & Jones, 2005). The empirical consequences are specific enough to be pedagogically actionable: slowly paced narration supports comprehension gains in non-native speakers, while rapid narration with simultaneous captions produces overload with no measurable benefit (R. E. Mayer et al., 2014). Across vocabulary research, dual-mode input consistently outperforms single-mode input, but three or more simultaneous modes generally undermine rather than enhance long-term retention, with the threshold varying by proficiency level (W. Li et al., 2022; Ramezanali et al., 2021). The compounded cognitive burden of L2 and multimedia processing must therefore be managed through architectural design that keeps working memory demands within productive range: **Criterion 2: Cognitive Load Management**, operationalised here by examining whether NotebookLM's architecture separates rather than simultaneously imposes its output types.

Effective multimodal L2 design must also reflect what the accumulated empirical literature identifies as consequential for learning outcomes across learner populations. The operational basis for input calibration rests more securely on evidence from proficiency-stratified multimedia studies than on any single theoretical model (W. Li et al., 2022; Ramezanali et al., 2021): proficiency level strictly moderates the effectiveness of multimodal input, with lower-proficiency learners requiring substantially more scaffolding and modal separation than advanced learners, a pattern attributable to the greater working memory demands that L2 processing imposes at lower proficiency levels (W. Li et al., 2022). Learner control over pacing consistently produces positive outcomes, while uniform, unmodifiable presentation conditions produce gains only for learners whose proficiency is sufficient to benefit from them (R. E. Mayer et al., 2014; R. E. Mayer & Fiorella, 2021; Ramezanali et al., 2021).

The practical demand this places on instructional design is that the evidence base must be actively operationalised rather than incidentally satisfied: **Criterion 3: Evidence Alignment**, operationalised here by examining whether NotebookLM's design decisions correspond to what the empirical literature identifies as consequential, including proficiency-differentiated modal sequencing and learner control over pacing.

CALL, Task-Based Language Teaching, and Digital Literacy Research

The cognitive and semiotic demands established by the preceding two traditions are necessary but not sufficient conditions for pedagogically purposeful technology use. Technology-mediated language use becomes pedagogically purposeful only when learners use language for authentic communicative purposes rather than decontextualised form practice (González-Lloret & Ortega, 2014), and digital tools do not merely deliver this kind of learning but actively reshape what counts as meaningful language use and what competences learners need to develop (Hafner et al., 2015). Underlying both arguments is

ISSN 2338-4778 (Print)
ISSN 2548-4192 (Online)

an ecological understanding of affordance in which affordances are not properties of objects or agents in isolation but emerge from the dynamic relationship between a tool, a user, and the institutional environment in which both are situated (Van Lier, 2004). This relational account has three analytically distinct layers: affordance as possibility refers to what a tool's structural design makes available independently of any particular user; affordance as actuality refers to what becomes available to a specific user shaped by their competences and institutional position; and affordance as uptake refers to the moment of purposeful engagement in which possibility becomes learning action (Blin, 2016; Van Lier, 2004). The distinction between possibility and uptake separates Criterion 4 from Criterion 5 in this framework: the former evaluates structural conditions built into the tool's design, while the latter assesses whether the ecological conditions for uptake are realisable in institutional practice.

A tool's structural features do not guarantee pedagogical value; uptake requires conditions, including teacher readiness, task design, and institutional ecology, that the tool itself does not provide (Blin, 2016). Multimodal approaches in digital EFL contexts produce consistent and measurable gains across vocabulary, comprehension, and integrated language skills when such ecological conditions are met, but the same tools produce no comparable benefit when tasks are decontextualised, scaffolding is absent, or institutional barriers including variable internet access, limited devices, and uneven teacher digital literacy prevent sustained engagement (Hafner et al., 2015; Mali & Salsbury, 2021; Mihaylova et al., 2022; Rahmanu & Molnár, 2024). These conditions characterise EFL settings across diverse geographic and institutional contexts and must be part of any serious evaluation of a tool's practical pedagogical potential.

Two further criteria follow. Technological sophistication and cognitive efficiency are pedagogically inert unless the tool supports goal-oriented, meaning-focused language use: **Criterion 4: Authentic Task Affordance**, operationalised by examining whether NotebookLM's architecture creates structural conditions (affordance as possibility, Van Lier, 2004) for genuine communicative or comprehension goals rather than decontextualised practice. Equally, a tool whose deployment depends on infrastructure, cost structures, or teacher expertise that most EFL settings cannot sustain cannot be assessed as pedagogically viable across the range of contexts it purports to serve: **Criterion 5: Contextual Accessibility**, operationalised by examining whether the ecological conditions necessary for affordance uptake (Blin, 2016; Van Lier, 2004) are realisable under the variable institutional circumstances that characterise EFL practice globally.

Productive Tensions among the Three Traditions

These three traditions do not converge entirely, and the tensions between them define the evaluative space within which any multimodal EFL tool must be situated. The most consequential tension runs between the social semiotic endorsement of modal multiplicity, grounded in the argument that each mode carries meaning no other mode can express, and the cognitive insistence that adding modes can impair rather than support learning when working memory is exceeded.

A related tension exists between the task-based emphasis on authentic communicative complexity and the CTML requirement for controlled, sequenced presentation conditions, a conflict that is especially acute where learners simultaneously manage the cognitive demands of a foreign language and those of multimodal processing. A third tension runs between Van Lier's (2004) ecological principle that affordances emerge from the relationship between tool, user, and context, and the social semiotics principle that modal richness is a structural property of the design itself. These two positions do not necessarily conflict, but they foreground different loci of pedagogical responsibility.

These tensions are not resolved in advance but are examined in the context of NotebookLM's specific architecture in the Results and Discussion sections that follow. Readers are directed to attend particularly to the intersection of Criterion 1 and Criterion 2, where the modal richness endorsed by social semiotics meets the cognitive load constraints established by CTML, and to the intersection of Criterion 4 and Criterion 5, where the structural conditions for authentic task affordance are evaluated alongside the institutional conditions required for that affordance to be taken up.

What the tensions establish at this stage is that the five criteria must be applied in productive tension with one another rather than independently: a tool that satisfies Criterion 1 while failing Criterion 2 does not meet the evaluative framework, nor does one that satisfies Criterion 4 while remaining inaccessible under the institutional and ecological conditions that characterise many EFL settings.

Table 1. Five Evaluative Criteria for Multimodal EFL Tools Derived

Criterion	Theoretical Basis	Core Demand for EFL Tools
C1: Modal Diversity and Transduction Capacity	(Cope & Kalantzis, 2000, 2020; Jewitt et al., 2016; Kress & Van Leeuwen, 2001, 2020; The New London Group, 1996)	The tool must activate two or more semiotic modes and enable transitions that reshape meaning across modal boundaries rather than merely restate it in a different surface format.
C2: Cognitive Load Management	(R. E. Mayer et al., 2014; R. E. Mayer & Fiorella, 2021; Plass & Jones, 2005)	The tool must manage the dual cognitive burdens of foreign-language and multimedia processing within the constraints of working memory.
C3: Evidence Alignment	(W. Li et al., 2022; R. E. Mayer et al., 2014; Ramezanali et al., 2021)	Tool design must reflect empirical findings on effective multimodal L2 learning, including proficiency-differentiated modal sequencing and learner control over pacing
C4: Authentic Task Affordance	(Blin, 2016; González-Lloret & Ortega, 2014; Hafner et al., 2015; Van Lier, 2004)	The tool must create structural conditions (affordance as possibility) for goal-oriented, meaning-focused language use rather than 6629ealizable6629lized form practice.
C5: Contextual Accessibility	(Blin, 2016; Mihaylova et al., 2022; Rahmanu & Molnár, 2024; Van Lier, 2004)	The tool must be operable under variable institutional conditions, with the ecological conditions for affordance uptake 6629ealizable across settings with limited infrastructure and uneven teacher digital readiness.

The tensions among the five criteria are managed through consistent application across the same corpus with the same evaluative questions, so that trade-offs and contradictions surface from the evidence rather than being predetermined by theoretical preference. Table 1 summarises the five criteria, their theoretical basis, and the core demand each imposes on multimodal EFL tools. These five criteria constitute a replicable, criterion-driven evaluative instrument for assessing multimodal AI tools against established EFL pedagogical principles. Their primary contribution is methodological: they provide a theoretically grounded framework that extends beyond NotebookLM to any AI tool whose multimodal features need to be evaluated against what the accumulated literature identifies as effective for language learning. The Method section that follows describes the methodological procedures used to operationalise this framework.

2. Method

This study adopts a narrative literature review (NLR) design, which is well-suited for synthesising heterogeneous literature around an emerging object of inquiry where empirical evidence is nascent, and the primary analytical task requires interpretive integration rather than statistical aggregation (Gregory & Denniss, 2018; Snyder, 2019;

ISSN 2338-4778 (Print)
ISSN 2548-4192 (Online)

Sukhera, 2022). NotebookLM's presence in the pedagogical literature combines empirical studies, practitioner reports, theoretical analyses, and critical assessments that do not yet share sufficient methodological homogeneity to support systematic aggregation, and an NLR accommodates this breadth without sacrificing analytical rigour (Ferrari, 2015; Pautasso, 2019). The five evaluative criteria established in the theoretical framework introduced above serve as the organising logic for that synthesis, adding a layer of transparency and replicability to the interpretive process through criteria that are theoretically derived and operationally defined.

The literature search was conducted using Publish or Perish (PoP) to query three databases: Google Scholar (n = 1,711), Scopus (n = 106), and PubMed (n = 43). Scopus was included for its interdisciplinary coverage of peer-reviewed education and technology research; PubMed was included because a substantial portion of early NotebookLM literature originated in clinical and health professions education contexts before being adopted by broader pedagogical communities. An additional targeted search was conducted independently in ERIC (n = 2) to supplement coverage of language education research not fully indexed in the preceding databases.

Two core search strings were applied, with Boolean variations explored iteratively across databases to maximise recall: (1) "NotebookLM" AND ("language learning" OR "EFL" OR "English language teaching"), and (2) ("generative AI" OR "AI tools") AND "multimodal" AND ("EFL" OR "language learning"). The first string targeted studies directly examining NotebookLM in educational contexts, while the second cast a broader net across the multimodal generative AI literature in language learning, providing the comparative frame for criterion evaluation. All searches were conducted within a temporal scope of January 2023 to March 2026, covering the period from NotebookLM's initial release through the completion of the search. The combined search across all four sources yielded 1,862 records.

Following the retrieval of records, six paired inclusion and exclusion criteria were established prior to screening. These criteria reflect the dual scope of the review: studies directly examining NotebookLM in EFL contexts, and studies examining multimodal generative AI affordances in language learning that provide the theoretical and empirical context for evaluating the five criteria. Table 2 presents the full set of criteria.

Table 2. Inclusion and Exclusion Criteria for Study Selection

Dimension	Inclusion Criteria	Exclusion Criteria
Topic and domain	Studies must address Google NotebookLM, multimodal generative AI tools, or multimodal affordances in EFL or language-learning contexts.	Articles from unrelated academic disciplines (e.g., health sciences, engineering, and general education without a language-learning focus) were excluded.
Study type	Included studies may report empirical research, theoretical or conceptual analyses, or critical pedagogical assessments, provided they offer a demonstrable analytical contribution to RQ1 or RQ2, understood as identifying features of the tool, measuring effects on language variables, or building an argument about the pedagogical conditions of tool use.	Articles with no analytical substance, such as blog posts, editorials without argument or data, and news items, were excluded.
Source type	Only peer-reviewed journal articles and conference proceedings were included.	Theses, preprints, presentation slides, videos, newsletters, and other non-peer-reviewed formats were excluded.
Publication	Only articles published between January	Articles published before January 2023

Dimension	Inclusion Criteria	Exclusion Criteria
period	2023 and March 2026 were considered.	or after March 2026 were excluded.
Language	Only English-language articles were included.	Articles written in languages other than English were excluded.
Accessibility	Articles must have been available for full-text review. Retrieval was attempted via institutional access, author copies via ResearchGate, and open-access repositories before an article was deemed inaccessible.	Articles behind paywalls with no accessible copy were excluded at Stage 4.

Screening proceeded across three stages with all decisions documented following the PRISMA flow diagram (Moher et al., 2009), adopted here as a reporting tool for transparency rather than a methodological claim of systematic review status, consistent with the principle that PRISMA's selection tracking concepts apply to any review incorporating structured screening procedures regardless of analytical design (Ferrari, 2015; Moher et al., 2009). Prior to Stage 1, automated deduplication removed 831 records, leaving 1,031 unique records for screening. Stage 1 screened all records by source identifiability, title relevance to EFL, language learning, AI tools, or multimodal pedagogy, and English-language publication; 705 records were excluded at this stage (no identifiable source: 12; title irrelevance: 656; non-English: 37), leaving 326 for abstract review. Stage 2 assessed all 326 abstracts for relevance to RQ1 or RQ2; 194 articles advanced to full-text review, and 132 were excluded.

Stage 3 applied full-text screening against the six criteria in Table 2, with all inclusion decisions made by the reviewer through direct reading and Google NotebookLM used solely for pre-reading content orientation under documented human oversight rather than as an analytical decision-maker (Adel & Alani, 2025; Bijker et al., 2024).

Documented oversight entailed generating a NotebookLM summary prior to reading, then making all inclusion and exclusion decisions exclusively through direct engagement with the full text, with the summary serving only as an orienting reference. A total of 144 articles were excluded at Stage 3, of which 32 were inaccessible due to paywall restrictions, introducing a potential open-access bias acknowledged as a limitation of the corpus. All screening was conducted by a single reviewer, which is a recognised practice in NLR designs where the individual reviewer's theoretical perspective shapes the interpretive process (Pautasso, 2019; Sukhera, 2022), mitigated here through pre-specified criteria and transparent documentation of all screening decisions (Ferrari, 2015; Green et al., 2006). The final corpus comprises 50 articles. Figure 1 presents the PRISMA flow diagram for the full selection process.

ISSN 2338-4778 (Print)
ISSN 2548-4192 (Online)

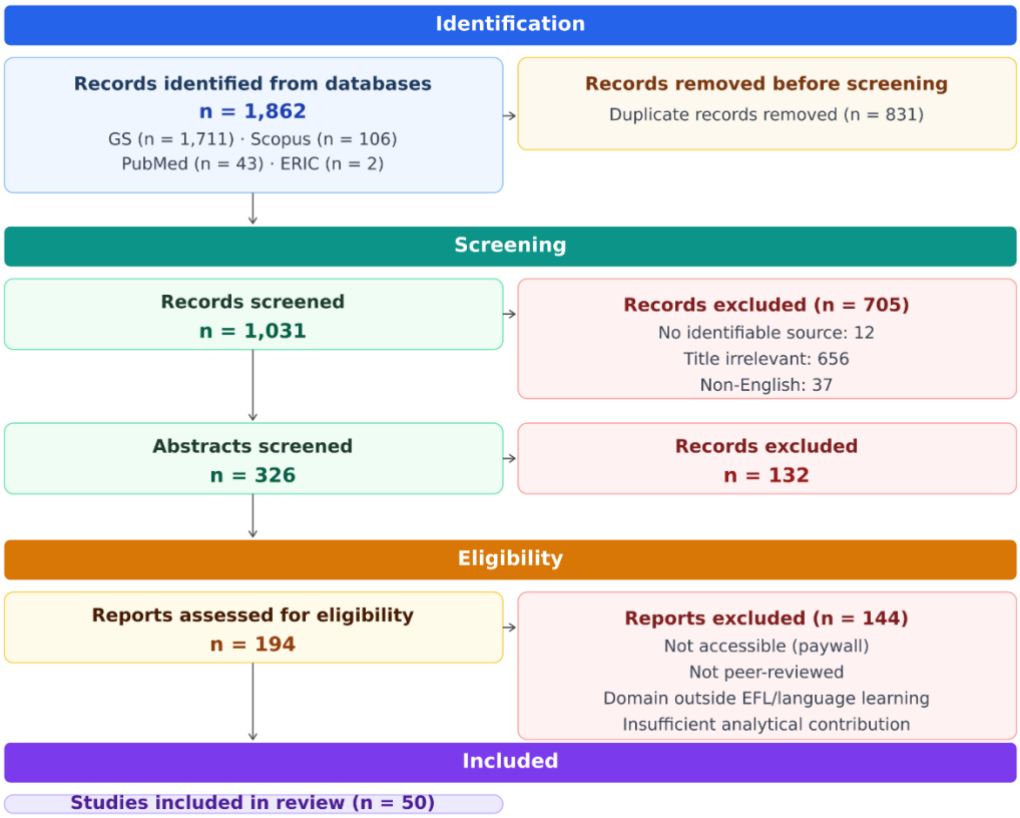

Figure 1. PRISMA 2020 flow diagram for article selection.

Data extraction was guided by the five evaluative criteria established in the theoretical framework introduced above. For each article, the following information was recorded: evidence type (empirical, theoretical-conceptual, critical-analytical, or systematic review), relevance to RQ1 or RQ2, the primary evaluative criterion addressed, a secondary criterion where applicable, and the article's evidential status with respect to Google NotebookLM. Articles were classified across three tiers: direct evidence if they examined, deployed, or evaluated NotebookLM explicitly; partial evidence if they referenced the tool in a limited capacity while primarily examining a related issue; or contextual evidence if their findings were relevant to one or more criteria without addressing NotebookLM directly.

For classification purposes, direct evidence required substantive engagement with NotebookLM as the primary object of analysis, whereas partial evidence was assigned when the tool was merely named and discussed, playing a secondary role within a broader argument. Direct and partial evidence constitute the primary basis for criterion verdicts, while contextual evidence establishes the comparative frame against which NotebookLM-specific findings are assessed.

The analysis was criterion-driven for RQ1, with each criterion operationalised as a consistent analytical question applied across the corpus. Each criterion received one of three verdicts according to a fixed decision rule. A criterion was judged met when direct and partial evidence confirmed that NotebookLM's architecture satisfied the criterion's core demand without a structural gap, with any remaining requirement attributable to ordinary pedagogical practice rather than to tool deficiency. A criterion was judged partially met when direct evidence confirmed the criterion for one architectural feature or dimension while leaving at least one other dimension structurally unaddressed, such that the gap originated in the tool's design rather than in its deployment.

A criterion was judged conditionally met when the tool's default architecture satisfied the criterion but a specific deployment decision, available within the same tool, could reverse that alignment, making the verdict dependent on instructional choices rather than on architecture alone. The boundary between partially met and conditionally met therefore rests on the locus of the limiting factor: a structural absence in the tool's design yields a partial verdict, whereas a deployment-dependent risk within an otherwise sound architecture yields a conditional one. For RQ2, the same corpus was examined inductively for recurring patterns of constraint and pedagogical risk, which were consolidated into three critical boundaries. Table 3 presents the mapping of all 50 corpus articles to their primary analytical category and evidence tier. Note that several articles contributed to more than one analytical category, as indicated by the asterisk; the total of 50 refers to unique articles, not to the total of category entries.

Table 3. Mapping of Corpus Articles to Analytical Categories and Evidence Tiers

Analytical Category	Evidence Tier	Articles
C1: Modal Diversity and Transduction Capacity	Direct (5)	(Brown & Saurez, 2025; Q. Liu & Zhao, 2025; Maranzana, 2025; Yeo et al., 2025; Zhao et al., 2026)
	Contextual (9)	(Christoforou, 2025; Ci & Jiang, 2025; Cope et al., 2025; Djekidel & Rogti, 2025; Fang et al., 2025; Huang et al., 2024; Labibah et al., 2025; D. Li et al., 2025; M. Liu et al., 2024; Nguyen & Nguyen, 2025)
C2: Cognitive Load Management	Direct (3)	(Alisoy, 2025; Q. Liu & Zhao, 2025*; Sturgeon Delia, 2025)
	Partial (2)	(Lee et al., 2025; Menon et al*., 2025)
	Contextual (2)	(Fan, 2025*; Ren et al., 2026)
C3: Evidence Alignment	Contextual (7)	(Erdiani et al., 2024; Fan, 2025; Q. Liu & Zhao, 2025; Menon et al., 2025; Moulieswaran, 2025; Wu & Li, 2025; Zhuang et al., 2025)
C4: Authentic Task Affordance	Direct (4)	(Jenkins & Hussain, 2026; Lindade, 2025; Maranzana, 2025; Panday-Shukla, 2024)
	Contextual (8)	(Compagnoni, 2025; Derakhshan & Park, 2026; Emerson, 2024; Guan et al., 2025*; Nurjamin et al., 2026; Sadykova & Kayumova, 2025; Utami & Rohadi, 2026; Zare et al., 2025)
C5: Contextual Accessibility	Direct (3)	(Aremu et al., 2025; Brown & Saurez, 2025; Sturgeon Delia, 2025)
	Contextual (5)	(Guan et al., 2025; Moulieswaran, 2025; Patty, 2025; Rashid et al., 2026; Stewart & Zheng, 2024*)
B1: Structural Constraints and Accuracy Risk	Direct (8)	(Albrecht-Crane, 2025; Jenkins & Hussain, 2026; Lindade, 2025; Panday-Shukla, 2024; Reyna, 2025; Venhrina, 2025; Willner, 2025; Yeo et al., 2025*)
	Partial (1)	(Soge & Mbato, 2025*)
	Contextual (1)	(Dewi, 2025)
B2: Cognitive Offloading and Suppression of Critical Framing	Direct (2)	(Albrecht-Crane, 2025; Alisoy, 2025)
	Partial (3)	(Lee et al., 2025; Rebolledo et al., 2025; Soge & Mbato, 2025)
	Contextual (6)	(Cope et al., 2025; Djekidel & Rogti, 2025; Fang et al., 2025; Jiang, 2024; Kang et al., 2026; M. Liu et al., 2024)
B3: Native-Speakerism and Linguistic Equity	Direct (3)	(Lindade, 2025; Maranzana, 2025; Yeo et al., 2025*)
	Contextual (4)	(Cope et al., 2025; Godwin-Jones, 2024; Nguyen & Nguyen, 2025; Stewart & Zheng, 2024*)

* Articles contributing to more than one analytical category

ISSN 2338-4778 (Print)
ISSN 2548-4192 (Online)

3. Result

The five evaluative criteria derived in the theoretical framework introduced above organise the analysis in this section. For each criterion, direct NotebookLM evidence carries the primary evidential weight, contextual evidence from the broader GenAI and multimodal EFL literature establishes the comparative frame, and the verdict reflects the totality of evidence rather than any single study. Because several articles contribute to more than one criterion, as indicated in Table 3, the evidential connections are noted where they arise. RQ1 is addressed through the criterion-by-criterion analysis below, with Table 4 summarising the verdicts and their key qualifiers at the close of this section.

NotebookLM's Multimodal Features and EFL Pedagogical Principles

Each of the five evaluative criteria derived in the theoretical framework above is applied here as a consistent analytical question across the corpus, with direct NotebookLM evidence carrying the primary evidential weight, partial evidence contributing supporting detail, and contextual evidence from the broader GenAI and multimodal EFL literature establishing the comparative frame against which NotebookLM-specific findings are assessed. The criterion verdicts that emerge from this analysis are not binary determinations but qualified assessments that reflect the totality of available evidence; where a criterion is partially or conditionally met, the qualifier identifies the specific architectural or pedagogical condition responsible. These five verdicts address RQ1 and provide the evidential basis from which the critical boundaries examined in the Discussion are derived.

Criterion 1: Modal Diversity and Transduction Capacity

Audio transduction is the clearest and most evidently robust affordance NotebookLM offers against this criterion. When source text is converted into conversational audio through the Audio Overview feature, the result is a reorganisation of content into a different register with different interactional cues rather than a simple reading aloud of the original (Q. Liu & Zhao, 2025; Maranzana, 2025; Yeo et al., 2025). This aligns with the conceptual conditions for transduction articulated by Kress & Van Leeuwen (2001): the register shift from academic text to conversational audio foregrounds different aspects of the content and produces meaning that the source text cannot produce on its own. Brown & Saurez (2025) confirmed this through a Universal Design for Learning (UDL) analysis, showing that learners can direct questions to the AI hosts in ways the written source does not permit; Yeo et al. (2025) documented how the audio discussion reorganises content into a talk-show format that extracts, embellishes, and extends rather than merely summarises; and Zhao et al. (2026) recorded how AI-driven audio from academic sources supports international students' listening and speaking preparation.

The broader GenAI literature corroborates these findings: transduction has been identified as a key scaffolding mechanism in L2 digital multimodal composing (Ci & Jiang, 2025), and active semiotic decision-making across modes is documented as a genuine learner behaviour when AI tools require it (Fang et al., 2025; D. Li et al., 2025; M. Liu et al., 2024). Cope et al. (2025) reinforce this point, arguing that AI-enabled modal transposition has become a fundamental process in contemporary language learning.

Where NotebookLM falls short is in the visual domain. Systematic review evidence confirms that AI-generated animated visuals improve EFL learning outcomes (Djekidel & Rogti, 2025), and empirical work documents improved reading comprehension through AI visual support in Indonesian EFL contexts (Labibah et al., 2025). Neither the attitude study by Huang et al. (2024) nor the systematic review by Nguyen & Nguyen (2025) examined NotebookLM directly, but together they establish that the visual mode carries substantial

evidential weight in EFL multimodal learning, and the tool's mind map feature does not address that weight. A mind map reorganises propositional content spatially but does not reshape meaning: the conceptual content is fully recoverable from the source text without transformation, which is reorganisation rather than transduction as specified by Cope & Kalantzis (2020).

The absence extends further: gestural and spatial modes are entirely outside the tool's generative scope, and the kind of active modal orchestration across speech, movement, space, and text that Christoforou (2025) documented among ESP learners using AI tools for video composition has no equivalent pathway within NotebookLM's architecture. This criterion is therefore partially satisfied: audio transduction is real and confirmed by direct evidence, while the remaining modes are either structurally limited to reorganisation or absent entirely. The audio dimension of this verdict also carries a qualification that is examined more fully in the third critical boundary: the accent defaults embedded in NotebookLM's audio output constrain the range of linguistic resources the tool makes available as legitimate meaning-making designs.

Criterion 2: Cognitive Load Management

The tension identified in the theoretical framework above between the social semiotic endorsement of modal multiplicity and the cognitive constraint that adding modes can impair rather than support learning surfaces directly here: the question is not whether NotebookLM offers multiple modes, but whether its architecture manages the cognitive cost of switching between them. NotebookLM's sequential architecture is its principal asset against this criterion. Learners engage with audio, text, or interactive Q&A in separate episodes rather than simultaneously, which distributes the dual burden of L2 and multimedia processing across time rather than compounding it (R. E. Mayer et al., 2014; R. E. Mayer & Fiorella, 2021).

Alisoy (2025) applied Cognitive Load Theory specifically to NotebookLM, arguing that its structured summaries and key-point lists externalise the organisational burden that would otherwise compete with comprehension for working memory capacity, mapping precisely onto the segmenting principle that R. E. Mayer & Fiorella (2021) identify as a key determinant of multimedia learning outcomes. Empirical weight was added by Q. Liu & Zhao (2025), whose quasi-experimental study found that converting a traditional news report into a naturally paced podcast format reduced anxiety and improved listening comprehension, complementing R. E. Mayer et al.'s (2014) finding that slowly paced narration supports comprehension gains in non-native speaker populations. Sturgeon Delia (2025) contributed practitioner-level confirmation, reporting reduced cognitive burden when students processed lecture content via AI-generated audio compared with written text alone.

Contextual evidence reinforces this pattern: Fan's (2025) empirical work on working memory in GenAI vocabulary processing established the direct relevance of dual coding to L2 contexts, Ren et al. (2026) documented the interaction between multimodal feedback design and cognitive load in AI-mediated writing, and Lee et al. (2025), who referenced NotebookLM explicitly, confirmed that cognitive load is an active pedagogical concern rather than a secondary consideration.

The counter-evidence warrants close attention. Menon et al. (2025), in a controlled experiment with AI-generated educational podcasts, found that embedding interactive reflection prompts into the audio disrupted learner flow and reduced enjoyment significantly compared to a standard listening mode. Crucially, the disruption did not arise from the audio format itself but from a deployment decision that imposed concurrent interactive demands, violating the segmenting principle at the level of instructional design

ISSN 2338-4778 (Print)
ISSN 2548-4192 (Online)

rather than at the level of tool architecture. The architecture itself is sound and well-aligned with the segmenting and cognitive load management principles identified by R. E. Mayer & Fiorella (2021) and Plass & Jones (2005) specifically for L2 learners; whether it remains aligned depends on how the tool is deployed. This criterion is therefore conditionally met: the default sequential architecture is pedagogically well-designed, but Interactive Mode requires scaffolding as a prerequisite rather than an optional enhancement.

Criterion 3: Evidence Alignment

Regarding pacing, NotebookLM's alignment with the empirical literature is reasonable. The natural, measured pacing of its audio generation, combined with standard playback controls for learner-directed adjustment, maps plausibly onto the slow-narration condition identified by R. E. Mayer et al. (2014) as supporting comprehension gains in non-native speakers, a correspondence corroborated by Q. Liu & Zhao (2025)'s podcast findings. This pacing alignment is consistent with evidence that working memory load in L2 contexts is directly sensitive to input rate (W. Li et al., 2022) and with the empirical principle that instructional input must be calibrated to learner readiness to remain productive (Ramezanali et al., 2021). Menon et al.'s (2025) results reinforce the same point from the negative direction: departures from evidence-aligned instructional design produce measurable negative outcomes, confirming that the evidence base has real predictive force in these contexts.

The proficiency differentiation evidence presents a substantially different picture, with consequential implications for NotebookLM. Expert consensus in a Fuzzy Delphi validation by Erdiani et al. (2024) established proficiency-differentiated design as an evidence-grounded principle rather than a theoretical preference, and an analytical audit by Moulieswaran (2025) highlighted that effective AI platforms actively utilise adaptive algorithms to adjust content and difficulty based on individual performance. This adaptability is not merely a design preference but a cognitive necessity: a meta-analysis by Ramezanali et al. (2021) demonstrated that learner proficiency strictly moderates the need for multimedia support, with lower-proficiency learners requiring significantly more scaffolding than advanced learners, a pattern attributable to the greater cognitive resources lower-ability learners must allocate to process the same input (W. Li et al., 2022).

NotebookLM's uniform output cannot address this demand. The tool generates identical content regardless of learner proficiency, and the adaptation identified by the evidence base as consequential is left entirely to teacher curation and task design. Wu & Li (2025) demonstrated that proficiency-sensitive AI feedback produces measurable differences in engagement patterns across learner levels, reinforcing the case that uniform output is not a neutral design choice but a consequential gap. Zhuang et al.'s (2025) description of a proficiency-sensitive AI writing feedback system, in which lower-performing students benefit most from explicit feedback while higher-performing students leverage informative feedback, illustrates what evidence-aligned design looks like in practice.

The pacing dimension aligns with the slow-narration condition and learner control is available; the proficiency differentiation dimension is structurally absent. This criterion is partially satisfied, and the gap is not incidental: it is a direct consequence of the tool's architecture, which produces fixed output from fixed sources with no mechanism for learner-adaptive differentiation.

Criterion 4: Authentic Task Affordance

NotebookLM's strongest correspondence with EFL pedagogical principles lies here, rooted in the architectural feature that distinguishes it most clearly from open-domain generative tools. Because all output is anchored to specific, teacher-selected content, the knowledge base for any task is predictable and controllable (Jenkins & Hussain, 2026), which makes authentic task design considerably more tractable than in tools where learners cannot verify whether AI output is grounded in the intended content. Maranzana (2025) captured the mechanism through the concept of "authenticized" materials: AI-generated audio from teacher-curated sources functions as controlled, purpose-specific L2 input, in which the teacher's source selection determines the communicative register, vocabulary range, and content specificity of everything the tool produces. Lindade (2025) approached the same point from classroom practice, arguing that closed-source generation makes authentic task alignment achievable rather than aspirational.

Panday-Shukla (2024) provides the most precise account of this architecture's dual character: the Retrieval-Augmented Generation model that restricts all output to user-uploaded documents makes the system substantially more reliable and reduces hallucination risk compared to open-domain tools, but it simultaneously places active vigilance demands on teachers, who are required to fact-check and edit all generated content before it is shared with learners, since source-grounded generation can still introduce plausible but inaccurate material into the task environment. The broader literature confirms that goal-directed communicative tasks produce genuine engagement in AI-assisted contexts only when task structures anchor AI use to specific communicative purposes (Compagnoni, 2025; Emerson, 2024) and that task design quality, rather than technical features, determines the quality of emotional engagement with multimodal AI tools (Derakhshan & Park, 2026).

The full range of contextual evidence consistently positions teacher task design as the constitutive condition for authentic task affordance, understood ecologically as a relationship of possibility between learners and their environment that signals an opportunity for meaningful action (Blin, 2016; Van Lier, 2004), across diverse instructional settings. Authentic language use patterns are documented in self-directed informal learning contexts when communicative purpose is present (Guan et al., 2025). The authenticity condition is confirmed across SDG-aligned literacy curricula (Nurjamin et al., 2026), GenAI literacy in teacher education (Utami & Rohadi, 2026), AI use with young EFL learners in co-production tasks (Sadykova & Kayumova, 2025), and essay writing tasks where structure supports a genuine communicative goal (Zare et al., 2025).

The corpus evidence indicates that teacher task design functions as a requirement inherent to task-based pedagogy as established by González-Lloret & Ortega (2014) and Hafner et al. (2015), a condition the reviewed studies report across diverse instructional settings rather than for NotebookLM specifically. The tool's closed-domain architecture offers one distinct advantage: it forecloses the open-ended hallucinations that undermine task authenticity in open-domain tools. This criterion is met, with the qualification that teacher task design operates here as a constitutive pedagogical requirement rather than as compensation for a tool deficiency. The interpretive grounds for distinguishing this verdict from the partial verdicts are developed in the Discussion.

Criterion 5: Contextual Accessibility

Contextual accessibility is evaluated here across three dimensions: technical and infrastructural access, the ecological conditions under which affordances actualise in institutional settings (Van Lier, 2004), and the linguistic and cultural accessibility of the tool's default outputs. Effective multimodal tools must be deployable across the full range

ISSN 2338-4778 (Print)
ISSN 2548-4192 (Online)

of institutional conditions that characterise EFL higher education globally, including settings where connectivity is unstable, devices are limited, and teacher digital readiness is uneven (Mali & Salsbury, 2021), and their benefits must be realisable for the diverse learner populations that populate those settings (Mihaylova et al., 2022; Rahmanu & Molnár, 2024).

The empirical evidence base for this criterion is thinner than for any other. The only corpus study with data from an explicitly under-resourced deployment is Aremu et al. (2025), who documented reductions in hesitation markers, improvements in turn management, and increased speaking confidence among adult A1 EFL learners in Ecuador, confirming that contextual accessibility is a realisable outcome when institutional support addresses infrastructure barriers. Consistent with Van Lier's (2004) ecological account that affordances actualise only when the institutional environment supports meaningful engagement (Blin, 2016), Patty (2025) argued that accessibility must be evaluated against pedagogical readiness and institutional ecology rather than technical access alone, a position operationalised by Rashid et al. (2026) Active Learning and GenAI Synergy Framework, which embeds teacher readiness and infrastructure capacity as explicit accessibility variables.

Secondary evidence completes the picture across all three dimensions. On the infrastructural dimension, deployment reliability is constrained by lack of local technical support and the high labor of preparing and editing AI materials outside scheduled teaching hours (Sturgeon Delia, 2025), compounded in settings where physical disruptions impede sustained use (Guan et al., 2025) and where outdated hardware and limited connectivity make adoption structurally difficult (Moulieswaran, 2025), conditions that mirror historically persistent barriers in EFL contexts including Indonesia (Mali & Salsbury, 2021). On the linguistic and cultural dimension, AI-generated English defaults to Standard English varieties in ways that actively disadvantage non-dominant dialect speakers, meaning accessibility must account for whom the tool's outputs are designed for rather than only whether the tool can be technically accessed (Stewart & Zheng, 2024).

On the learner diversity dimension, the interactive audio features aligned with Universal Design for Learning principles specifically benefit neurodivergent learners by preventing information overload (Brown & Saurez, 2025). The corpus reports cost structure and platform consolidation as strengths (Moulieswaran, 2025; Sturgeon Delia, 2025), and reports internet dependency, curation labor, and linguistic default norms as barriers (Guan et al., 2025; Stewart & Zheng, 2024; Sturgeon Delia, 2025). This criterion is partially satisfied: the strengths apply at the level of cost and consolidation, while the barriers apply at the level of infrastructure and linguistic default. The empirical literature contains no study of NotebookLM's contextual accessibility situated in Indonesia, the institutional location of this review and a major EFL context. The interpretive significance of the linguistic and cultural dimension reported here is developed as the third critical boundary in the following section.

Table 4. Summary of Criterion Verdicts for NotebookLM Against the Five Evaluative Criteria.

Criterion	Verdict	Key Qualifier
Modal Diversity and Transduction Capacity	Partially met	Audio transduction confirmed (Yeo et al., 2025); visual output structurally limited to reorganisation (Cope & Kalantzis, 2020); gestural and spatial modes absent; audio defaults constrain linguistic diversity (Lindade, 2025; Yeo et al., 2025)
Cognitive Load Management	Conditionally met	Default sequential architecture aligned with cognitive load mitigation (Alisoy, 2025); Interactive Mode without scaffolding disrupts learner flow (Menon et al., 2025)
Evidence Alignment	Partially met	Audio pacing aligned with slow-narration condition (Q. Liu & Zhao, 2025; R. E. Mayer et al., 2014); proficiency differentiation

Criterion	Verdict	Key Qualifier
		structurally absent (W. Li et al., 2022; Ramezanali et al., 2021)
Authentic Task Affordance	Met	Source-grounding via a closed-domain system directly supports authentic task design (Jenkins & Hussain, 2026; Maranzana, 2025); teacher task design is constitutively required.
Contextual Accessibility	Partially met	Cost structure is favorable, but internet dependency and curation labor are real barriers (Aremu et al., 2025; Sturgeon Delia, 2025); UDL-aligned audio significantly benefits neurodivergent learners (Brown & Saurez, 2025)

4. Discussion

The criterion verdicts reported in the Results section reveal a consistently conditional correspondence between NotebookLM's features and EFL pedagogical principles. This Discussion addresses RQ2 by identifying three critical boundaries that emerge inductively from the same corpus, accounting for the conditions under which that correspondence weakens or fails. The section concludes by integrating the findings across the five criteria and three boundaries, positioning them against prior literature, and deriving implications for practice and research.

Critical Boundaries of NotebookLM's Affordances

The criterion analysis confirmed a meaningful but consistently conditional correspondence between NotebookLM's features and EFL pedagogical principles. Three critical boundaries emerge from the same corpus. The structural constraints of source-grounded generation connect directly to the partial verdicts on modal diversity and evidence alignment; the suppression of Critical Framing intensifies the tension already identified between cognitive load management and authentic task affordance; and the native-speakerist audio default extends the unresolved dimensions of both modal diversity and contextual accessibility. These are not peripheral limitations but structural and pedagogical conditions that define the range of contexts and tasks for which the tool can be deployed responsibly.

Boundary 1: Structural Constraints and Accuracy Risk

The architecture that produces NotebookLM's most valued affordances simultaneously imposes its most significant constraints. Source-grounded generation operates through a retrieval-augmented architecture in which all output is generated exclusively from uploaded documents, and this design choice simultaneously defines the limits of what the tool can produce (Albrecht-Crane, 2025; Reyna, 2025). Source-boundedness reduces generic hallucinations but introduces a different epistemic risk: learners may develop a false sense of comprehensiveness that obscures knowledge gaps beyond the uploaded documents, while copyright restrictions and cost barriers limit which sources can be legitimately uploaded in the first place (Reyna, 2025). The risk is not confined to the content scope. In multilingual contexts, the architecture produces linguistic instability, with critical words omitted when the tool operates in a language other than English (Venhrina, 2025), a finding that carries particular weight for EFL settings where source materials are not exclusively in English.

The pedagogical burden this imposes on educators is substantial: high-quality source curation is not an optional enhancement, but the primary determinant of the tool's educational value (Jenkins & Hussain, 2026), and hallucination awareness must be treated as a constitutive teacher competency rather than a supplementary skill (Dewi, 2025; Panday-Shukla, 2024). Where that oversight is applied thoughtfully, the closed-domain system can streamline complex workflows such as standards alignment (Willner, 2025);

ISSN 2338-4778 (Print)
ISSN 2548-4192 (Online)

where it is absent, the risks compound rather than diminish. The accuracy risk dimension extends across multiple study types, giving it corpus-wide rather than isolated evidential weight. Academically plausible misrepresentation is the specific form this risk takes in NotebookLM: students have reported that AI output shaped their understanding of a topic in ways they later recognised as distorted (Soge & Mbato, 2025), audio content embellishes source material in ways that diverge from what the documents actually say despite source-grounding (Yeo et al., 2025), and accuracy limitations are constitutive of the audio generation process rather than incidental to it (Lindade, 2025).

Framed through Van Lier's (2004) ecological account, which holds that a tool's structural properties delimit affordance actualisation regardless of user intent (Blin, 2016), this boundary establishes that source-grounded generation shifts the form of accuracy risk rather than eliminating it. The criteria most directly affected are modal diversity, where the modal range is bounded by what uploaded sources contain; evidence alignment, where proficiency-adaptive generation is architecturally foreclosed; and authentic task affordance, where task authenticity is threatened precisely when the conditions under which C4 is met, namely teacher curation, fact-checking, and source quality, cannot be assumed. The verdict that Criterion 4 is met is therefore conditional on the oversight described by Panday-Shukla (2024): in its absence, the architectural advantage of source-grounding does not translate into authentic task conditions but into a plausible facsimile of them.

Boundary 2: Cognitive Offloading and the Suppression of Critical Framing

The source-grounded architecture examined in Boundary 1 produces not only accuracy risk but a second pedagogical consequence that operates through a different mechanism: by compressing and organising content with minimal friction, the same architecture that reduces hallucination risk simultaneously reduces the interpretive resistance that deep learning requires. A productive reduction in cognitive load removes extraneous processing demands that compete with comprehension for limited working memory resources (R. E. Mayer et al., 2014; R. E. Mayer & Fiorella, 2021). A counterproductive one removes the interpretive struggle that makes comprehension deep rather than superficial. NotebookLM's compression architecture is well designed for the first outcome and carries structural risk for the second (Albrecht-Crane, 2025; Alisoy, 2025). The mechanism is specific: statistical text compression combined with a frictionless consumer interface removes the productive resistance that schema formation and critical reading competence require, a structural feature rather than an incidental one (Albrecht-Crane, 2025).

The vulnerability this creates falls disproportionately on learners whose proficiency limits their ability to evaluate formal academic prose, and qualitative data from Indonesian EFL master's students confirms that academically plausible AI output can shape thinking in ways learners only later recognise as distorted (Soge & Mbato, 2025). At the task level, the pattern is consistent: the faster AI produces organised content, the less students engage with source material themselves (Rebolledo et al., 2025), and learner agency increases when AI processes require active rather than passive engagement (Fang et al., 2025). This substitution tendency is corroborated across composing, presentation design, and writing contexts, where cognitive engagement in GenAI-assisted tasks constitutes a qualitatively different activity from traditional engagement in ways that require explicit instructional attention (Kang et al., 2026; Lee et al., 2025; M. Liu et al., 2024), and where the question of where human agency sits in relation to AI efficiency is constitutive of what any task actually practises (Jiang, 2024).

What this boundary threatens most directly is Critical Framing as Cope & Kalantzis (2009) defined it: the stage in the Multiliteracies pedagogical cycle where learners interrogate meaning designs for their social and cultural contexts, purposes, and effects. Frictionless compression provides no natural resistance to provoke that interrogation, and deliberate instructional design alongside programs of critical AI literacy are required to reintroduce the analytical distance the tool removes (Cope et al., 2025). The problem is cross-modal rather than text-specific, with reduced engagement with interpretive demands documented for AI-generated visual content as well (Djekidel & Rogti, 2025). Without deliberate critical framing, authentic task affordance risks becoming surface task completion and cognitive load management risks reducing germane alongside extraneous load when interpretive friction disappears entirely.

Boundary 3: Native-Speakerism and Linguistic Equity

The reinforcement of native-speakerist norms in NotebookLM is a demonstrated effect rather than an inferred risk. Its AI hosts default exclusively to American or global-north accents and do not respond to user prompts for accent diversity (Lindade, 2025; Yeo et al., 2025), and this default is not acoustically neutral: AI-generated English actively reproduces and normalises hegemonic linguistic conventions in ways that position learners in relation to a standard built into the tool's output by default, regardless of their own linguistic backgrounds or the diversity of the EFL classrooms they inhabit (Stewart & Zheng, 2024). The structural dimension of this problem extends beyond accent to authority: when tool outputs carry implicit linguistic authority, learner agency in linguistic choice-making is constrained because learners may not feel entitled to question what the tool produces as standard (Godwin-Jones, 2024). This tendency further complicates the multiliteracies landscape, as AI systems that default to dominant norms narrow rather than expand the range of linguistic resources learners consider legitimate meaning-making tools (Cope et al., 2025; Nguyen & Nguyen, 2025).

The only counter-strategy documented in the corpus is deliberate selection of linguistically and culturally diverse source texts as a partial corrective to audio defaults, a practice Maranzana (2025) captures under the concept of "authenticized" materials. This mitigation depends entirely on teacher awareness and the willingness to actively curate against the tool's default trajectory, and the tool provides no structural support for it, consistent with the pattern identified across all three boundaries. The criteria most directly affected are contextual accessibility, where accessibility must encompass linguistic and cultural dimensions rather than only technical access, and modal diversity and transduction, where The New London Group's (1996) principle that diverse linguistic and cultural resources are equally valid Available Designs sits in tension with audio output that defaults to dominant English varieties. While the evidence for this boundary is more direct than previously documented in the literature, the extent to which these defaults affect learning outcomes across different EFL contexts remains an open empirical question.

Integrating the Findings

This section draws the findings together across three movements: an analysis of cross-criteria patterns and the dominant mediating variable they share, a positioning of the review's findings against prior literature on GenAI in EFL and multimodal CALL, and a consolidated set of implications for practitioners and researchers.

ISSN 2338-4778 (Print)
ISSN 2548-4192 (Online)

Patterns Across the Five Criteria

The five criterion verdicts, read together, describe a tool whose pedagogical correspondence is real and theoretically accountable but consistently conditional. No criterion is met without qualification: four are partially or conditionally satisfied, and the criterion for authentic task affordance is met with a condition intrinsic to task-based pedagogy rather than specific to this tool. Teacher mediation is the variable that runs through all five criteria without exception. As recent studies emphasise, generative AI tools become educationally meaningful only through guided interaction and explicit pedagogical framing (Utami & Rohadi, 2026).

Modal diversity requires teacher-orchestrated task sequences that direct learners across modes sequentially rather than simultaneously, since the tool's architecture offers modalities but does not determine how or in what order learners engage them (Maranzana, 2025; Yeo et al., 2025); authentic task affordance requires teacher task design alongside active fact-checking of source-grounded outputs (Panday-Shukla, 2024); cognitive load management depends on scaffolding decisions around Interactive Mode, since without explicit guidance multimodal and AI-driven tasks can overwhelm learners or produce misinterpretations (Labibah et al., 2025; Ren et al., 2026); evidence alignment for proficiency differentiation requires teacher curation to compensate for what the tool does not structurally provide; and contextual accessibility requires source preparation, infrastructure management, and deliberate counter-curation against native-speakerist audio defaults, since NotebookLM's AI hosts retain American or global-north accents and do not respond to prompts for accent diversity (Lindade, 2025; Yeo et al., 2025). The practical implication is substantive: the tool is not an autonomous learner-facing resource but a teacher-dependent instrument whose value is proportional to the quality of the instructional design surrounding it (Dewi, 2025; Utami & Rohadi, 2026).

The verdict on authentic task affordance warrants a brief explanation, because it might otherwise seem inconsistent with the conditional nature of the other four. The criteria that are partially or conditionally met share the same structural logic: the tool falls short of what the criterion demands, and teacher effort compensates for a gap. An authentic task affordance is different in kind. By utilising a retrieval-augmented generation model that strictly restricts generative output to user-provided source documents, NotebookLM directly addresses the criterion's core demand rather than approximating it, significantly reducing hallucination risk while generating highly relevant "authenticized" input (Jenkins & Hussain, 2026; Maranzana, 2025). Teacher task design is constitutively required, but by what task-based language teaching has always required, not because the tool is deficient. As summarised in Table 4 in the Results section above, the criterion verdicts describe a tool whose pedagogical correspondence is real and theoretically accountable but consistently conditional.

Positioning Against Prior Literature

Against the ChatGPT-centric research that currently dominates the GenAI-EFL field, NotebookLM represents a qualitatively distinct affordance profile rather than an incremental variation on the same tool type. Its source-grounded architecture makes it more controllable, more teacher-dependent, and better suited to specific task types than open-domain tools. In contrast, the same architecture introduces constraints that open-domain tools do not impose in the same form: bounded generativity, instability in non-English source processing (Venhrina, 2025), and substantial curation labor. The choice between them should therefore be task-specific rather than categorical. Against the broader multimodal CALL literature, the picture is more qualified. The scholarly consensus on multimodal digital tools concerns platforms that engage multiple modes with real

pedagogical function across all of them, not tools that are rich in one mode and absent in others (Hafner et al., 2015; Mihaylova et al., 2022; Rahmanu & Molnár, 2024). NotebookLM satisfies that standard for audio; the remaining modes are not comparably addressed.

On the question of accuracy, source-grounding partially addresses hallucination concerns but does not resolve them. Across study types, the evidence converges on a consistent finding: source-grounding shifts the form of accuracy risk rather than eliminating it, with instability documented in non-English source processing (Venhrina, 2025), academically plausible misrepresentation confirmed through qualitative data (Soge & Mbato, 2025), and audio embellishment observed despite source constraints (Yeo et al., 2025). The risk is narrower in scope than in open-domain tools. However, it is not absent, and learners who cannot evaluate formal academic language are particularly vulnerable to it, a finding with particular relevance for EFL contexts where proficiency variation is the norm rather than the exception.

The cognitive offloading risk identified in Boundary 2 is not unique to NotebookLM. However, its frictionless compression architecture intensifies a tendency documented across the broader GenAI-EFL literature: the faster AI produces organised content, the less students engage with source material themselves (Rebolledo et al., 2025), and the substitution of AI processing for learner processing has been confirmed across writing, composing, and presentation contexts (Kang et al., 2026; M. Liu et al., 2024). What distinguishes NotebookLM is that its design optimises precisely for frictionless content delivery, making the suppression of Critical Framing a structural tendency rather than an incidental risk.

In contrast to the multiliteracies literature, this positions the tool in tension with the critical framing stage of the pedagogical cycle identified by Cope & Kalantzis (2000) as essential for deep learning. The native-speakerist defaults identified in Boundary 3 are similarly consistent with a pattern documented across AI language tools more broadly (Godwin-Jones, 2024; Stewart & Zheng, 2024), but the absence of any user-adjustable accent or variety parameter in NotebookLM makes this boundary more structurally fixed than in tools that allow greater output customisation.

Implications for Practice

Deployment decisions for EFL practitioners should be guided by the tool's specific features rather than by a general assessment of its potential. The Audio Overview is the strongest feature for EFL purposes and is appropriate for listening comprehension, L2 input preparation, and anxiety reduction where the source register suits the learner level, but its use requires explicit instruction on accuracy limitations: academically plausible misrepresentation is possible even within the source-grounded domain (Soge & Mbato, 2025; Yeo et al., 2025). Interactive mode carries real promise for active listening practice but requires scaffolding as a prerequisite rather than an enhancement, since interactive elements introduced without adequate instructional design disrupt learner flow rather than deepen engagement (Menon et al., 2025). The closed-domain architecture suits controlled, content-specific tasks such as source-specific reading preparation, seminar preparation, and content-based listening, while it is poorly suited to open-ended creative writing or exploratory synthesis.

Two classroom scenarios illustrate how these features translate into practice. In a content-based listening sequence, a teacher uploads a curated set of reading passages on a single topic, generates an Audio Overview, and assigns it as pre-class listening preparation, then opens the following session by asking learners to identify two claims in the audio that the source passages do not actually support. This task converts the accuracy risk documented in Boundary 1 into an object of analysis rather than an unexamined hazard

ISSN 2338-4778 (Print)
ISSN 2548-4192 (Online)

(Soge & Mbato, 2025; Yeo et al., 2025). In a seminar preparation sequence, a teacher who anticipates that lower-proficiency learners will struggle with uniform output generates the Audio Overview, then provides a tiered task sheet in which lower-proficiency learners answer scaffolded comprehension questions while higher-proficiency learners evaluate the framing choices in the audio discussion. This division compensates at the level of task design for the proficiency differentiation that the tool does not provide architecturally (W. Li et al., 2022; Ramezanali et al., 2021), and it directs each learner group toward the modal engagement appropriate to its level rather than presenting a single undifferentiated resource.

Across all contexts, source curation must deliberately include linguistically and culturally diverse materials to counter audio defaults that privilege American and global-north accents and fail to respond to prompts for diversity (Lindade, 2025; Yeo et al., 2025). In a multilingual EFL setting, a teacher might pair a source text written in a dominant English variety with a second source authored in a regional or postcolonial variety, then ask learners to compare how the Audio Overview renders each, using the contrast to make the tool's default norm visible and available for discussion rather than absorbed without comment.

Beyond task design and source curation, the cognitive offloading risk documented in Boundary 2 requires a third layer of practitioner response: explicit critical AI literacy instruction that prepares learners to engage with AI-generated content analytically rather than deferentially, interrogating the framing, selection, and omissions in tool outputs rather than accepting them as authoritative summaries of source material (Albrecht-Crane, 2025; Cope et al., 2025). This is not a supplementary enhancement but a prerequisite for responsible deployment, since the frictionless design that makes the tool accessible is the same design feature that suppresses the productive interpretive friction that critical reading requires.

Directions for Future Research

The empirical agenda this review identifies follows directly from the gaps in the corpus. The most consequential gap is the complete absence of empirical data on NotebookLM's text-based features in EFL reading or writing contexts: briefing documents, FAQ generation, and study guides are among the tool's most used features in practice, yet their effects on comprehension, writing preparation, or vocabulary development have not been measured with objective language outcomes across all 1,862 records retrieved. The methodological profile of the corpus warrants equal attention: self-report and perception data dominate, single-semester timeframes cannot distinguish novelty effects from sustained learning, and longitudinal designs with objective pre- and post-measures are needed. Context gaps are equally substantial, with K-12 and secondary education, non-Asian EFL settings, and non-English-L1 learner populations all significantly underrepresented, each of which marks a domain where the findings of this review may not generalise without further empirical work. Indonesia warrants particular attention as a priority research context.

As the institutional location of this review and one of the world's largest EFL populations, it is absent from the empirical literature on NotebookLM's contextual accessibility, and the infrastructure, connectivity, and teacher readiness conditions that characterise Indonesian EFL higher education make it a consequential test case for the accessibility claims this review has assessed only at a theoretical level. A final priority concerns linguistic equity: the extent to which native-speakerist audio defaults measurably affect EFL learners' linguistic self-concept and productive language use across different institutional settings remains an open empirical question, since the evidence reviewed here

establishes the structural tendency but cannot speak to its learning outcome consequences across the diverse populations NotebookLM is increasingly deployed to serve.

5. Conclusion

This review contributes a theoretically grounded evaluative framework for assessing AI multimodal tools in EFL pedagogy, one that is criterion-driven, replicable in its evaluative logic, and applicable beyond NotebookLM to any tool whose affordances need to be mapped against established pedagogical principles. Applied to a corpus of 50 articles, the framework confirms that NotebookLM's alignment with EFL pedagogical principles is real but consistently conditional: strongest when the source-grounded architecture directly supports authentic task design, and most constrained when the tool's generative defaults conflict with the demands of proficiency differentiation, Critical Framing, and linguistic equity.

Across all five criteria and all three boundaries, teacher mediation is not an enhancement but the constitutive condition for affordance actualisation. The tool functions as a teacher-dependent instrument, and its pedagogical value is determined by the instructional design built around it rather than solely by its technical capabilities. The most consequential practical priority that follows is therefore clear: responsible deployment requires, above all, explicit critical AI literacy instruction as a prerequisite, alongside deliberate source curation and scaffolded task design, since their absence does not merely limit the tool's effectiveness but actively introduces the risks of cognitive offloading, uncritical acceptance of plausible misrepresentation, and the reinforcement of hegemonic linguistic norms documented across the corpus.

Several limitations of this review define its scope. As a narrative literature review, the analysis reflects interpretive synthesis rather than statistical aggregation, and single-reviewer screening, while consistent with the NLR design and supported by documented criteria, introduces the possibility of selection judgments that a second reviewer might have made differently. A further methodological limitation concerns the use of Google NotebookLM as a pre-reading orientation tool at Stage 3 of the screening process. Although its role was restricted to content summarisation under documented human oversight, the possibility that the tool's summaries inadvertently shaped the reviewer's reading of borderline articles cannot be fully excluded.

The search strategy did not include MLA International Bibliography or LLBA, and articles indexed in those sources but not retrieved may have altered some criterion verdicts at the margins. The corpus is shaped by what the field has produced in a short window: self-report and perception data dominate, the literature is concentrated in higher education and Asian EFL contexts, and Indonesia, the institutional location of this review and one of the world's largest EFL populations, is absent from the empirical literature on NotebookLM's contextual accessibility. NotebookLM is also an actively developing platform, and the features documented in the reviewed literature may have changed since then. These constraints do not undermine the analysis, but they define its generalisability: the review establishes a theoretically grounded baseline for a tool whose empirical literature is still forming, and its primary contribution, the evaluative framework, is offered as a starting point for the institutional conversations that responsible AI integration in EFL education has not yet had.

ISSN 2338-4778 (Print)
ISSN 2548-4192 (Online)

6. References

Adel, A., & Alani, N. (2025). Can generative AI reliably synthesise literature? Exploring hallucination issues in ChatGPT. *AI & SOCIETY*, *40*(8), 6799–6812. https://doi.org/10.1007/s00146-025-02406-7

Albrecht-Crane, C. (2025). Thinking Smarter, not Harder? Google NotebookLM's Misalignment Problem in Education. *Proceedings of the 43rd ACM International Conference on Design of Communication*, 121–127. https://doi.org/10.1145/3711670.3764628

Alisoy, H. (2025). Can NotebookLM Support English Language Learners? A Theoretical Perspective on AI Tools in Education. *Porta Universorum*, *1*(6), 25–55. https://doi.org/10.69760/portuni.0106003

Aremu, I. O., Espinosa, K. E. P., Cañizares, F. I., & Tenesaca, J. R. B. (2025). AI-powered podcast interventions for enhancing speaking skills in English Language Teaching (ELT) Adult A1 students. *Arandu UTIC*, *12*(3), 2702–2723. https://doi.org/10.69639/arandu.v12i3.1509

Bijker, R., Merkouris, S. S., Dowling, N. A., & Rodda, S. N. (2024). ChatGPT for Automated Qualitative Research: Content Analysis. *Journal of Medical Internet Research*, *26*(1), e59050. https://doi.org/10.2196/59050

Blin, F. (2016). Towards an "ecological" CALL theory: Theoretical perspectives and their instantiation in CALL research and practice. In L. Murray & F. Farr (Eds.), *The Routledge handbook of language learning and technology*. Routledge.

Brown, V., & Saurez, D. (2025). NotebookLM: Revolutionizing Learning for Students with Neurodivert Challenges using AI and Universal Design Principles. *FDLA Journal*, *9*(1). https://nsuworks.nova.edu/fdla-journal/vol9/iss1/22

Christoforou, M. (2025). Gen AI-assisted multimodal meaning *Design*: Exercising a pedagogic metalanguage of transposition. *Pedagogies: An International Journal*, 1–25. https://doi.org/10.1080/1554480X.2025.2522884

Ci, F., & Jiang, L. (2025). The integration of generative artificial intelligence into digital multimodal composing in second language classrooms: A scoping review from the perspective of tasks. *Digital Applied Linguistics*, *2*, 102939. https://doi.org/10.29140/dal.v2.102939

Compagnoni, I. (2025). Pedagogical Implications of AI-Enhanced Digital Storytelling in EFL Education. *International Journal of Linguistics*, *17*(5), 1–24. https://doi.org/10.5296/ijl.v17i5.22773

Cope, B., & Kalantzis, M. (Eds.). (2000). *Multiliteracies: Literacy Learning and the Design of Social Futures* (0 ed.). Routledge. https://doi.org/10.4324/9780203979402

Cope, B., & Kalantzis, M. (2009). "Multiliteracies": New Literacies, New Learning. *Pedagogies: An International Journal*, *4*(3), 164–195. https://doi.org/10.1080/15544800903076044

Cope, B., & Kalantzis, M. (2020). *Making Sense: Reference, Agency, and Structure in a Grammar of Multimodal Meaning* (1st ed.). Cambridge University Press. https://doi.org/10.1017/9781316459645

Cope, B., Kalantzis, M., & Zapata, G. C. (2025). Language Learning after Generative AI. In G. C. Zapata, *Generative AI Technologies, Multiliteracies, and Language Education* (1st ed., pp. 14–39). Routledge. https://doi.org/10.4324/9781003531685-2

Derakhshan, A., & Park, Y. (2026). The Role of Multimodal AI Technologies in EFL Studentsrsquo; Perceived Positive and Negative Achievement Emotions: An Existential Positive Psychology (EPP) Perspective.)3(*17*, ستارهای زبانی. https://doi.org/10.48311/lrr.2025.118514.83043

Dewi, F. (2025). Leveraging Generative AI in ELT: Teachers' Integration Strategies and Pedagogical Adaptations. *JOLLT Journal of Languages and Language Teaching*, *13*(2), 600–615. https://doi.org/10.33394/jollt.v13i2.13670

Djekidel, O. N. E., & Rogti, M. (2025). AI-Generated Animated Visuals and Their Effect on EFL Learning Outcomes: A Systematic Review with Pedagogical Implications for Practice. *Journal of Studies in Language, Culture and Society (JSLCS)*, *8*(4), 257–278.

Emerson, N. (2024). AI-enhanced collaborative story writing in Japanese university EFL classes. *Technology in Language Teaching & Learning*, *6*(3), 1764. https://doi.org/10.29140/tltl.v6n3.1764

Erdiani, N., Hashim, H., & Sulaiman, N. A. (2024). Application of Fuzzy Delphi Method to Identify the Construct for Designing and Developing the Multimodal Learning Framework for Writing Skills in ESL Context. *International Journal of Communication Networks and Information Security (IJCNIS)*, *16*(1 (Special Issue)), 1085–1097.

Fan, K. (2025). Generative AI and Second Language Vocabulary Processing: A Cognitive Study of Chinese EFL Learners. *Journal of Humanities and Social Sciences Studies*, *7*(7), 103–111. https://doi.org/10.32996/jhsss.2025.7.7.12

Fang, X., Ng, D. T. K., Li, D. X., Tao, Y. Y., Wu, H. K., & CHU, S. K. W. (2025). GenAI-supported digital multimodal composing: Facilitating EFL students' GenAI literacy, writing engagement, motivation, and abilities. *AI, Brain and Child*, *1*(1), 23. https://doi.org/10.1007/s44436-025-00020-4

Ferrari, R. (2015). Writing narrative style literature reviews. *Medical Writing*, *24*(4), 230–235. https://doi.org/10.1179/2047480615Z.000000000329

Godwin-Jones, R. (2024). Distributed agency in second language learning and teaching through generative AI - Language Learning and Technology. *Language Learning & Technology*, *28*(2). https://www.lltjournal.org/item/10125-73570/

González-Lloret, M., & Ortega, L. (Eds.). (2014). *Technology-mediated TBLT: Researching Technology and Tasks* (Vol. 6). John Benjamins Publishing Company. https://doi.org/10.1075/tblt.6

Green, B. N., Johnson, C. D., & Adams, A. (2006). Writing narrative literature reviews for peer-reviewed journals: Secrets of the trade. *Journal of Chiropractic Medicine*, *5*(3), 101–117. https://doi.org/10.1016/S0899-3467(07)60142-6

Gregory, A. T., & Denniss, A. R. (2018). An Introduction to Writing Narrative and Systematic Reviews—Tasks, Tips and Traps for Aspiring Authors. *Heart, Lung and Circulation*, *27*(7), 893–898. https://doi.org/10.1016/j.hlc.2018.03.027

Guan, L., Zhang, E. Y., & Gu, M. M. (2025). Examining generative AI–mediated informal digital learning of English practices with social cognitive theory: A mixed-methods study. *ReCALL*, *37*(3), 315–331. https://doi.org/10.1017/S0958344024000259

Hafner, C. A., Chik, A., & Jones, R. H. (2015). Digital literacies and language learning. *Language Learning & Technology*, *19*(3), 1–7. https://doi.org/10.64152/10125/44426

Huang, H.-W., Chen, X., & Sankey, A. (2024). Leveraging Multimodal GenAI Chatbots in EFL Learning: Learning Attitudes and User Experiences. *Proceedings of the 2024 International Conference on Artificial Intelligence and Teacher Education*, 22–29. https://doi.org/10.1145/3702386.3702406

Jenkins, M. R., & Hussain, F. W. (2026). Judging a book without a cover: An exploratory conceptual evaluation of notebooklm as a textbook alternative using established textbook adoption criteria. *Marketing Education Review*, *0*(0), 1–14. https://doi.org/10.1080/10528008.2026.2618166

Jewitt, C., Bezemer, J., & O'Halloran, K. (2016). *Introducing Multimodality* (0 ed.). Routledge. https://doi.org/10.4324/9781315638027

ISSN 2338-4778 (Print)
ISSN 2548-4192 (Online)

Jiang, J. (2024). When generative artificial intelligence meets multimodal composition: Rethinking the composition process through an AI-assisted design project. *Computers and Composition*, *74*, 102883. https://doi.org/10.1016/j.compcom.2024.102883

Kang, C., Yen, P., Guo, C., Li, M., Mai, W., Yen, C., & Xie, Y. (2026). Exploring the Pedagogical Affordances and Constraints of AI-Supported Presentation Design in EFL Contexts (Pt. 102-114). *Research Journal of English Language and Literature (RJELAL)*, *14*(1). https://doi.org/10.33329/rjelal.14.1.102

Kress, G. (2009). *Multimodality* (0 ed.). Routledge. https://doi.org/10.4324/9780203970034

Kress, G., & Van Leeuwen, T. (2001). *Multimodal discourse: The modes and media of contemporary communication*. Arnold ; Oxford University Press.

Kress, G., & Van Leeuwen, T. (2020). *Reading Images: The Grammar of Visual Design* (3rd ed.). Routledge. https://doi.org/10.4324/9781003099857

Labibah, H. F., Mujiyanto, J., Rukmini, D., & Whidiyanto. (2025). AI-enhanced multimodal reading: Exploring its impact on efl learners' reading comprehension and engagement in Madrasah. *Indonesian EFL Journal*, *11*(3), 693–704. https://doi.org/10.25134/v94fzg98

Lai, C., & Li, G. (2011). Technology and Task-Based Language Teaching: A Critical Review. *CALICO Journal*, *28*(2), 498–521.

Lee, Y.-J., Davis, R. O., & Choi, J.-I. (2025). Integrating generative AI into EFL writing: University students' strategies and perceptions. *Online Journal of Communication and Media Technologies*, *15*(4), e202541. https://doi.org/10.30935/ojcmt/17545

Li, D., Xia, S., & Guo, K. (2025). Using AI to enhance digital multimodal composing: EFL learners' semiotic decision-making, self-efficacy, enjoyment, and continuance intention. *Language Learning & Technology*, 1–15. https://doi.org/10.64152/10125/73636

Li, W., Yu, J., Zhang, Z., & Liu, X. (2022). Dual Coding or Cognitive Load? Exploring the Effect of Multimodal Input on English as a Foreign Language Learners' Vocabulary Learning. *Frontiers in Psychology*, *13*, 834706. https://doi.org/10.3389/fpsyg.2022.834706

Lindade, C. (2025). Talk and Talk: (Re) Introducing EFL Learners to Podcasts through NotebookLM. *eJournal*, *7*(27).

Liu, M., Zhang, L. J., & Biebricher, C. (2024). Investigating students' cognitive processes in generative AI-assisted digital multimodal composing and traditional writing. *Computers & Education*, *211*, 104977. https://doi.org/10.1016/j.compedu.2023.104977

Liu, Q., & Zhao, Y. (2025). The impact of NoteBookLM generated podcasts on EFL news listening comprehension, flow experience, and anxiety: A quasi-experimental study. *Proceedings of the 2nd Guangdong-Hong Kong-Macao Greater Bay Area Education Digitalization and Computer Science International Conference*, 665–670. https://doi.org/10.1145/3746469.3746573

Lo, C. K., Yu, P. L. H., Xu, S., Ng, D. T. K., & Jong, M. S. (2024). Exploring the application of ChatGPT in ESL/EFL education and related research issues: A systematic review of empirical studies. *Smart Learning Environments*, *11*(1), 50. https://doi.org/10.1186/s40561-024-00342-5

Mali, Y. C. G., & Salsbury, T. L. (2021). Technology integration in an Indonesian EFL writing classroom. *TEFLIN Journal - A Publication on the Teaching and Learning of English*, *32*(2), 243. https://doi.org/10.15639/http:/teflinjournal.v32i2/243-266

Maranzana, S. (2025, November 7). From Authentic to "Authenticized": Reconceptualizing Language Teacher Identity and AI-Driven Materials. *Innovation in Language Learning*.

Conference Proceedings. Innovation in Language Learning 2025. https://conference.pixel-online.net/library_scheda.php?id_abs=7665

Martin, R., & Johnson, S. (2023, July 12). *Introducing NotebookLM*. Google. https://blog.google/innovation-and-ai/technology/ai/notebooklm-google-ai/

Mayer, R. E., & Fiorella, L. (Eds.). (2021). *The Cambridge Handbook of Multimedia Learning* (3rd ed.). Cambridge University Press. https://doi.org/10.1017/9781108894333

Mayer, R. E., Lee, H., & Peebles, A. (2014). Multimedia Learning in a Second Language: A Cognitive Load Perspective. *Applied Cognitive Psychology*, *28*(5), 653–660. https://doi.org/10.1002/acp.3050

Menon, V., Cherney, A., Cloude, E. B., Zhang, L., & Do, T. D. (2025). Evaluating the Impact of LLM-guided Reflection on Learning Outcomes with Interactive AI-Generated Educational Podcasts. In J. Wilson, C. Ormerod, & M. Beiting Parrish (Eds.), *Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con): Full Papers* (pp. 99–106). National Council on Measurement in Education (NCME). https://aclanthology.org/2025.aimecon-main.11/

Mihaylova, M., Gorin, S., Reber, T. P., & Rothen, N. (2022). A Meta-Analysis on Mobile-Assisted Language Learning Applications: Benefits and Risks. *Psychologica Belgica*, *62*(1), 252. https://doi.org/10.5334/pb.1146

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. *Journal of Clinical Epidemiology*, *62*(10), 1006–1012. https://doi.org/10.1016/j.jclinepi.2009.06.005

Moulieswaran, N. (2025). Integrating Generative AI-Powered Tools for Vocabulary Instruction: Strategies for Indian ESL Learners – A Conceptual Analysis. *Indian Journal of Language and Linguistics*, *6*(4), 13–23. https://doi.org/10.54392/ijll2542

Nguyen, T. N. Q., & Nguyen, T. T. (2025). AI literacy in language learning through digital multimodal composing: A systematic review. *VNU Journal of Foreign Studies*, *41*(3), 142–156. https://doi.org/10.63023/2525-2445/jfs.ulis.5517

Nurjamin, A., Nurjamin, L. R., Fajriah, Y. N., Nurjamin, A. K., & Firdaus, H. A. (2026). Integrating Generative Artificial Intelligence (AI)-Based Multimodal Learning in Education to Enhance Literacy Aligned with Sustainable Development Goals (SDGs). *ASEAN Journal of Educational Research and Technology*, *5*(1), 71–88.

Panday-Shukla, P. (2024). Google NotebookLM: Creating Engaging Language Lessons. *The FLTMAG*. https://doi.org/10.69732/DZWQ2675

Patty, J. (2025). Integrating Critical Literacy and Multimodal Pedagogy in English Language Education. *J-Shelves of Indragiri (JSI)*, *7*(2), 176–198. https://doi.org/10.61672/jsi.v7i2.3341

Paul, M., & Dhami, D. (2025, September 8). *6 ways to use NotebookLM to master any subject*. Google. https://blog.google/innovation-and-ai/models-and-research/google-labs/notebooklm-student-features/

Pautasso, M. (2019). The Structure and Conduct of a Narrative Literature Review. In M. Shoja, A. Arynchyna, M. Loukas, A. V. D'Antoni, S. M. Buerger, M. Karl, & R. S. Tubbs (Eds.), *A Guide to the Scientific Career* (1st ed., pp. 299–310). Wiley. https://doi.org/10.1002/9781118907283.ch31

Plass, J. L., & Jones, L. C. (2005). Multimedia Learning in Second Language Acquisition. In R. Mayer (Ed.), *The Cambridge Handbook of Multimedia Learning* (1st ed., pp. 467–488). Cambridge University Press. https://doi.org/10.1017/CBO9780511816819.030

Rahmanu, I. W. E. D., & Molnár, G. (2024). Multimodal immersion in English language learning in higher education: A systematic review. *Heliyon*, *10*(19), e38357. https://doi.org/10.1016/j.heliyon.2024.e38357

ISSN 2338-4778 (Print)
ISSN 2548-4192 (Online)

Ramezanali, N., Uchihara, T., & Faez, F. (2021). Efficacy of Multimodal Glossing on Second Language Vocabulary Learning: A Meta-analysis. *TESOL Quarterly*, *55*(1), 105–133. https://doi.org/10.1002/tesq.579

Rashid, S., Malik, S., & Ghauri, F. (2026). The Active Learning–GenAI Synergy Framework: Ethical Integration of Generative AI in EFL/ESL Writing in Resource-Constrained Contexts. *F1000Research*, *14*, 1210. https://doi.org/10.12688/f1000research.171435.2

Rebolledo, R., Veas, C., & Gisbert, M. (2025). Refining EFL Essay Writing with Multimodal GenAI Tools. *Computer-Assisted Language Learning Electronic Journal*, *26*(6), 146–171.

Ren, Y., Wong, W. L., Barrot, J., & Zhang, L. J. (2026). Learning Styles, Engagement and Anxiety in AI-Mediated Writing: A Multimodal Feedback Study. *International Journal of Applied Linguistics*. https://doi.org/10.1111/ijal.70120

Reyna, J. (2025, July 24). *The Potential of Google NotebookLM for Teaching and Learning*. eLearn 2025.

Sadykova, G., & Kayumova, A. (2025). AI-Powered Image and Audio Generators for Very Young EFL Learners. *Iranian Journal of Language Teaching Research*, *13*(2), 1–24.

Snyder, H. (2019). Literature review as a research methodology: An overview and guidelines. *Journal of Business Research*, *104*, 333–339. https://doi.org/10.1016/j.jbusres.2019.07.039

Soge, E. P., & Mbato, C. L. (2025). *The perceived impacts of generative AI on indonesian EFL master's students' motivation in academic writing*. https://e-journal.uniflor.ac.id/index.php/JPM/article/view/6120

Stewart, N., & Zheng, Y. (Danson). (2024). Generative AI's recolonization of EFL classrooms: The case of continuation writing. *Australian Review of Applied Linguistics*, *47*(3), 383–409. https://doi.org/10.1075/aral.24091.ste

Sturgeon Delia, C. (2025). A Pilot Practitioner's Inquiry into Students' Reflections on AI-Generated Podcasts in Higher Education. *Education Sciences*, *15*(12), 1680. https://doi.org/10.3390/educsci15121680

Sukhera, J. (2022). Narrative Reviews: Flexible, Rigorous, and Practical. *Journal of Graduate Medical Education*, *14*(4), 414–417. https://doi.org/10.4300/JGME-D-22-00480.1

The New London Group. (1996). A Pedagogy of Multiliteracies: Designing Social Futures. *Harvard Educational Review*, *66*(1), 60–93. https://doi.org/10.17763/haer.66.1.17370n67v22j160u

Utami, D., & Rohadi, T. (2026). Integrating AI and Generative AI Literacy in English Teacher Education for Young Learners. *JEPAL (Journal of English Pedagogy and Applied Linguistics)*, *6*(2), 276–288. https://doi.org/10.32627/jepal.v6i2.1838

Van Lier, L. (Ed.). (2004). *The Ecology and Semiotics of Language Learning: A Sociocultural Perspective* (Vol. 3). Springer Netherlands. https://doi.org/10.1007/1-4020-7912-5

Venhrina, O. (2025). Do RAG systems provide an advantage? An empirical study of chatgpt and notebooklm effectiveness in generating academic tests. In *Transformation of the scientific area in the context of contemporary challenges: Scientific monograph*. Publishing House "Baltija Publishing." https://doi.org/10.30525/978-9934-26-631-7

Willner, L. S. (2025). AI-Powered Instructional Planning for Integrated Content, Literacy, and English Language Development for K-12 Multilingual Learners. *GATESOL Journal*, *34*(1), 17–34. https://doi.org/10.52242/gatesol.199

Wu, X., & Li, R. (2025). Perceived supports, achievement emotions, and engagement in multimodal GAI chatbot-assisted language learning: A sequential mixed-methods

study. *Computer Assisted Language Learning*, 1–24. https://doi.org/10.1080/09588221.2025.2569353

Yeo, M. A., Moorhouse, B. L., & Wan, Y. (2025). From Academic Text to Talk-Show: Deepening Engagement and Understanding with Google NotebookLM. *Teaching English as a Second or Foreign Language--TESL-EJ*, *28*(4). https://doi.org/10.55593/ej.28112int

Zare, J., Ranjbaran Madiseh, F., & Derakhshan, A. (2025). Generative AI and English essay writing: Exploring the role of ChatGPT in enhancing learners' task engagement. *Applied Linguistics*, 1–21.

Zhao, X., Chen, X., Rollins, M., Carratù, M., & Shallari, I. (2026). Empowering International Students' Listening and Speaking Skills via AI-Driven Audio in NotebookLM: Reflections Across Disciplines. *Proceedings of the 59th Hawaii International Conference on System Sciences*. https://scholarspace.manoa.hawaii.edu/items/d51cf229-8aba-4177-b3b4-1879cc7fa184

Zhuang, Y., Zhao, R., Xie, Z., & Yu, P. L. H. (2025). Enhancing language learning through generative AI feedback on picture-cued writing tasks. *Computers and Education: Artificial Intelligence*, *9*, 100450. https://doi.org/10.1016/j.caeai.2025.100450