Mirlam: Jurnal Pendidikan Bahasa dan Sastra Indonesia
e-ISSN: 2722-757X
Vol 7(2) (2026): 55-68 DOI: https://doi.org/10.30598/mirlamvol7no2hlm55-68

ChatGPT as a Grammar Tutor for Indonesian as a Foreign Language (BIPA): Evaluating Pedagogical Quality

Nia Budiana[1*], Didik Hartono[1], Widya Caterine Perdhani[1]

[1]Faculty of Cultural Studies, Universitas Brawijaya, Malang, Indonesia
*nia_budiana@ub.ac.id

Article Info

Submited: 09 August 2026
Accepted: 16 August 2026
Available Online: 19 August 2026
Published: 20 August 2026

Abstract

Learners of Bahasa Indonesia untuk Penutur Asing (BIPA), like language learners generally, consult ChatGPT for grammatical explanations outside class hours, yet the pedagogical quality of those explanations remains unexamined. This study evaluates ChatGPT (GPT-4o) explanations of four grammatical concepts that BIPA error-analysis research identifies as persistent sources of difficulty: the multifunctional di-, the meN- prefixation system, the suffixes -kan and -i, and the circumfixes me-kan and me-i. Using qualitative content analysis, outputs were coded against a seven-dimension rubric—accuracy, clarity, completeness, organisation, example quality, readability, and pedagogical usefulness—each operationalised through explicit four-level descriptors. Ratings function as ordinal descriptive summaries of qualitative judgements, not as inferential evidence, and no generalisation beyond the analysed corpus is claimed. Of 28 ratings, 22 fell in the highest band and 6 in the second: accuracy, clarity, organisation, and example quality were uniformly strong, whereas readability was weakest, followed by pedagogical usefulness. The recurring shortcoming was not linguistic error but pedagogical omission: explanations did not anticipate the affixation misconceptions documented among BIPA learners, and treated related affixes as unconnected topics. On the evidence of these four explanations, ChatGPT is better positioned to supplement structured BIPA instruction than to substitute for it.

Keywords: *BIPA; ChatGPT; Grammatical Explanation; Pedagogical Quality; Qualitative Content Analysis*

INTRODUCTION

The rapid development of artificial intelligence, and of large language models (LLMs) in particular, has altered the conditions under which foreign languages are learned. ChatGPT no longer functions merely as a retrieval system; it operates as an interactive tutor that generates grammatical explanations, supplies contextual examples, and responds to language questions in real time (Kasneci et al., 2023). Because that capability is available continuously and at no cost to the learner, it has been absorbed into language study far faster than research has been able to evaluate it.

The empirical literature on ChatGPT in second and foreign language learning has grown quickly but unevenly. Systematic reviews converge on the same structural imbalance: Lo et al. (2024), reviewing seventy empirical studies in ESL and EFL settings, report a pronounced skew toward writing; Fang and Han (2025), synthesising seventy-one studies, find that learner and teacher perception studies dominate while evidence of instructional efficacy remains scarce; and Koç and Savaş (2025), in a meta-synthesis of fifty-seven studies published between 2010 and 2024, reach a comparable conclusion for AI chatbots more broadly (Supriyono et al. 2025). Where outcomes have been measured, the findings are encouraging but bounded. Karataş et al. (2024) documented gains in writing, grammar, and vocabulary among Turkish university learners alongside explicit concern about over-reliance; Zare et al. (2025) found increased task motivation that partially decayed at delayed post-test; Yan (2024) and Guo and Wang (2024) examined how learners process ChatGPT-generated feedback in writing; and Fitria (2023) reviewed the application of ChatGPT to English essay writing. What these studies share is an object of enquiry: the learner's response to the tool. The explanation itself is rarely treated as an artefact whose quality can be judged.

That omission matters most where instruction is thinnest. Bahasa Indonesia untuk Penutur Asing (BIPA), Indonesian taught as a foreign language, is not a single classroom experience. Learners enter BIPA programmes from widely differing first-language backgrounds, at proficiency levels benchmarked against Common European Framework of Reference (CEFR)-referenced descriptors from A1 to C2 (Sudaryanto & Widodo, 2020), and through institutional arrangements ranging from short intensive courses to semester-length university programmes (Ardiyanti, 2024). The field itself has expanded rapidly, as a bibliometric survey of the BIPA literature confirms (Syafruddin et al., 2025). Learning also takes place outside any programme: learners consult AI tools during independent study, between class sessions, or before arriving in Indonesia at all. The extent of this has not been quantified for BIPA specifically, but reviews of second and foreign language learning document rapid and largely unsupervised uptake of such tools among learners generally (Fang & Han, 2025; Koç & Savaş, 2025; Lo et al., 2024). In those unsupervised moments the quality of an AI explanation is not a convenience but the only correction available.

The grammatical territory in which this matters is well mapped. Affixation is identified across the BIPA error-analysis literature as the primary source of difficulty for foreign learners of Indonesian (Shofia & Suyitno, 2020). Najiba et al. (2023) documented substantial affixation errors in the narrative writing of BIPA students, particularly in meN- verb formation and in the confusion of di- as a passive verbal prefix with di as a preposition. Kusmiatun et al. (2025) identified seven distinct error patterns in Thai BIPA learners' use of the prefix /me-/, ranging from prefix omission to improper sound elision and morpheme substitution. Fiiarum and Susanto (2023) found morphological errors persisting even among advanced-level BIPA learners, and Mukhibun and Saddhono (2023) traced the allomorphic complexity of Indonesian confixation in learner texts. The particular concepts recur across these studies: the particle di-, which functions as a passive verbal prefix (*dibaca*, 'is read'), a locative preposition (*di rumah*, 'at home'), and an adverbial phrase marker (*di sini*, 'here'); and the meN- prefix, whose five allomorphs (me-, men-, mem-, meng-, meny-) are governed by phonological assimilation and elision rules that generate the elision and substitution errors catalogued in the studies cited above.

Research on technology in BIPA has begun to address this environment, but from a distance. Nasrullah et al. (2024) reviewed the contribution of e-learning platforms, language applications, and immersive technologies to BIPA accessibility, and Rahmatika et al. (2025) surveyed BIPA teachers' attitudes toward digital game-based language learning. Both are concerned with delivery infrastructure and practitioner disposition rather than with what an AI system actually says when a learner asks it a question about Indonesian morphology. To date, the AI-and-BIPA literature consists largely of conceptual argument and literature review; the output itself has not been placed under evaluation.

Outside BIPA, by contrast, rubric-based evaluation of LLM output is an established practice, and its findings are strikingly consistent. In medicine, Goodman et al. (2023) had thirty-three physicians rate 284 ChatGPT answers for accuracy and completeness and found a high median accuracy alongside a minority of substantially incorrect responses, while Wei et al. (2024) synthesised how primary studies have operationalised medical response quality across accuracy, completeness, and readability. In science education, Choi (2025) reported that 94.2% of ChatGPT's answers to elementary science questions were

scientifically valid but only 12.8% were curriculum-appropriate, the shortfall arising from overly advanced terminology; Küchemann et al. (2024) similarly found ChatGPT-generated physics assessment items accurate yet poorly aligned to student misconceptions. In statistics education, Shukla et al. (2025) had five biostatistics experts rate six LLMs on a five-point scale across five explanatory dimensions, including clarity and pedagogical value. In language education, Steiss et al. (2024) found trained raters scoring human feedback on student writing above ChatGPT's overall, with the largest gaps on prioritisation rather than on accuracy; Saricaoglu and Bilki (2025) reported ChatGPT-4 feedback on argumentative essays to be highly accurate but variable in specificity; and Yang and Chen (2025) showed that model version and prompt language materially affect the quality of automated corrective feedback. Alqahtani and Abumalik (2025) come closest to the present concern, analysing ChatGPT responses to grammatical queries and reporting accuracy ranging from 0% to 80% together with instructor concern about learner over-reliance, though their focus is English grammar.

A single pattern runs through this body of work: linguistic or factual correctness is achieved far more reliably than instructional fitness. That distinction has a long pedigree in grammar pedagogy. Thornbury (1999) held that effective grammatical explanation requires clarity of rule formulation, accuracy of terminology, quality of examples, and relevance to learner need. Ellis (2006) argued from a second language acquisition perspective that an explanation must not only be accurate but must facilitate cognitive processing, and that explanations anticipating common misconceptions support acquisition more effectively than those that do not—a position he subsequently refined in reconsidering the place of form-focused instruction (Ellis, 2016). Celce-Murcia (1991) insisted that grammar be presented in contexts of authentic use rather than in isolation. More recent work substantiates and complicates these claims: meta-analyses report substantial overall effects for form-focused instruction with only minor differences between explicit and implicit variants (Kang et al., 2019; Li & Sun, 2024); Kissling (2021) showed that learners systematically misinterpret pedagogical rules in ways rule-writers do not anticipate; Ryu and Choi (2024) provided eye-movement evidence that targeted metalinguistic explanation aids grammatical uptake without impairing comprehension; and Pawlak (2021) and Hwang (2023) document how far grammar-teaching practice still stands from the research that should inform it. Comprehensibility itself is now well theorised through cognitive load research, which specifies how unexplained terminology, absent signalling between core and peripheral content, and undifferentiated information volume impose avoidable processing demands (Castro-Alonso et al., 2021; Sweller et al., 2019).

Setting these strands side by side makes the gap precise. Studies of LLM output quality have built and applied evaluation rubrics, but almost entirely in medicine, science, and statistics, and where they enter language education they evaluate feedback on learner writing rather than expository grammatical explanation. Studies of AI in second language learning have measured perceptions, motivation, and writing outcomes, but have seldom treated the explanation as the unit of analysis. The small number of studies that do evaluate AI-generated grammatical explanation address English. And research on AI in BIPA has remained conceptual, addressing infrastructure and teacher attitudes rather than output. No study, to the authors' knowledge, has evaluated the quality of AI-generated explanations of Indonesian grammar against the specific errors that BIPA learners are documented to make. Two consequences follow: the field has no instrument calibrated to BIPA, and it has no evidence about whether the accuracy–pedagogy dissociation observed elsewhere also holds for Indonesian morphology.

The present study addresses that gap. It evaluates ChatGPT explanations of four grammatical concepts selected from the BIPA error-analysis literature, using a seven-dimension rubric whose dimensions are derived from grammar pedagogy and cross-checked against established LLM-evaluation practice, and whose score levels are operationalised through explicit descriptors. Two research questions guide the study: (1) How is the quality of ChatGPT-generated explanations of these BIPA grammatical concepts characterised across the seven evaluative dimensions? (2) What are the strengths and limitations of ChatGPT in explaining these concepts for BIPA learning purposes? By anchoring evaluation in documented BIPA learner difficulty rather than in general Indonesian-language pedagogy, the study takes a first step toward a systematic, BIPA-grounded framework for assessing LLM-generated Indonesian-language content.

METHODS

This study employed qualitative content analysis (Krippendorff, 2018; Mayring, 2014, 2021) in its deductive, category-application variant, in which a category system derived from theory is applied systematically to a defined corpus and each assignment is documented against explicit coding rules. The design is qualitative in its object, its unit of analysis, and its warrant: the object is the meaning and instructional function of a text, the unit is a complete explanatory output, and the warrant rests on the traceability of each judgement to textual evidence rather than on sampling. Numerical ratings are used

within this design, and their status requires statement. Each rating is an ordinal code standing for a qualitative judgement made against a written descriptor; the means and percentage distributions reported below are descriptive summaries of the coded corpus and nothing more. They are included because they make the profile of strengths and weaknesses comparable across dimensions and concepts, and because they expose the pattern of judgements to inspection. They are not inferential statistics: no hypothesis is tested, no sampling frame exists, no confidence interval or significance claim is made, and no conclusion is generalised beyond the four outputs analysed. Differences between numerically close values carry no evidential weight in themselves and are interpreted only where the underlying textual justifications support a substantive distinction.

The corpus consists of four text outputs generated by ChatGPT (GPT-4o) in response to four standardised prompts. The four grammatical concepts were not selected to represent Indonesian grammar as a whole but were drawn from the affixation errors that BIPA learners are documented to make (Fiiarum & Susanto, 2023; Kusmiatun et al., 2025; Najiba et al., 2023; Shofia & Suyitno, 2020). They comprise: the distinction between di- as a passive verbal prefix, as a preposition, and as an adverbial marker; the meN- prefixation system and its allomorphic variation; the suffixes -kan and -i and their functional differences; and the circumfixes me-kan and me-i and their contrasting semantic roles. All four prompts followed a single uniform frame—a request to explain the concept in Indonesian, including definitions, rules, and example sentences—and are reproduced in full in Appendix A so that the outputs can be traced to their elicitation. Each prompt was submitted in a separate conversation session to prevent contextual carryover between concepts, with no custom instructions, no system prompt, and no follow-up turns; default model settings were used throughout, and all outputs were collected in June 2026.

The prompts were deliberately level-neutral: none specified a proficiency level, a first-language background, or an instructional context. This is a design property rather than an oversight: it is intended to approximate the conditions under which a learner consulting the system independently would query it, an assumption of the design rather than an observed pattern of use. It nonetheless raises the question of who such an explanation is calibrated for. Table 1 therefore sets out an analytical alignment between the four concepts and CEFR-referenced BIPA proficiency bands, derived from the morphological complexity of each concept and from published analyses of how the concepts are sequenced in standard BIPA materials (Ardiyanti, 2024; Kurniasih, 2021; Sudaryanto & Widodo, 2020). The alignment is analytical and is offered as an interpretive frame for the findings; it is not an empirically validated placement, and no learners at these levels were tested.

Table 1. Analytical Alignment of the Four Grammatical Concepts with CEFR-Referenced BIPA Proficiency Bands

Grammatical concept	Morphological complexity	Proposed CEFR-referenced BIPA band	Basis of the alignment
1. Multifunctional *di-* (passive prefix / preposition / adverbial marker)	One form, three functions; distinguished orthographically by attached versus separate writing	A1–A2, extending to B1 for co-occurrence cases	Locative expression and basic passive forms belong to beginner content in standard BIPA materials; clauses containing both functions require later syntactic control
2. *meN-* prefixation system	One prefix, five allomorphs, phonologically conditioned elision, loanword exceptions	A2–B1	Active verb formation is core elementary content; elision rules and loanword exceptions extend into the intermediate band
3. Suffixes *-kan* and *-i*	Two suffixes with overlapping and partly contrastive argument-structure functions	B1–B2	Requires prior control of verb valency and of the distinction between direct and indirect objects
4. Circumfixes *me-kan* and *me-i*	Simultaneous prefixation and suffixation; contrastive semantics; presupposes concepts 2 and 3	B1–B2	Presupposes control of the *meN-* system and of the suffixes treated separately, and is therefore sequenced after both

The seven-dimension rubric is the study's central instrument, and it was constructed in three steps. First, candidate dimensions were derived from the criteria that grammar pedagogy specifies for effective explanation. Thornbury's (1999) account of rule formulation, terminological accuracy, exemplification, and structural progression yielded clarity, accuracy, example quality, and organisation. Ellis's (2006, 2016) emphasis on cognitive processing and on the anticipation of misconception yielded pedagogical usefulness,

and his concern with the adequacy of the account given yielded completeness. Celce-Murcia's (1991) insistence on contextualised authentic use was folded into example quality as a criterion of contextual variety rather than treated as a separate dimension. Second, these candidates were cross-checked against the dimensions recurrently used in rubric-based LLM evaluation, so that the instrument would be commensurable with existing work: accuracy and completeness recur across expert-rated LLM studies (Goodman et al., 2023; Wei et al., 2024); accessibility or readability recurs as a distinct construct because factually correct output is repeatedly found to be pitched above its intended reader (Choi, 2025); and pedagogical value appears explicitly among the expert-rated dimensions used by Shukla et al. (2025). Third, only dimensions attested in both literatures were retained, and overlapping candidates were collapsed, yielding the seven dimensions set out in Table 2.

Two features of the instrument require explicit statement. First, the four-point scale was constructed by the authors for this study. It was informed by the general practice of ordinal expert rating in LLM evaluation, but it is not an adaptation of any single published instrument, and it does not reproduce the scale length or the dimension set of any of the studies cited above. Second, readability was deliberately not operationalised through automated readability indices, since existing Indonesian readability formulas are calibrated on first-language readers and are not validated for foreign learners. It was instead defined through the constructs of cognitive load theory (Castro-Alonso et al., 2021; Sweller et al., 2019): unglossed technical terminology, absence of signalling between core and supplementary content, and undifferentiated information volume are treated as sources of avoidable processing demand, and the descriptors in Table 3 refer to these directly.

Table 2. Derivation and Operational Definition of the Seven Evaluative Dimensions

No	Dimension	Definition as applied in this study	Derivation
1	Accuracy	Conformity of rule statements, form–function mappings, and examples to standard Indonesian grammar, including the marking of attested exceptions.	Terminological accuracy in Thornbury (1999) and Ellis (2006); recurrent as a rated dimension in expert evaluation of LLM output (Goodman et al., 2023; Wei et al., 2024)
2	Clarity	Unambiguity of each rule formulation and explicitness of the criterion by which competing forms are distinguished.	Clarity of rule formulation in Thornbury (1999); clarity and accessibility among the expert-rated dimensions in Shukla et al. (2025)
3	Completeness	Coverage of the functions and forms of the target concept, including the boundary cases in which they interact.	Adequacy of the account given in Ellis (2006); completeness as a standard rated dimension in LLM evaluation (Goodman et al., 2023; Wei et al., 2024)
4	Organisation	Consistency and signposting of the expository progression across sub-parts, and presence of an integrative synthesis.	General-to-specific progression in Thornbury (1999)
5	Example quality	Grammaticality, contextual plausibility, range, and contrastive pairing of the examples supplied.	Exemplification criteria in Thornbury (1999); contextualised authentic use in Celce-Murcia (1991)
6	Readability	Processing demand imposed on a reader without linguistic training, assessed through glossing of metalanguage, signalling between core and supplementary content, and proportionality of information volume.	Accessibility treated as a construct distinct from accuracy in LLM evaluation (Choi, 2025); operationalised through cognitive load theory rather than through readability indices (Castro-Alonso et al., 2021; Sweller et al., 2019)
7	Pedagogical usefulness	Provision beyond description for correct application: anticipation of documented learner error, and connection of the concept to directly related concepts.	Misconception anticipation in Ellis (2006, 2016); system-oriented grammar teaching in Celce-Murcia (1991); pedagogical value among the expert-rated dimensions in Shukla et al. (2025)

Because the boundary between adjacent score levels is where a rubric of this kind is most vulnerable, each of the seven dimensions was operationalised through written descriptors for all four levels, set out in Table 3. The decision rule separating the two levels that occur in this corpus is stated positively: a rating of 4 requires that no weakness against the descriptor can be identified in the output, whereas a rating of 3 requires that at least one specific weakness be named, whether as a locatable feature of the output or as a specifiable absence from it, and that the weakness leave the core function of the dimension intact. Under this rule a rating cannot be assigned without nameable textual grounds, and a reader who disagrees with a rating can locate the evidence on which it turned.

Table 3. Operational Descriptors for Each Score Level of Each Dimension

Dimension	4 — Very Good	3 — Good	2 — Fair	1 — Poor
Accuracy	All rule statements, form–function mappings, and examples conform to standard Indonesian grammar; no misleading generalisation is made; attested exceptions are noted where relevant.	All core rule statements are correct, but at least one exception, restriction, or peripheral case is omitted or stated too broadly, without generating a false rule.	At least one rule statement or example is incorrect, or would license an ungrammatical form, although the overall account remains recognisable.	Central rule statements are wrong or mutually inconsistent; following the explanation would systematically produce ungrammatical forms.
Clarity	Each rule is stated in a single unambiguous formulation, and the criterion distinguishing competing forms is made explicit and is directly testable by the learner.	Rules are intelligible, but at least one formulation is ambiguous or obliges the reader to infer the criterion of distinction.	Formulations are frequently ambiguous; the reader cannot reliably determine which form applies in a given case.	The text cannot be resolved into applicable rule statements.
Completeness	All functions and forms of the concept attested in the BIPA reference literature are covered, together with the boundary cases in which those functions co-occur within a single construction. Relations to separately treated concepts fall under pedagogical usefulness.	All principal functions and forms are covered, but boundary or co-occurrence cases are not addressed.	At least one principal function or form of the concept is absent.	Coverage is fragmentary; the concept is not recognisably treated.
Organisation	A consistent progression from definition to rule to exemplification to synthesis is applied uniformly to every sub-part and closes with an integrative summary.	The progression is broadly coherent but is applied inconsistently across sub-parts, or no closing synthesis is provided.	A structure is discernible but does not support cumulative comprehension; related material is dispersed.	No discernible structure.
Example quality	Examples are grammatical, contextualised in plausible use, varied across the range of the rule, and paired contrastively wherever the concept is defined by opposition.	Examples are grammatical and relevant but limited in range, or insufficiently contrastive where the concept is defined by opposition.	Examples are present but do not illustrate the stated rule, or are confined to a single unrepresentative type.	Examples are absent, ungrammatical, or contradict the rule.
Readability	Metalanguage is either avoided or glossed at first use; core content is distinguished from supplementary content by explicit signalling; length is	The text imposes avoidable processing demand through unglossed technical terminology, absence of signalling between core and supplementary content, or	Such demands are severe enough that a learner without linguistic training is unlikely to extract the rule unaided.	The text is not accessible to the reader for whom it is intended.

	proportionate to the concept.	undifferentiated length, but the rule remains recoverable by an attentive reader without assistance.		
Pedagogical usefulness	The explanation goes beyond description to support application—by naming a documented learner error and showing why it is wrong, by linking the concept explicitly to a directly related concept, or by supplying a stated decision procedure—and leaves unaddressed no documented misconception, and no relation to a neighbouring concept, that its own content directly raises.	The explanation supports correct comprehension, and may supply some means of application, but it leaves unaddressed a documented misconception, or a relation to a neighbouring concept, that its own content directly raises.	The explanation conveys information without any provision for application; the learner is given no means of self-monitoring.	The explanation is likely to induce or reinforce error.

Analysis proceeded in four stages. Each output was first read in full to establish an overall reading. Each was then coded against the seven dimensions in a fixed order, with every rating recorded against the specific feature of the output, or the specific absence from it, on which the rating turned. Ratings were then interpreted across concepts to identify patterns of strength and weakness, with particular attention to whether an identified weakness corresponded to a misconception already documented in the BIPA error-analysis literature. Finally, the patterns were examined for their sources: rather than stopping at the observation that certain dimensions scored lower, the analysis asked what property of the outputs, of the interaction format, or of the model's task made those dimensions systematically weaker, and what follows for BIPA practice.

The limits of this procedure should be stated plainly rather than mitigated. Coding was carried out by one rater. No second coder participated, and consequently no inter-rater agreement coefficient is reported; the study makes no claim to inter-subjective reliability, and its ratings should be read as one systematically documented reading rather than as a consensus judgement. What the design offers instead is auditability. The category system was fixed in advance of coding and applied in identical order to every output; each rating was recorded in the coding sheet against the passage of the output on which it turned, and representative extracts are quoted in the discussion below; the descriptors against which every rating was made are published in full in Table 3; and the complete rating matrix with its justifications is reported in Table 4 rather than summarised, so that each judgement can be checked against the descriptor that governed it. This satisfies the transparency requirement that Krippendorff (2018) treats as a precondition of reliability without satisfying the reliability requirement itself, and independent recoding by a second rater is accordingly identified below as the first priority for replication.

RESULT AND DISCUSSION

The complete rating matrix, together with the identified weakness underlying every rating below the highest level, is presented in Table 4, and the resulting profile of the four concepts across the seven dimensions in Figure 1. Of the twenty-eight ratings, twenty-two fell in the highest band and six in the second; none fell in the lower two bands. The discussion that follows does not restate those values but examines four patterns that emerge from them, quotes the passages of the outputs on which the ratings turned, and considers what produces each pattern.

Table 4. Rating Profile Across Four Grammatical Concepts and Seven Evaluative Dimensions

Dimension	1. di-	2. meN-	3. -kan / -i	4. me-kan / me-i	Identified weakness underlying any rating below 4
Accuracy	4	4	4	4	No weakness identified in any output. Attested exceptions, including the loanword exceptions to *meN-* elision (*mengkritik*, *memprotes*, *mentransfer*), were correctly marked.
Clarity	4	4	4	4	No weakness identified in any output. Within each concept, the criteria distinguishing competing forms were made explicit through diagnostic questions, negative evidence, minimal pairs, and sentence-component analysis.
Completeness	3	4	4	4	*di-*: clauses in which two *di-* functions co-occur (for example *Barang itu diletakkan di atas meja*) are not addressed.
Organisation	4	4	4	4	No weakness identified in any output.
Example quality	4	4	4	4	No weakness identified in any output.
Readability	4	3	3	3	*meN-*: undifferentiated length, with no signalling between core and supplementary content. *-kan/-i*: 'causative', 'benefactive', 'locative', 'applicative', and 'intensive' used without gloss. *me-kan/me-i*: undifferentiated length, with no signalling between core and supplementary content.
Pedagogical usefulness	3	4	4	3	*di-*: the separation error *di makan* for *dimakan*, documented by Najiba et al. (2023), is not named. *me-kan/me-i*: the relation between the suffix *-kan* and the circumfix *me-kan* is not addressed. The two ratings of 4 rest on negative evidence naming documented elision errors plus an active–passive cross-link (*meN-*), and on contrastive minimal pairs plus a stated decision procedure (*-kan/-i*).
Concept mean	**3.71**	**3.86**	**3.86**	**3.71**	Overall mean 3.79. Of 28 ratings, 22 (78.6%) at level 4 and 6 (21.4%) at level 3; none at level 2 or 1. Dimension means are given in Figure 1.

The first and most consequential pattern is a systematic dissociation between linguistic accuracy and pedagogical utility. Four dimensions—accuracy, clarity, organisation, and example quality—were rated at the highest level for every concept without exception. Two dimensions, readability and pedagogical usefulness, never reached that level uniformly, and completeness fell short once. This division is not arbitrary. The four dimensions that reached ceiling can each be satisfied from the language system alone: whether a rule statement is correct, whether it is unambiguously formulated, whether the exposition proceeds coherently, and whether the examples instantiate the rule are all properties determinable from knowledge of Indonesian. The accuracy of the *meN-* account is representative: ChatGPT states that 'the prefix men- is used with base words beginning with c, d, j, t, z' and that 'when the base word begins with t, the letter t is usually elided', and it goes on to mark the loanword exceptions *mengkritik*, *memprotes*, and *mentransfer*, which are precisely the cases where the elision rule does not apply. The two dimensions that did not reach ceiling cannot be satisfied from the language system alone, because both require inference about a reader—what this learner already knows, what terminology will be opaque to them, and what they are likely to get wrong. The asymmetry is plausibly related to the evidential base available to a general-purpose model: text describing Indonesian grammar is abundant, whereas documentation of what foreign learners of Indonesian do with it is comparatively scarce and largely confined to the specialist literature reviewed above. That is an inference about the likely source of the pattern rather than a claim about the composition of any model's training data, which is not public. Whatever its source, the pattern is not peculiar to Indonesian: Choi (2025) found the same split in elementary science, where 94.2% of responses were scientifically valid but only 12.8% were curriculum-appropriate, and Steiss et al. (2024) found human feedback outperforming ChatGPT on prioritisation rather than on correctness. What the present analysis adds is that the split reappears along the same fault line in a low-resource pedagogical context.

The second pattern concerns readability, the weakest dimension in this corpus, and it resolves into two distinct causes that a single rating obscures. In the explanation of *-kan* and *-i*, the difficulty was lexical. The output summarises the contrast as 'general meaning: causative and benefactive (-kan); locative, applicative, or intensive (-i)', deploying five technical labels without gloss in a text otherwise addressed to a general reader. In the explanations of *meN-* and of *me-kan/me-i*, the difficulty was structural: the output was

long and undifferentiated, with no marker distinguishing what a learner must grasp from what is supplementary. These are not variants of one problem. The first is a register mismatch, correctable by the learner through a follow-up prompt or by an instructor in a single sentence. The second is not correctable in that way, because it arises from the absence of any criterion for omission: to signal that one part of an explanation matters more than another, a system must know who is reading, and a level-neutral prompt supplies no such information. Cognitive load theory distinguishes these mechanisms precisely, the first raising extraneous load through unfamiliar symbols and the second violating the signalling principle by which learners are directed to essential material (Castro-Alonso et al., 2021; Sweller et al., 2019). Three of the four outputs were rated at ceiling on completeness while falling below it on readability, and in each case these are the same textual property assessed against two criteria. Comprehensiveness and accessibility trade off, and in these outputs the trade-off is resolved toward comprehensiveness—the appropriate resolution for a reference text and the wrong one for a first encounter with a rule.

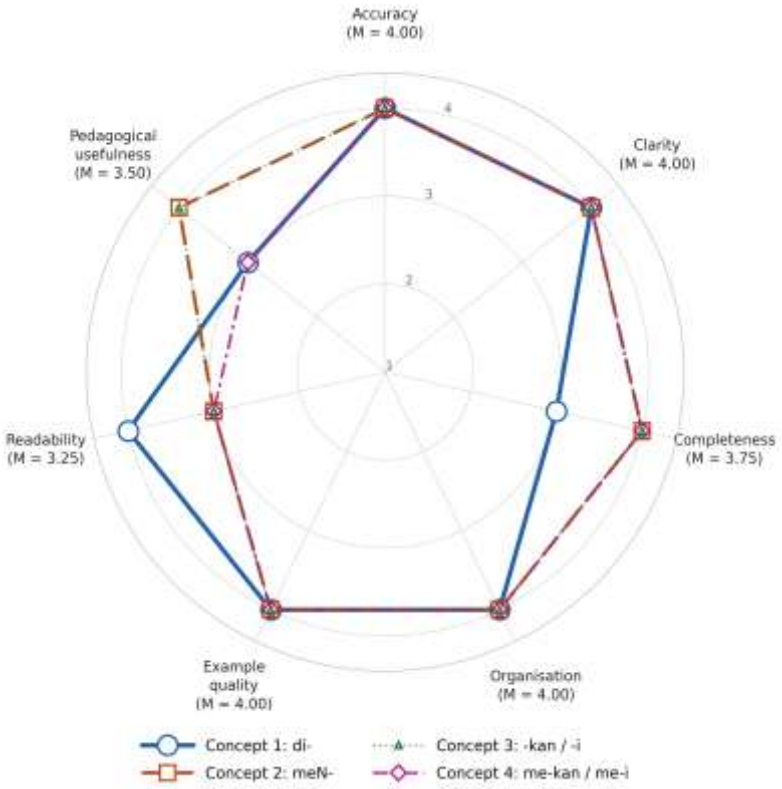

Figure 1. Rating Profiles of the Four Grammatical Concepts Across the Seven Dimensions, with Dimension Means

Note. Ratings are on the 1–4 scale defined in Table 3. Dimension means (M) are shown beside each axis label; per-concept means are given in Table 4. Series are distinguished by colour, line style, and marker shape so that the figure remains legible in greyscale; markers are nested by size so that coincident traces remain visible. Concepts 2 and 3 have identical rating profiles and coincide at every dimension.

The third pattern is that the failures of pedagogical usefulness are failures of anticipation rather than of content, and that they are predictable in advance. Both lower ratings on this dimension arose in this way. In the *di-* explanation, the three functions were distinguished accurately and a serviceable diagnostic was supplied—'is there a verb after di? If yes, it is usually written together… is there a place word after di? If yes, it is written separately'—yet the output nowhere mentions that learners write *di makan* for *dimakan*, the separation error documented in BIPA learner writing by Najiba et al. (2023) and the one the orthographic contrast under discussion most readily generates. In the *me-kan*/*me-i* explanation, the semantic contrast was analysed carefully through sentence-component analysis ('primary object: buku; the book is made to enter the bag; focus is on the object being moved; therefore *memasukkan* is used'), but the question a learner arrives with after reading about the suffixes—what distinguishes *-kan* from *me-kan*—was neither posed nor answered. The mechanism is the same in both cases: the system answers the question asked, and by definition a learner able to formulate the question that would expose their misconception no longer holds it. An explanation that describes the correct form without naming the competing incorrect form leaves the learner's existing hypothesis undisturbed, which is the condition under which Kissling (2021) found learners systematically misreading pedagogical rules, and the converse of the targeted metalinguistic intervention that Ryu and Choi

(2024) found effective. Ellis (2006, 2016) had already identified misconception anticipation as constitutive of effective explanation rather than as an optional refinement; what these outputs show is that it is the one requirement an accurate, well-organised, well-exemplified explanation can fail entirely while appearing complete.

The fourth pattern is that the two outputs rated at ceiling on pedagogical usefulness are the two that went beyond description to equip the learner for application, and that they did so in different ways. The *meN-* output supplied negative evidence, presenting incorrect forms alongside correct ones—'me- + tulis → menulis, not mentulis; me- + kirim → mengirim, not mengkirim; me- + simpan → menyimpan, not mensimpan'—which names the elision and substitution errors that fall among the patterns Kusmiatun et al. (2025) document among BIPA learners, and it closed with an active–passive comparison that tied *meN-* to *di-*, presenting the two affixes as complements within one system. The *-kan*/*-i* output supplied contrastive minimal pairs ('memasukkan: to put something in – saya memasukkan buku ke tas; memasuki: to enter a place – mahasiswa memasuki ruang kelas') together with a set of diagnostic questions constituting a stated decision procedure. That these pairs are themselves built on *meN-*-prefixed forms is worth noting, since it is part of why the boundary between the suffix and the circumfix is left indistinct across the two outputs. Neither of the two lower-rated outputs did either. With four outputs there is no basis for a covariation claim, and none is made here; what can be said is that in this corpus the property distinguishing the higher from the lower ratings on this dimension was not accuracy, which was invariant, but whether the explanation provided something beyond a correct description. The cross-concept integration seen in the *meN-* output is worth separating out, because it implicates the interaction format as much as the model. Each prompt was submitted in an isolated session, a decision taken to protect the outputs from contextual carryover, and one that also approximates the concept-by-concept manner in which a learner consulting the system independently is likely to proceed. The session boundary that secures methodological independence is the same boundary that prevents a cumulative account from being built. Celce-Murcia (1991) argued that grammar should be taught as a system rather than as isolated items; the affordance that makes such a tool convenient, that one may ask a single question and receive a single self-contained answer, works structurally against that principle. The integration in the *meN-* output was spontaneous, not systematic, and nothing in the interaction guarantees its recurrence.

Read against the proficiency alignment in Table 1, these patterns are not evenly distributed. The two concepts aligned with B1–B2, the suffixes and the circumfixes, are the two whose explanations were weakened by unglossed metalanguage and by undifferentiated length respectively, and they are also the concepts that presuppose control of the material treated at lower bands. Learners at those levels are not assumed by BIPA curricula to command Indonesian linguistic metalanguage (Ardiyanti, 2024; Sudaryanto & Widodo, 2020). Meanwhile the concept aligned with the lowest band, *di-*, produced the only explanation rated at ceiling for readability and the one that most conspicuously omitted a documented beginner error. A level-neutral output is calibrated to no learner in particular, and the mismatch falls differently at different levels. This reading is analytical rather than empirical, and its status should not be overstated: it follows from the alignment proposed in Table 1 and from the textual evidence in Table 4, not from any observation of learners.

Taken together, these findings carry a specific implication for BIPA practice rather than for Indonesian-language pedagogy in general. Because the four explanations analysed were consistently accurate, clearly formulated, well organised, and well exemplified, such output is a defensible first point of reference for a BIPA learner needing a quick and correct account between class sessions. Because the weaknesses fall on the two dimensions that determine whether an explanation forestalls an error rather than merely describing the correct form afterwards, it cannot be left to operate unsupervised as a substitute for instruction. For BIPA instructors, this means treating AI output as a first draft requiring a second pass, one that names what the output does not: the co-occurrence of *di-* functions within a single clause, the separation error *di makan* for *dimakan*, and the relation between the suffix *-kan* and the circumfix *me-kan*. For BIPA materials developers, it identifies a concrete and bounded design task: short supplementary notes for each affix family that name the misconception the AI leaves unaddressed, written to be consulted alongside an AI explanation rather than in place of it. For researchers, the rubric in Tables 2 and 3 is offered as an instrument that can be applied to other models and other concepts, and whose descriptors can be contested and revised.

Several limitations bound these claims. The study evaluated four concepts, one model, and one output per concept, elicited through a single prompt formulation; a different formulation, a different model version, or a repeated session might yield different outputs, and the study provides no evidence about that variability. Coding was performed by a single rater, so the ratings are auditable but not inter-subjectively verified. The dimensions of the rubric are not fully independent: a single textual absence may bear on more than one of them, and the descriptors in Table 3 assign each such absence to the dimension whose defining criterion it violates most directly, a decision that is stated rather than validated. The proficiency alignment in Table 1 is analytical rather than empirical. Most importantly, the study evaluated explanations, not learning: it

establishes properties of a text and says nothing directly about what BIPA learners understand, retain, or misapply after reading it. Replication with a second independent coder, extension across models, prompt formulations, repeated sessions, and level-specified prompts, and learner-facing designs using think-aloud protocols or comprehension tasks are all required before the present findings can be treated as more than a first characterisation.

CONCLUSION

This study asked how well ChatGPT explains the grammatical concepts that BIPA error-analysis research identifies as persistently difficult, and where its strengths and limitations lie for BIPA purposes. Across the four concepts examined, the evaluation found explanations that were linguistically accurate, clearly formulated, systematically organised, and effectively exemplified, but consistently weaker on the two dimensions that require modelling the learner rather than the language: readability and pedagogical usefulness. The shortcoming was one of omission rather than error. The explanations described correct forms without naming the incorrect forms that BIPA learners are documented to produce, and treated each affix as a self-contained topic rather than as an element of an integrated morphological system, with the result that an explanation could be complete and correct while leaving a learner's misconception entirely undisturbed. The contribution of the study is therefore twofold. It supplies an operationalised, BIPA-anchored instrument for judging AI-generated Indonesian grammatical explanation, with published descriptors that can be applied, contested, and revised by others; and it shows that in this context the criterion separating an adequate explanation from a useful one is the anticipation of documented learner error rather than linguistic correctness. These claims hold for one model, one prompt formulation, and one rater's documented reading of four outputs, and they describe the explanations analysed rather than the behaviour of ChatGPT in general. They also say nothing directly about what BIPA learners actually understand from such explanations, which is a question the rubric developed here is intended to support rather than to answer.

ACKNOWLEDGEMENT

This research received no external funding. The authors declare no conflict of interest and thank the BIPA programme at Universitas Brawijaya for institutional support during this study. All citations and the reference list were managed using Mendeley reference management software.

REFERENCES

Alqahtani, F. A., & Abumalik, A. M. (2025). From Assistance to Over-Reliance: Rethinking Chatgpt's Role in Grammar Learning. *Forum for Linguistic Studies*, *7*(8), 523–539. https://doi.org/10.30564/fls.v7i8.10519

Ardiyanti, W. N. (2024). The Strengths and Weaknesses of The BIPA Curriculum As An Indonesian Teaching Guideline for Foreign Speakers: A Review of The Literature 2018–2023. *Indonesian Journal of Curriculum and Educational Technology Studies, 12*(2), 108–122. https://doi.org/10.15294/ijcets.v12i2.13939

Castro-Alonso, J. C., de Koning, B. B., Fiorella, L., & Paas, F. (2021). Five Strategies for Optimizing Instructional Materials: Instructor- and Learner-Managed Cognitive Load. *Educational Psychology Review, 33*(4), 1379–1407. https://doi.org/10.1007/s10648-021-09606-9

Celce-Murcia, M. (1991). Grammar Pedagogy in Second and Foreign Language Teaching. *TESOL Quarterly, 25*(3), 459–480. https://doi.org/10.2307/3586980

Choi, Y. (2025). Exploring The Scientific Validity of Chatgpt's Responses in Elementary Science for Sustainable Education. *Sustainability, 17*(7), Article 2962. https://doi.org/10.3390/su17072962

Ellis, R. (2006). Current Issues in the Teaching of Grammar: An SLA perspective. *TESOL Quarterly, 40*(1), 83–107. https://doi.org/10.2307/40264512

Ellis, R. (2016). Focus on Form: A Critical Review. *Language Teaching Research, 20*(3), 405–428. https://doi.org/10.1177/1362168816628627

Fang, S., & Han, Z. (2025). On the Nascency of ChatGPT in Foreign Language Teaching and Learning. *Annual Review of Applied Linguistics, 45*, 148–178. https://doi.org/10.1017/S026719052510010X

Fiiarum, F. A. K., & Susanto, G. (2023). Kesalahan Berbahasa Indonesia dalam Tulisan Pemelajar BIPA Tingkat Mahir. *Journal of Language, Literature, and Arts, 3*(4), 557–568. https://doi.org/10.17977/um064v3i42023p557-568

Fitria, T. N. (2023). Artificial Intelligence (AI) technology in OpenAI ChatGPT Application: A Review of ChatGPT in Writing English Essay. *ELT Forum: Journal of English Language Teaching, 12*(1), 44–58. https://doi.org/10.15294/elt.v12i1.64069

Goodman, R. S., Patrinely, J. R., Stone, C. A., Jr., Zimmerman, E., Donald, R. R., Chang, S. S., Berkowitz, S. T., Finn, A. P., Jahangir, E., Scoville, E. A., Reese, T. S., Friedman, D. L., Bastarache, J. A., van der Heijden, Y. F., Wright, J. J., Ye, F., Carter, N., Alexander, M. R., Choe, J. H., … Johnson, D. B. (2023). Accuracy and Reliability of Chatbot Responses to Physician Questions. *JAMA Network Open, 6*(10), Article e2336483. https://doi.org/10.1001/jamanetworkopen.2023.36483

Guo, K., & Wang, D. (2024). To Resist It or to Embrace It? Examining Chatgpt's Potential to Support Teacher Feedback in EFL Writing. *Education and Information Technologies, 29*(7), 8435–8463. https://doi.org/10.1007/s10639-023-12146-0

Hwang, H.-B. (2023). Is Evidence-Based L2 Pedagogy achievable? The Research–Practice Dialogue in Grammar Instruction. *The Modern Language Journal, 107*(3), 734–755. https://doi.org/10.1111/modl.12864

Kang, E. Y., Sok, S., & Han, Z. (2019). Thirty-Five Years of ISLA on Form-Focused Instruction: A Meta-analysis. *Language Teaching Research, 23*(4), 428–453. https://doi.org/10.1177/1362168818776671

Karataş, F., Abedi, F. Y., Ozek Gunyel, F., Karadeniz, D., & Kuzgun, Y. (2024). Incorporating AI in Foreign Language Education: An Investigation into ChatGPT's Effect on Foreign Language Learners. *Education and Information Technologies, 29*(15), 19343–19366. https://doi.org/10.1007/s10639-024-12574-6

Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., … Kasneci, G. (2023). ChatGPT for Good? On Opportunities and Challenges of Large Language Models for Education. *Learning and Individual Differences, 103*. https://doi.org/10.1016/j.lindif.2023.102274

Kissling, E. M. (2021). From rule-based explicit instruction to explicit knowledge: A Pilot Study on How L1 English Speakers Interpret Pedagogical Rules about Spanish Preterite and Imperfect. *Instructed Second Language Acquisition, 5*(1), 40–68. https://doi.org/10.1558/isla.18098

Koç, F. Ş., & Savaş, P. (2025). The Use of Artificially Intelligent Chatbots in English Language Learning: A Systematic Meta-Synthesis Study of Articles Published Between 2010 And 2024. *ReCALL, 37*(1), 4–21. https://doi.org/10.1017/S0958344024000168

Krippendorff, K. (2018). *Content Analysis: An Introduction to Its Methodology* (4th ed.). SAGE.

Küchemann, S., Rau, M., Schmidt, A., & Kuhn, J. (2024). ChatGPT's quality: Reliability and Validity of Concept Inventory Items. *Frontiers in Psychology, 15*, Article 1426209. https://doi.org/10.3389/fpsyg.2024.1426209

Kurniasih, D. (2021). Analisis Bahan Ajar Bahasa Indonesia bagi Penutur Asing (BIPA) Sahabatku Indonesia Tingkat Dasar. *Madah: Jurnal Bahasa dan Sastra, 12*(1), 25–45. https://doi.org/10.31503/madah.v12i1.305

Kusmiatun, A., Fadilla, A. R., & Islahuddin. (2025). Analisis Kesalahan Penggunaan Prefiks /me-/ pada Tugas Menulis Pemelajar BIPA Thailand. *Dialektika: Jurnal Bahasa, Sastra, dan Pendidikan Bahasa dan Sastra Indonesia, 12*(1), 57–66. https://doi.org/10.15408/dialektika.v12i1.40478

Li, F., & Sun, Y. (2024). Effects of Different Forms of Explicit Instruction on L2 Development: A Meta-Analysis. *Foreign Language Annals, 57*(1), 229–255. https://doi.org/10.1111/flan.12726

Lo, C. K., Yu, P. L. H., Xu, S., Ng, D. T. K., & Jong, M. S.-Y. (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), Article 50. https://doi.org/10.1186/s40561-024-00342-5

Mayring, P. (2014). *Qualitative Content Analysis: Theoretical Foundation, Basic Procedures and Software Solution*. Klagenfurt. https://nbn-resolving.org/urn:nbn:de:0168-ssoar-395173

Mayring, P. (2021). *Qualitative Content Analysis: A Step-by-Step Guide*. SAGE.

Mukhibun, A., & Saddhono, K. (2023). The Indonesian Confix /pe-+-an/ in Thammasat University Student Descriptive Text: Allomorph, Usage, and Grammatical Meaning. *Aksis: Jurnal Pendidikan Bahasa dan Sastra Indonesia, 7*(2), 117–127. https://doi.org/10.21009/AKSIS.070202

Najiba, N., Wurianto, A. B., & Isnaini, M. (2023). Bentuk Afiksasi pada Teks Narasi Mahasiswa BIPA: Kajian Terhadap Hasil Tulis Mahasiswa BIPA Asal Afghanistan Angkatan Tahun 2021 di Universitas Muhammadiyah Malang. *GHANCARAN: Jurnal Pendidikan Bahasa dan Sastra Indonesia, 5*(1), 1–14. https://doi.org/10.19105/ghancaran.v5i1.7065

Nasrullah, R., Prayogi, A., & Jannah, R. (2024). Digital Transformation in BIPA Learning: Increasing Accessibility and Effectiveness Through Technology. *LITERATUR: Jurnal Bahasa dan Sastra, 6*(2), 67–96. https://doi.org/10.47766/literatur.v6i2.3373

Pawlak, M. (2021). Teaching Foreign Language Grammar: New Solutions, Old Problems. *Foreign Language Annals, 54(*4), 881–896. https://doi.org/10.1111/flan.12563

Rahmatika, L., Sariasih, Y., Islam, M. M., Solihati, T. A., & Haryanto, S. (2025). BIPA Teachers' Perspectives on Digital Game-Based Language Learning (DGBLL): Attitudes, Benefits and Challenges in Teaching Indonesian as a Foreign Language. *Indonesian Journal of Applied Linguistics, 15*(1), 183–197. https://doi.org/10.17509/dc5c2131

Ryu, H., & Choi, S. (2024). Metalinguistic Explanations and Their Impact on Incidental Grammar Acquisition: An Eye-Movement Study. *Language Teaching Research*. https://doi.org/10.1177/13621688241286668

Saricaoglu, A., & Bilki, Z. (2025). The Capacity of ChatGPT-4 for L2 Writing Assessment: A Closer Look at Accuracy, Specificity, and Relevance. *Annual Review of Applied Linguistics, 45*, 253–273. https://doi.org/10.1017/S0267190525100160

Shofia, N. K., & Suyitno, I. (2020). Problematika Belajar Bahasa Indonesia Mahasiswa Asing. *BASINDO: Jurnal Kajian Bahasa, Sastra Indonesia, dan Pembelajarannya, 4*(2), 204–214. https://journal2.um.ac.id/index.php/basindo/article/view/7955

Shukla, M., Pandey, D., Kaur, S., Agarwal, M., Goyal, A., & Sharma, H. (2025). Evaluating the Accuracy and Explanatory Quality of Large Language Models ChatGPT, Claude, DeepSeek, Gemini, Grok, and Le Chat in Statistical Test Selection for Hypothesis Testing Decisions. *Cureus, 17*(10), Article e94949. https://doi.org/10.7759/cureus.94949

Steiss, J., Tate, T., Graham, S., Cruz, J., Hebert, M., Wang, J., Moon, Y., Tseng, W., Warschauer, M., & Olson, C. B. (2024). Comparing the Quality of Human and ChatGPT Feedback of Students' Writing. *Learning and Instruction, 91*, Article 101894. https://doi.org/10.1016/j.learninstruc.2024.101894

Sudaryanto, S., & Widodo, P. (2020). Common European Framework of Reference for Languages (CEFR) dan Implikasinya bagi Buku Ajar BIPA. *Jurnal Idiomatik: Jurnal Pendidikan Bahasa dan Sastra Indonesia, 3*(2), 80–87. https://doi.org/10.46918/idiomatik.v3i2.777

Supriyono, A. Y., Rumalean, I., Triyana, E., Kurniawan, S. A., & Juwanda. (2025). Pemanfaatan AI (Artificial Intelligence) dalam Penulisan Karya Ilmiah Bagi Mahasiswa. *Gaba-Gaba: Jurnal Pengabdian kepada Masyarakat dalam Bidang Bahasa dan Seni, 5*(1), 1-9. https://doi.org/10.30598/gabagabavol5iss1pp1-9

Sweller, J., van Merriënboer, J. J. G., & Paas, F. (2019). Cognitive Architecture and Instructional Design: 20 Years Later. *Educational Psychology Review, 31*(2), 261–292. https://doi.org/10.1007/s10648-019-09465-5

Syafruddin, S., Pratiwi, B., Izzati, A. N., & Sianipar, I. F. (2025). Teaching Indonesian for Foreign Speakers (BIPA): A Bibliometric Analysis. *Issues in Language Studies, 14*(1), 55–73. https://doi.org/10.33736/ils.7551.2025

Thornbury, S. (1999). *How to Teach Grammar*. Pearson Education.

Wei, Q., Yao, Z., Cui, Y., Wei, B., Jin, Z., & Xu, X. (2024). Evaluation of ChatGPT-Generated Medical Responses: A Systematic Review and Meta-Analysis. *Journal of Biomedical Informatics, 151*, Article 104620. https://doi.org/10.1016/j.jbi.2024.104620

Yan, D. (2024). Comparing Individual Vs. Collaborative Processing of ChatGPT-Generated Feedback: Effects on L2 Writing Task Improvement and Learning. *Language Learning & Technology, 28*(1), 1–19. https://doi.org/10.64152/10125/73597

Yang, C. T.-Y., & Chen, H. H.-J. (2025). ChatGPT and L2 Chinese Writing: Evaluating the Impact of Model Version and Prompt Language on Automated Corrective Feedback. *Computer Assisted Language Learning, 39*(4), 1119–1147. https://doi.org/10.1080/09588221.2025.2453205

Zare, J., Al-Issa, A., & Ranjbaran Madiseh, F. (2025). Interacting with ChatGPT in essay writing: A study of L2 Learners' Task Motivation. *ReCALL, 37*(3), 385–402. https://doi.org/10.1017/S0958344025000035

APPENDIX A. FULL TEXT OF THE FOUR PROMPTS

The four prompts below were submitted to ChatGPT (GPT-4o) in June 2026. Each was entered as the opening turn of a separate conversation session, with no custom instructions, no system prompt, and no follow-up turns, under default model settings. All four follow a single uniform frame, varying only in the grammatical concept named.

Prompt 1 (multifunctional di-)

"Explain the difference between the functions of di- as a passive verb prefix, di as a preposition, and di as a marker of adverbial phrases of place in Indonesian. Include definitions, rules, and example sentences."

Prompt 2 (meN- prefixation system)

"Explain the meN- prefixation system in Indonesian and its allomorphic variations. Include definitions, rules, and example sentences."

Prompt 3 (suffixes -kan and -i)

"Explain the suffixes -kan and -i in Indonesian and the functional differences between them. Include definitions, rules, and example sentences."

Prompt 4 (circumfixes me-kan and me-i)

"Explain the circumfixes me-kan and me-i in Indonesian and their contrasting semantic roles. Include definitions, rules, and example sentences."