المجلة الليبية للدراسات الأكاديمية المعاصرة (Libyan Journal of Contemporary Academic Studies)
LJCAS, eISSN: 5970-3005
Volume 4, Issue 1, 2026, Pages: 386-395

# The Problem of Surface-Level Fluency in AI-Generated Texts: A Discourse-Based Method to Evaluate Second Language Writing

Zeena Al-Asi *
Department of English, Faculty of Education, Zuwara, University of Zawia, Libya
***Email (for reference researcher):** zynhalasy7@gmail.com

# مشكلة الطلاقة السطحية في النصوص المولدة بالذكاء الاصطناعي: منهج قائم على تحليل الخطاب لتقييم الكتابة باللغة الثانية

زينة العاصي *
قسم اللغة الإنجليزية، كلية التربية - زوارة، جامعة الزاوية، ليبيا

**Received:** 03-01-2025; **Accepted:** 14-03-2026; **Published:** 28-03-2026

**Abstract:**
AI tools can produce writing that looks smooth at first glance. The sentences are neat. The grammar is often clean. The wording may sound mature. Yet strong surface fluency does not always mean strong writing. This problem matters in second language writing because teachers, raters, and automated systems may reward polish more than meaning. The present paper addresses that risk through a discourse-based method for evaluating second language writing. The paper draws on public theory, public corpora, and published experiments. Key sources include Halliday and Hasan's work on cohesion, Coh-Metrix and TAACO, TOEFL11, the Write & Improve Corpus 2024, ASAP 2.0, GPT-based scoring studies, and the DECOR benchmark. The paper argues that AI-generated texts often compress local errors while leaving wider discourse limits in place. These limits include weak task focus, thin support, unstable topic movement, and shallow conclusions. In response, the paper proposes a Surface-to-Discourse Method. The method separates surface fluency from discourse depth, scores both layers, and then computes a Fluency-Discourse Gap. A small gap signals balance. A large gap signals polished but thin writing. The paper also shows how the method can guide feedback, revision, and fairer classroom judgment. The main claim is simple. Smooth language should not be treated as a full sign of writing quality. In second language settings, whole-text meaning still needs close reading.

**Keywords:** AI-generated text, second language writing, discourse analysis, writing evaluation, coherence, cohesion, automated writing evaluation.

**الملخص**
يمكن لأدوات الذكاء الاصطناعي إنتاج كتابة تبدو متقنة من النظرة الأولى؛ فجملها منسقة، وقواعدها سليمة غالباً، ومفرداتها تبدو ناضجة. ومع ذلك، فإن الطلاقة السطحية القوية لا تعني دائماً جودة الكتابة؛ وتبرز أهمية هذه المشكلة في سياق كتابة اللغة الثانية(L2) ، حيث قد يميل المعلمون والمقيمون والأنظمة الآلية إلى مكافأة المظهر المصقول للنص أكثر من المعنى الجوهري. تعالج هذه الورقة تلك المخاطر عبر طرح منهجية قائمة على الخطاب لتقييم الكتابة باللغة الثانية، مستندة في ذلك إلى النظريات الأكاديمية، والمدونات اللغوية العامة، والتجارب المنشورة. وتعتمد الدراسة على مصادر محورية تشمل أعمال "هاليداي وحسن" في التماسك النصي، وأدوات (Coh-Metrix) و(TAACO)، ومدونات (TOEFL11) و (Write & Improve 2024)، و(ASAP 2.0)، بالإضافة إلى دراسات التقييم القائمة على نماذج (GPT) ومعيار.(DECOR)
تجادل الورقة بأن النصوص المولدة بالذكاء الاصطناعي غالباً ما تضغط الأخطاء الموضعية وتخفيها، لكنها تبقي على قيود خطابية أوسع؛ تشمل ضعف التركيز على المهمة المطلوبة، وهشاشة الدعم بالحجج، وعدم استقرار الانتقال بين الموضوعات، والنتائج الضحلة. واستجابةً لذلك، تقترح الورقة "منهجية السطح إلى الخطاب(Surface-to-Discourse Method) "، التي تفصل بين الطلاقة السطحية وعمق الخطاب، وتقوم بقياس كل طبقة على حدة، ثم حساب "فجوة الطلاقة والخطاب". فبينما تشير الفجوة الصغيرة إلى التوازن، تعكس الفجوة الكبيرة كتابة مصقولة ولكنها جوفاء. كما توضح الورقة كيف يمكن لهذا المنهج توجيه التغذية الراجعة، والمراجعة، والتقييم العادل داخل الفصول الدراسية. إن المطلب الرئيسي للبحث بسيط: لا ينبغي اعتبار اللغة المصقولة دليلاً كاملاً على جودة الكتابة، ففي سياقات اللغة الثانية، لا يزال المعنى الشامل للنص يتطلب قراءة فاحصة ودقيقة.

**الكلمات المفتاحية :**النصوص المولدة بالذكاء الاصطناعي؛ اللغة الثانية؛ الطلاقة السطحية؛ تحليل الخطاب؛ تقييم الكتابة؛ التماسك النصي؛ فجوة الطلاقة والخطاب.

## 1. Introduction

Generative AI has changed how writing is produced, revised, and judged. Many systems can now produce short essays in seconds. These essays often look strong on first reading. Their sentences are well formed. Their grammar is often stable. Their wording may sound confident. This kind of polish can shape human judgment very quickly. That first impression can hide a deeper problem. A text may sound fluent while still saying little. It may keep a clear sentence shape but lose a clear line of thought. It may repeat one idea in new words without adding support. It may move from paragraph to paragraph without real development. In short, the writing may look finished at the sentence level while remaining weak at the discourse level.

This issue matters for second language writing. Many L2 writers already work under pressure from grammar, vocabulary, and time. If AI tools help them clean local errors, the result may appear strong before deeper review begins. Teachers may then face a hard task. They must decide whether the text shows real writing growth or only local repair. Automated systems face the same difficulty. If a system rewards polish more than discourse, it may give high scores to texts that are smooth but thin.

Work in discourse studies has long shown that writing quality depends on more than sentence correctness. Halliday and Hasan (1976) described cohesion as a key resource for text building. Later tools such as Coh-Metrix and TAACO made it easier to study cohesion and related discourse features across larger sets of texts (McNamara et al., 2014; Crossley et al., 2016). Research in L2 writing also found that cohesive and discourse features are linked to writing quality, though not in a simple way (Crossley et al., 2016; Kim & Crossley, 2018).

Recent AI studies sharpen this issue. GPT-based systems can rate writing with promising accuracy in some settings (Mizumoto & Eguchi, 2023; Yancey et al., 2023). GPT-4 also shows strength in discourse coherence rating (Naismith et al., 2023). Yet newer work also reports weaker specificity when AI comments on coherence and content-related writing traits (Saricaoglu & Bilki, 2025). Zhang et al. (2024) further note that many writing systems still rely on basic surface features when dealing with coherence.

The present paper responds to that gap. It asks a direct question. How can evaluators keep the useful part of fluency judgment without confusing it with full writing quality? The paper offers a practical answer through a discourse-based method called the Surface-to-Discourse Method. The method gives surface fluency and discourse depth separate scores. It then compares them through a Fluency-Discourse Gap. The purpose is not to punish fluent writing. The purpose is to stop fluency from standing in for meaning.

This paper is a method paper grounded in public evidence. It does not report a new large-scale live experiment. Instead, it builds its argument from public corpora, public benchmarks, and published studies. This choice keeps the paper transparent and reproducible. It also fits the current need in L2 writing research. Before stronger assessment systems can be built, the field needs a clearer way to name the problem and judge it.

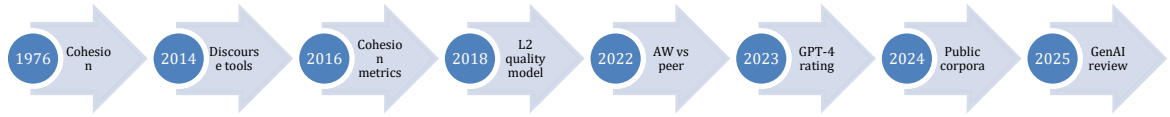


**Figure 1** Key theory and public resource milestones behind the study.

## 2. Background and review of related work

Writing evaluation has often moved between two levels of attention. One level focuses on local form. It asks whether sentences are correct, smooth, and readable. The other level focuses on whole-text meaning. It asks whether ideas develop, connect, and reach a clear purpose. These two levels support each other, but they are not the same.

The discourse tradition helps explain that difference. Halliday and Hasan (1976) showed that texts hold together through ties such as reference, conjunction, substitution, ellipsis, and lexical cohesion. Their work made an important point. A text is not just a set of correct sentences. It is a network of relations across the whole piece. That insight still matters for L2 writing today.

Computational tools later expanded this line of work. Coh-Metrix was designed to measure text and discourse features at scale (McNamara et al., 2014). TAACO added a wide set of local, global, and overall cohesion indices, and it was validated against expert judgments of coherence and essay quality (Crossley et al., 2016). These tools made discourse analysis easier to apply across large learner corpora.

Research using these tools has shown that discourse features matter for writing quality. Crossley et al. (2016) studied 57 L2 university writers over a semester. Their results showed that local, global, and text cohesion features explained 36% of the variance in judgments of text cohesion and 42% of the variance in overall writing quality. They also found that cohesion features predicted whether an essay came from the start or end of the course with 71% accuracy. This finding suggests that discourse growth is real, measurable, and tied to stronger writing.

Kim and Crossley (2018) extended this picture in a larger standardized writing setting. Using 480 TOEFL iBT responses, they modeled lexical, syntactic, and cohesive features together. Their structural model explained 81.7%

of the variance in L2 writing quality. One key result is especially useful here. Higher-rated essays showed greater lexical overlap between paragraphs. That is a discourse signal, not only a grammar signal.

L2 writing classrooms also show the same tension. Chen and Cui (2022) compared automated writing evaluation and peer feedback in a continuation writing task. Students who received peer feedback showed greater gains in cohesive device use and cohesive chain formation than students who received AWE feedback. This does not mean automated help is useless. It means local correction alone may not push discourse growth far enough.

The spread of generative AI changed the scene again. Systems such as ChatGPT and GPT-4 can score or comment on essays with high speed. Mizumoto and Eguchi (2023) used GPT-based scoring on all 12,100 essays in TOEFL11. They found a useful level of accuracy and reliability, and they also showed that linguistic features could improve scoring. Yancey et al. (2023) found that GPT-4 could rate short L2 essays almost as well as modern AWE methods when calibration examples were given. Yet performance varied by first language. This result warns against simple trust in one score.

Other studies move closer to discourse. Naismith et al. (2023) trained GPT-4 to assess written discourse coherence. Their findings showed strong agreement with human ratings. GPT-4 also outperformed traditional NLP coherence metrics in that study. This result is important because it suggests that newer models can engage with discourse better than older scoring systems.

Still, newer evidence also shows limits. Saricaoglu and Bilki (2025) studied ChatGPT-4 as an L2 writing assessor across task response, coherence and cohesion, lexical resource, and grammatical range and accuracy. Their results showed high accuracy across all four dimensions. However, specificity was much weaker in task response and coherence and cohesion. In plain terms, the system often saw that something was wrong, but it did not always explain the problem in a clear and useful way.

A related line of work now targets coherence repair. Zhang et al. (2024) introduced DECOR, a benchmark for incoherence detection, reasoning, and rewriting in L2 English writing. Their paper states that existing automated writing evaluation systems mainly use basic surface linguistic features to detect coherence. DECOR is important because it shifts the field from simple detection to reason-based repair.

Recent review work places these findings in a wider frame. Li (2025) argues that GenAI research in L2 writing shows promise, but also reveals clear gaps in feedback evaluation, discourse comparison, and ethical use. One point in that review is especially relevant here. High scoring agreement does not prove that human and AI raters are using the same construct. That issue stands at the center of the present paper.

**Table 1** Key published studies that support the present paper

| Study | Public setting | Main value for this paper |
|---|---|---|
| 1976 | Cohesion in English | Shows that text quality depends on links across the whole text, not only sentence correctness. |
| 2016 | L2 university essays | Finds that cohesion features relate to text quality and writer development. |
| 2018 | TOEFL iBT responses | Shows that cohesive features remain important in strong L2 quality models. |
| 2022 | Continuation writing task | Shows that peer feedback can improve cohesion and coherence more than AWE alone. |
| 2023 | TOEFL11 and BEA studies | Shows that GPT-based scoring can be useful, but variation and construct issues remain. |
| 2024 | DECOR benchmark | Moves coherence work from simple detection toward explanation and rewriting. |
| 2025 | Recent reviews and validation work | Warns that high agreement does not prove full discourse coverage. |

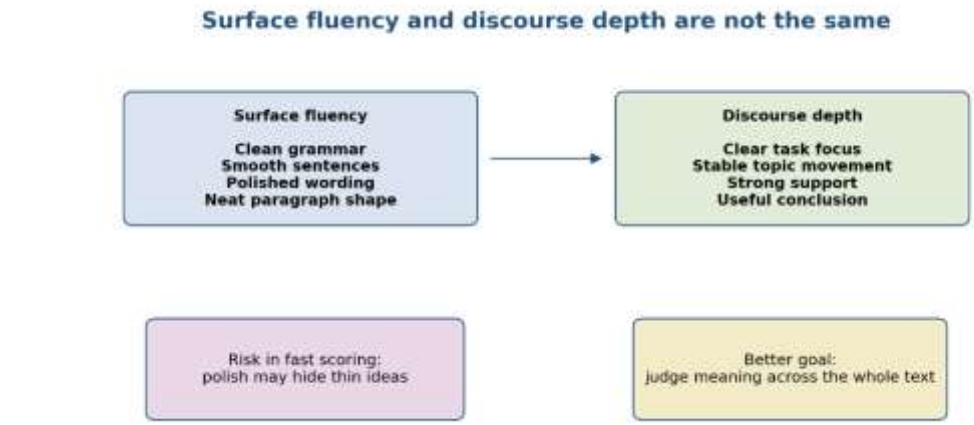


**Figure 2** Surface fluency and discourse depth are related, but not the same.

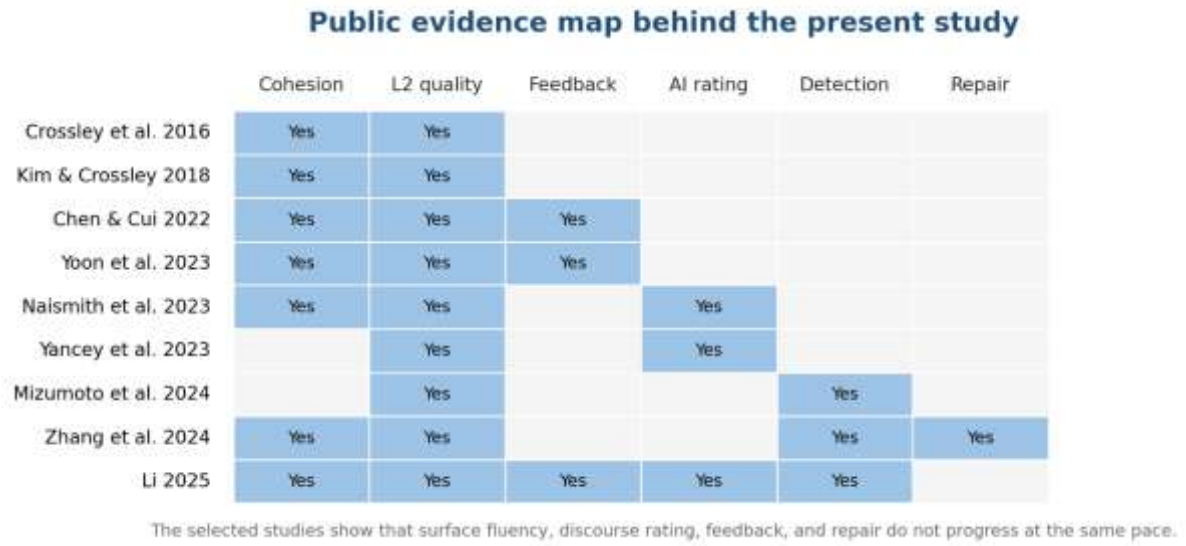


**Figure 3** Public evidence map for the discourse-based method.

**3. The problem of surface-level fluency in AI-generated texts**

The core problem can be stated simply. AI-generated texts often gain quick trust because they look fluent. Yet this fluency may stay near the surface. It may not reflect strong meaning work across the whole text. That mismatch becomes risky when evaluators treat surface polish as a full sign of writing quality.

Surface-level fluency includes several visible traits. A text may have clean grammar. It may use longer or more varied sentences. It may avoid spelling problems. It may sound formal and controlled. Paragraphs may also look neat on the page. These features matter, especially in L2 writing. Writers need help with accuracy and sentence control. The problem begins when these features dominate judgment.

Discourse depth asks different questions. Does the text answer the task fully? Does each paragraph add a clear step in the argument? Do ideas connect across the whole piece? Does support move beyond general statements? Does the conclusion do more than restate the opening? These are harder questions. They take slower reading. They also depend on context, purpose, and logic.

Generative AI can reduce local friction very well. It can repair grammar. It can suggest transitions. It can smooth tone. It can turn short clauses into polished sentences. For that reason, AI-generated or AI-revised texts may look stronger than they are. The local surface becomes cleaner, while the discourse core changes little. Writers and teachers may then confuse repair with development.

This confusion is not only a classroom issue. It affects scoring validity. If an automated system gives high scores to polished but underdeveloped essays, the score may reward the wrong construct. If teachers rely on a fast first impression, they may do the same. Li (2025) notes that high agreement between human and AI scoring does not prove shared construct coverage. Saricaoglu and Bilki (2025) make a similar point from another angle. Their study showed that ChatGPT-4 could be accurate, but its feedback on coherence and cohesion was often too general.

The problem is sharper in L2 contexts for three reasons. First, many L2 writing rubrics still include grammar, vocabulary, and coherence in the same score band. This can blur the line between local control and whole-text meaning. Second, L2 writers may use AI mainly for sentence repair, which changes the visible layer faster than

the deeper layer. Third, some AI detectors may misread L2 writing patterns, which creates another source of unfair judgment (Li, 2025).

Public benchmarks support this concern. DECOR was created because coherence detection and repair in L2 writing remained underdeveloped (Zhang et al., 2024). The Write & Improve Corpus 2024 also shows why draft history matters. The corpus includes multiple versions of the same learner essays, not only final texts (Nicholls et al., 2024). This makes it possible to study how local correction and wider discourse growth interact over time.

The phrase surface-level fluency is useful because it names a real assessment trap. It does not dismiss fluency. Good writing needs fluency. Rather, the phrase points to a layer problem. Fluency is one layer. Discourse depth is another. If the first layer improves much faster than the second, the gap itself becomes meaningful.

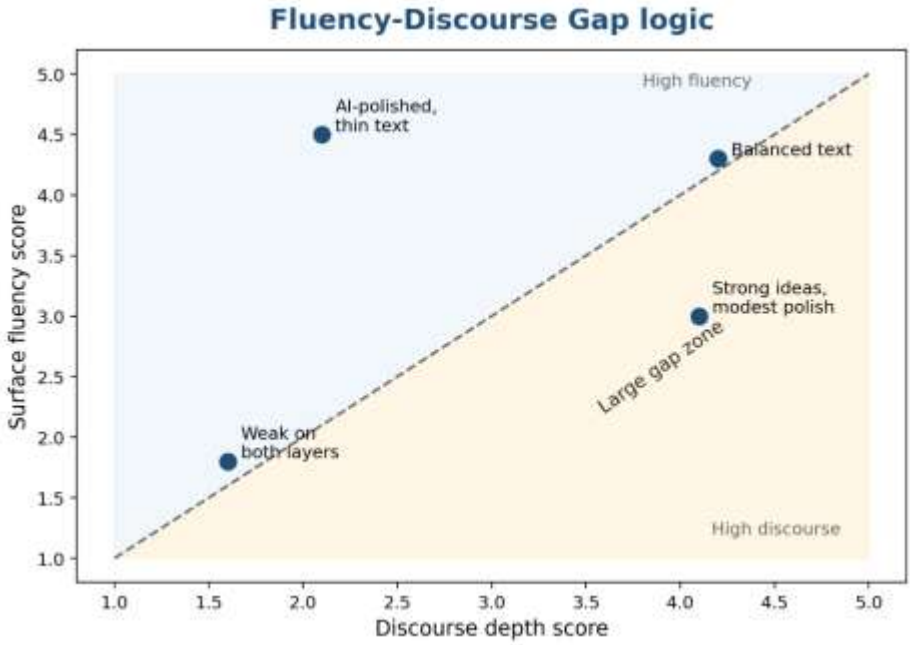


**Figure 4** The logic of the Fluency-Discourse Gap.

**4. A discourse-based method for evaluating second language writing**

This paper proposes a practical procedure called the Surface-to-Discourse Method, or SDM. The method is designed for teachers, raters, researchers, and tool designers. It can be used with human judgment alone or with computational support. Its main goal is to separate two kinds of strength that are often merged too early.

The first part of the method scores surface fluency. This score focuses on local control. It asks whether grammar is mostly stable, whether sentence phrasing is smooth, whether word choice is broadly fitting, and whether paragraph shape is neat enough for easy reading. The score does not ask whether the argument is deep. It asks whether the language surface creates a fluent reading experience.

The second part scores discourse depth. This score focuses on whole-text meaning. It asks whether the response keeps a clear task focus, whether ideas move in a steady order, whether support is specific enough, whether paragraph links are meaningful, and whether the closing section adds a real ending. This score may use discourse tools, but it also needs close reading.

The method then compares the two scores. The comparison produces a Fluency-Discourse Gap. The gap can be expressed in a simple way.

Fluency-Discourse Gap = Surface Fluency Score - Discourse Depth Score.

If the gap is small, the text is fairly balanced. If the gap is large and positive, the text is polished but thin. If both scores are low, the text needs broad support. If discourse depth is stronger than surface fluency, the writer may have meaningful ideas that need language repair.

The value of the gap is diagnostic. It turns a vague feeling into a clear signal. Many teachers already sense that some texts sound better than they think. SDM gives that feeling a usable structure. It also prevents one score from hiding another.

The method uses four steps. First, the evaluator reads the whole text once without scoring. This first pass is only for general sense. Second, the evaluator scores surface fluency on a five-point scale. Third, the evaluator scores discourse depth on a five-point scale. Fourth, the evaluator compares the two scores and writes a short note that explains the gap.

The method works best when the discourse score is anchored to visible indicators. Five indicators are proposed here. The first is task focus. The text should answer the prompt and keep its central purpose. The second is topic movement. Each paragraph should add a clear step. The third is support quality. Claims should be explained or

illustrated. The fourth is connectedness. Links across sentences and paragraphs should build meaning, not only rhythm. The fifth is ending value. The conclusion should do more than repeat prior lines.

The surface score also needs clear anchors. Four indicators are proposed. The first is grammar control. The second is sentence smoothness. The third is lexical fit. The fourth is paragraph cleanliness. These are simpler to judge, and AI systems already handle much of this layer.

SDM can be used in several ways. In classrooms, teachers can use it after AI-assisted drafting. In assessment, raters can use it as a second-pass check after a holistic score. In research, corpus analysts can compare surface and discourse profiles across prompts or groups. In tool design, developers can use the method to decide when a system should offer revision help instead of only a score.

The method also fits current public resources. Coh-Metrix and TAACO can support cohesion inspection (McNamara et al., 2014; Crossley et al., 2016). TOEFL11 provides a large public learner set for scoring experiments (Blanchard et al., 2013). The Write & Improve Corpus 2024 adds multi-draft learner data (Nicholls et al., 2024). ASAP 2.0 expands access to source-based essays for model testing (The Learning Agency Lab, 2024). DECOR provides a benchmark for incoherence detection and rewriting (Zhang et al., 2024).

The method is modest on purpose. It does not claim to replace full rubric systems. It also does not claim that discourse can be reduced to one number. Its value lies elsewhere. It helps evaluators avoid a common mistake. That mistake is to let fluent surfaces stand in for whole-text quality.

**Table 2** The Surface-to-Discourse Method scoring frame.

| Layer | Main indicators | What the evaluator asks |
|---|---|---|
| Surface fluency | Grammar control, sentence smoothness, lexical fit, paragraph cleanliness | Does the text read smoothly at the local level? |
| Discourse depth | Task focus, topic movement, support quality, connectedness, ending value | Does the text develop and connect ideas across the whole response? |
| Gap reading | Difference between the two scores | Is the text balanced, polished but thin, or idea-rich but under-polished? |

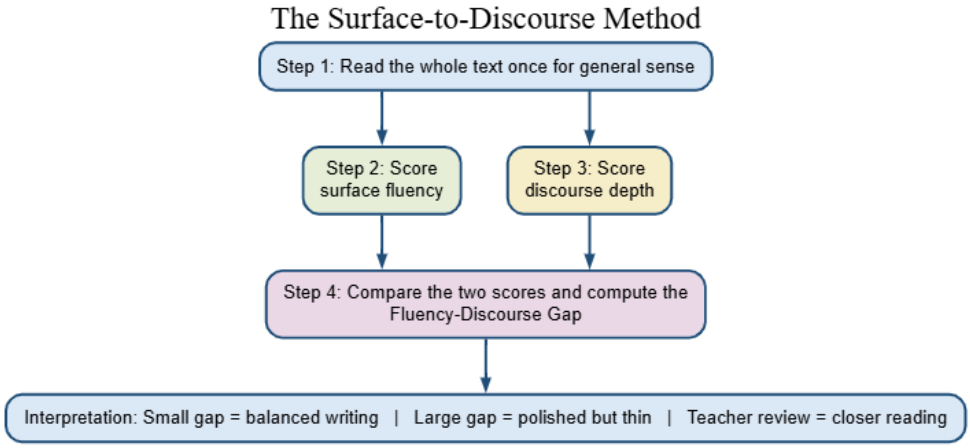


**Figure 5** Architecture of the Surface-to-Discourse Method.

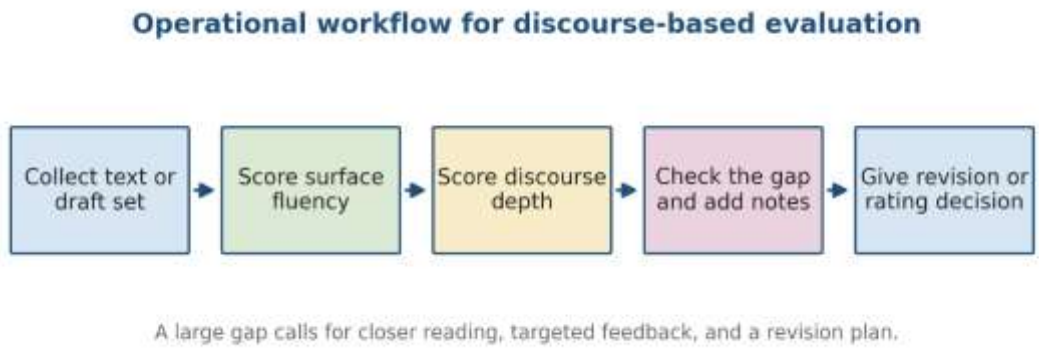


**Figure 6** Workflow for applying the discourse-based method.

**5. Public experiments and practical evidence behind the method**

A method paper needs a public evidence base. SDM is supported by three kinds of evidence. The first comes from discourse theory and cohesion research. The second comes from L2 writing assessment studies. The third comes from newer GPT-based scoring and repair research.

The theory base begins with cohesion. Halliday and Hasan (1976) described the semantic ties that help texts hold together. Their work does not reduce writing quality to grammar. It treats connected meaning as a central property of text. Later computational work made those ties measurable. Coh-Metrix was designed for automated evaluation of text and discourse, with attention to cohesion, readability, and related text features (McNamara et al., 2014). TAACO then made local and global cohesion measures easier to use in research practice (Crossley et al., 2016).

The L2 writing evidence base shows that discourse matters for quality. Crossley et al. (2016) found that cohesion features were linked to both text cohesion judgments and overall writing quality. Their results support two parts of SDM. First, discourse signals can be measured. Second, those signals explain quality beyond local correctness. Kim and Crossley (2018) add a stronger modeling result. Their work showed that cohesive features, together with lexical and syntactic features, explained most of the variance in L2 writing quality. Yet their findings also imply that no single surface layer is enough.

Classroom feedback research shows how this issue appears in practice. Chen and Cui (2022) compared AWE and peer feedback on cohesion and coherence in a continuation writing task. Students with peer feedback achieved greater gains in cohesive device use and cohesive chain formation. This result supports a practical claim of SDM. Revision plans should return writers to meaning work, not only sentence repair.

GPT-based scoring studies add both promise and caution. Mizumoto and Eguchi (2023) used GPT scoring on TOEFL11 and found useful accuracy and reliability. Their study suggests that large language models can support essay scoring at scale. Yancey et al. (2023) found that GPT-4 could rate short L2 essays nearly as well as modern AWE methods when calibration examples were provided. Yet their results also varied by first language. This matters because a clean score can still hide uneven fairness.

Naismith et al. (2023) move closer to the discourse concern of this paper. Their GPT-4 system produced discourse coherence ratings that matched human ratings well and outperformed traditional coherence metrics. This finding matters for SDM in two ways. It suggests that discourse-aware AI evaluation is now feasible. It also shows that newer systems can move beyond shallow surface counting.

Even so, the newer literature still points to a specificity problem. Saricaoglu and Bilki (2025) found that ChatGPT-4 was very accurate across L2 writing dimensions, including coherence and cohesion. However, its feedback in those discourse-related dimensions was often less specific than its feedback on grammar or lexis. This result supports one of the main claims of the present paper. AI may detect discourse trouble, but it may still describe that trouble weakly.

The repair literature adds another missing piece. Zhang et al. (2024) introduced DECOR because L2 writers often struggle with coherence and because earlier systems did little more than detect surface problems. DECOR includes expert annotations for incoherence detection, reasoning, and rewriting. The benchmark is especially useful for future SDM work because it ties discourse diagnosis to actual revision.

Public corpora make practical testing possible. TOEFL11 contains 12,100 essays by writers from 11 language backgrounds and has long been useful for scoring and educational NLP work (Blanchard et al., 2013). The Write & Improve Corpus 2024 includes 5,050 user-prompt sets, more than 23,000 essays, draft histories, CEFR labels, and grammatical annotations (Nicholls et al., 2024). ASAP 2.0 adds about 24,000 source-based essays and was released to support better essay scoring and feedback development (The Learning Agency Lab, 2024). Together, these resources can support future validation work on the proposed method.

This paper therefore does not build SDM from intuition alone. The method stands on a converging public record. Discourse theory explains why whole-text meaning matters. L2 writing studies show that discourse signals predict quality. Feedback studies show that sentence repair is not enough. GPT studies show that AI can help, but may still blur construct boundaries. Benchmarks such as DECOR show the next step: detection, explanation, and repair at the discourse level.

**Table 3** Public corpora, benchmarks, and resources that can support validation work.

| Resource | Public scope | Use for this paper |
|---|---|---|
| TOEFL11 | 12,100 essays from 11 L1 groups | Useful for scoring tests and comparison of learner profiles. |
| Write & Improve Corpus 2024 | 5,050 prompt-user sets and more than 23,000 essays | Useful for multi-draft revision tracking and CEFR-linked analysis. |
| ASAP 2.0 | About 24,000 source-based essays | Useful for prompt-linked scoring and feedback testing. |
| DECOR | Benchmark for incoherence detection, reasoning, and rewriting | Useful for discourse diagnosis and repair studies. |

**Table 4** Published public experiments and what they show.

| Study | Practical or experiment | What it shows |
|---|---|---|
| Crossley et al. (2016) | L2 course essays over time | Cohesion features help explain essay quality and development. |
| Kim and Crossley (2018) | Modeled 480 TOEFL responses | Cohesive features remain strong inside broad quality models. |
| Chen and Cui (2022) | Compared AWE with peer feedback | Meaning-focused feedback can improve cohesion and coherence more clearly. |
| Mizumoto and Eguchi (2023) | GPT scoring on TOEFL11 | Large language models can support scoring, but need careful framing. |
| Naismith et al. (2023) | GPT-4 coherence rating | Newer models can judge discourse coherence better than older metrics. |
| Zhang et al. (2024) | DECOR benchmark | Discourse repair should include detection, reasoning, and rewriting. |

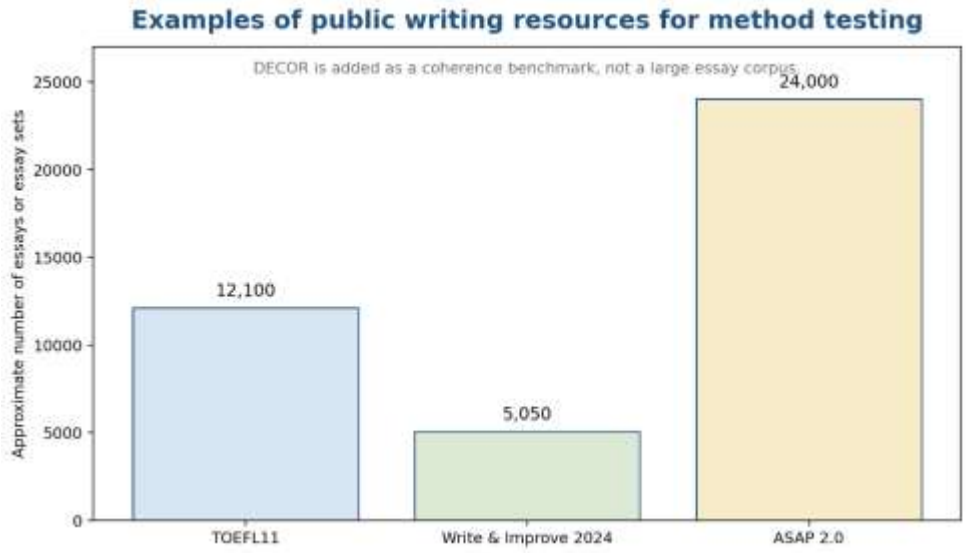


Figure 7 Public resources that can support future validation work.

**6. An illustrative application of the method**

To show how SDM works, this section presents an illustrative comparison. The example is not a new experiment. It is a demonstration of the scoring logic.

Imagine a short argumentative response written with AI help. The text has clean grammar, varied sentences, and polished transitions. It opens with a broad claim about online learning. The second paragraph repeats the same claim in new wording. The third paragraph adds one example, but the example is generic. The conclusion restates the opening line almost word for word.

A fast reader may score this text highly because it feels smooth. Under SDM, the evaluator first reads without scoring. The second pass then gives the text a surface fluency score. Because grammar is stable and sentence flow is smooth, the text may receive 4.5 out of 5 on surface fluency.

The third pass then scores discourse depth. Here the rating changes. The task is answered only in a general way. Topic movement is limited. Support is thin. The conclusion adds little. Paragraph links are present, but they mostly act as formal connectors. A discourse depth score of 2.5 or 3 would therefore be more fitting.

The Fluency-Discourse Gap in this case is large. The text is polished, but its meaning work is limited. That gap guides feedback. The evaluator should not spend most of the next comment cycle on grammar. The next step should target support, focus, and development.

Now imagine a different text by an L2 writer without AI polish. The grammar has visible errors. Some sentences are short. Yet the response answers the task clearly. Each paragraph adds a new point. The examples are specific. The conclusion returns to the claim with a stronger final insight.

A fast holistic score may punish this text more than it should. SDM makes the profile clearer. The surface fluency score may be 2.8 or 3. The discourse depth score may be 4 or 4.2. The gap now runs in the other direction. The writer has strong ideas that need local language support. This is a different teaching case.

The practical value of SDM lies here. Two texts may look very different on the page, yet both need targeted feedback. One needs discourse growth. The other needs language repair. A single overall score can hide that distinction.

The method also supports revision planning. After AI-assisted drafting, the writer can check whether AI mostly improved sentence shape. If yes, the next revision round should work on meaning. This may include adding support, improving paragraph purpose, or reworking the conclusion. The revision loop in this paper makes that process visible.

**Table 5** Example scoring profile under the discourse-based method.

| Case | Surface fluency | Discourse depth | Gap reading |
|---|---|---|---|
| AI-polished but thin text | 4.5/5 | 2.5-3.0/5 | Large positive gap. The draft sounds polished, but ideas stay thin. |
| Idea-strong learner text | 2.8-3.0/5 | 4.0-4.2/5 | Negative gap. Meaning is strong, but local language needs repair. |

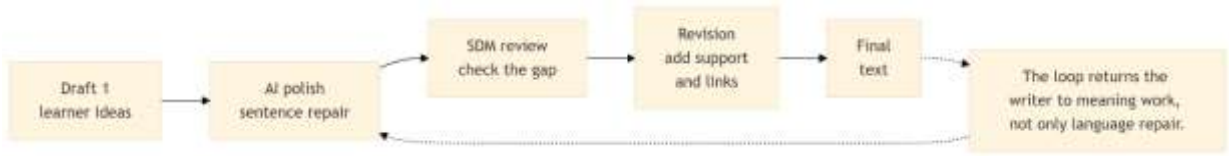


**Figure 8** Revision loop after AI-assisted drafting.

## 7. Discussion

The paper makes a narrow but important claim. In L2 writing, smooth language should not be read as full writing quality. AI tools increase the urgency of that claim because they can improve local fluency very quickly.

The present method does not reject AI. In fact, it treats AI as useful for one part of the writing process. Sentence repair, vocabulary suggestions, and phrasing help can reduce local burden. This is valuable for L2 writers. The problem begins only when that local improvement is treated as a sign that discourse development also improved.

The method also has implications for scoring fairness. If teachers and systems reward surface polish too heavily, they may over-score AI-polished but underdeveloped texts. At the same time, they may under-score learner texts that show strong meaning but weak sentence control. SDM tries to reduce both errors.

There are limits to the method. Discourse depth still requires judgment. Even with tools, no simple index captures all of coherence, support, audience fit, and rhetorical force. The method also needs validation across prompts, levels, and language backgrounds. Public corpora make that validation possible, but the work remains to be done.

Another limit concerns AI change itself. Models are improving fast. A gap that is large today may shrink in some domains tomorrow. Yet this does not weaken the method. It makes it more useful. As models improve, evaluators need a clearer way to check whether surface gains and discourse gains are moving together.

## 8. Pedagogical and assessment implications

In classroom use, the method can support a two-pass response style. The first pass gives the writer a balanced profile. The second pass gives targeted next steps. This can save teacher time because feedback becomes more selective. It can also help students see why a polished draft may still need more work.

In formal assessment, the method can work as a validity check. If a script receives a high holistic score but also shows a large Fluency-Discourse Gap, the rater can review the script more closely. This does not force a lower score. It simply slows a risky judgment.

In tool design, the method suggests a shift in output style. Instead of returning one overall number, systems should show separate layer scores. They should also explain what kind of revision is needed next. When the gap is large, the tool should prioritize discourse prompts, not only grammar prompts.

## 9. Conclusion

The rise of AI-generated and AI-revised writing has made one old problem more visible. Texts can look fluent before they become strong. In L2 writing, that gap matters because teachers, raters, and systems often work under time pressure. Fast judgment can then reward polish more than meaning.

This paper addressed that problem through a discourse-based method. The proposed Surface-to-Discourse Method separates surface fluency from discourse depth, then compares them through a Fluency-Discourse Gap. The method is simple, but it rests on a strong public base. Discourse theory, cohesion tools, public corpora, GPT-based scoring studies, and coherence benchmarks all support the need for a layered view of writing quality.

The main message is direct. AI can improve sentence form quickly. It cannot be assumed to improve whole-text meaning at the same pace. Evaluation must therefore look at both layers. When the surface grows faster than the discourse, the gap itself becomes the finding.

**Compliance with ethical standards**

*Disclosure of conflict of interest*

The author(s) declare that they have no conflict of interest.

**References**

1. Blanchard, D., Tetreault, J., Higgins, D., Cahill, A., & Chodorow, M. (2013). *TOEFL11: A corpus of non-native English* (ETS Research Report No. RR-13-24). ETS Research Report Series. https://doi.org/10.1002/j.2333-8504.2013.tb02331.x
2. Chen, M., & Cui, Y. (2022). The effects of AWE and peer feedback on cohesion and coherence in continuation writing. *Journal of Second Language Writing, 57*, 100915. https://doi.org/10.1016/j.jslw.2022.100915
3. Crossley, S. A., Kyle, K., & McNamara, D. S. (2016). The development and use of cohesive devices in L2 writing and their relations to judgments of essay quality. *Journal of Second Language Writing, 32*, 1-16. https://doi.org/10.1016/j.jslw.2016.01.003
4. Aisha Mohamed Ahmed. (2026). Globalization and Technology in English: Structural Shifts, Digital Usage, and Cultural Implications. African Union Journal of Academic and Research Studies, 1(1), 40-55.Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in English*. Longman.
5. Kim, M., & Crossley, S. A. (2018). Modeling second language writing quality: A structural equation investigation of lexical, syntactic, and cohesive features in source-based and independent writing. *Assessing Writing, 37*, 39-56. https://doi.org/10.1016/j.asw.2018.03.002
6. Li, S. (2025). Generative AI and second language writing. *Digital Studies in Language and Literature, 2*(1). https://doi.org/10.1515/dsll-2025-0007
7. McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge University Press. https://doi.org/10.1017/CBO9780511894664
8. Mizumoto, A., & Eguchi, M. (2023). Exploring the potential of using an AI language model for automated essay scoring. *Research Methods in Applied Linguistics, 2*(2), 100050. https://doi.org/10.1016/j.rmal.2023.100050
9. Naismith, B., Mulcaire, P., & Burstein, J. (2023). Automated evaluation of written discourse coherence using GPT-4. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)* (pp. 394-403). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.bea-1.32
10. Nicholls, D., Caines, A., & Buttery, P. (2024). *The Write & Improve Corpus 2024: Error-annotated and CEFR-labelled essays by learners of English*. University of Cambridge Repository. https://doi.org/10.17863/CAM.112997
11. 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
12. The Learning Agency Lab. (2024). *ASAP 2.0 dataset*. https://the-learning-agency-lab.com/learning-exchange/asap-2-0-dataset/
13. Yancey, K. P., Laflair, G., Verardi, A., & Burstein, J. (2023). Rating short L2 essays on the CEFR scale with GPT-4. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)* (pp. 576-584). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.bea-1.49
14. Zhang, X., Diaz, A., Chen, Z., Wu, Q., Qian, K., Voss, E., & Yu, Z. (2024). DECOR: Improving coherence in L2 English writing with a novel benchmark for incoherence detection, reasoning, and rewriting. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (pp. 11436-11458). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.emnlp-main.639