Synonym Substitution and Sentence Restructuring Tools: A New Frontier of Contract Cheating - A Critical Review and Conceptual Analysis

Parhlad Singh Ahluwalia, Academician, Dr. Bhimrao Ambedkar Law University, Jaipur, Rajasthan, India

Abstract

The global expansion of digital writing assistance has quietly produced a new and under-examined vector of academic dishonesty: automated synonym-substitution ("word-spinning") and sentence-restructuring ("paraphrasing" or "spinning") software. Unlike classical contract cheating, in which a third party is commissioned to write an entire assignment, these tools allow a student or author to take an existing text — their own prior submission, a peer's work, an internet source, or increasingly the output of a generative AI system — and mechanically transform its surface form while leaving its underlying ideas, structure, and argumentative sequence intact. Because most first-generation plagiarism detectors rely on lexical and n-gram matching, such transformations frequently defeat similarity checks even though no genuine intellectual transformation has occurred. This paper synthesises the literature on contract cheating, textual plagiarism, and machine paraphrasing to argue that synonym-substitution and restructuring tools constitute a distinct and escalating category of contract cheating — one that is arguably more corrosive to assessment validity than essay mills because it is free, instantaneous, and normalised as a "writing aid." Drawing on published detection benchmarks, taxonomies of obfuscation strategies, and institutional policy documents (including India's UGC 2018 anti-plagiarism regulations), the paper develops a conceptual taxonomy of spinning technologies, reviews empirical detection-accuracy data, and examines the problem from student, educator, technologist, and policy perspectives. The analysis shows that detection performance varies enormously by obfuscation technique — from near-perfect detection of naïve word-spinners to substantial detection failure against large-language-model (LLM) paraphrasers — and that regulatory frameworks written for a copy-paste era are poorly equipped to classify AI-mediated rewriting. The paper concludes with a multi-layered set of recommendations spanning assessment design, detection technology, disclosure policy, and digital literacy education.

Keywords: *contract cheating; synonym substitution; sentence restructuring; paraphrasing tools; word spinning; plagiarism detection; academic integrity; large language models*

1. Introduction

Academic integrity has always rested on a simple contract: the work submitted for credit is the work of the person receiving that credit. Every technology that intervenes between a writer's ideas and the words on the page — the typewriter, the internet search engine, the online essay bank, and now the generative language model — has forced a renegotiation of that contract. Plagiarism itself has never been a binary phenomenon. As Sinha, Singh and Kumar observed in their widely cited review of unethical publication practices, plagiarism exists on a spectrum, from wholesale reproduction of another author's work to the far more common and far more ambiguous practice of "substituting the synonyms and editing the original text" [1]. It is precisely this ambiguous middle ground — cosmetic rewriting that leaves the intellectual content of a source essentially untouched — that has metastasised into a commercial and technological ecosystem over the past decade.

The term "contract cheating" was coined by Clarke and Lancaster to describe the outsourcing of assessed work to a third party for payment [2]. In their original formulation, the defining feature was economic: a contractor is remunerated to produce work that a student then passes off as their own. Subsequent research has repeatedly documented the scale of this problem. In a systematic review of 65 studies conducted between 1978 and 2018, Newton found that self-reported contract cheating rose from an average of roughly 3.5% across the full historical period to 15.7% among studies published between 2014 and 2018, and extrapolated this to approximately 31 million students worldwide [3]. Draper and colleagues' investigation of essay mill business models further showed how professionalised, low-margin "writer" economies have grown around this demand [4], while Medway and colleagues' covert investigation of UK essay mills catalogued the reassurance techniques — money-back guarantees, plagiarism reports, "undetectability" promises — that these services use to convert anxious students into paying customers [5].

However, the classical essay-mill model requires money, time, and a degree of premeditated risk (engaging an unknown third party, transmitting payment, waiting for delivery). Synonym-substitution and sentence-restructuring tools remove all three frictions. They are free or nearly free, they operate in seconds, and — critically — they are marketed not as cheating tools but as writing-improvement, paraphrasing, or "plagiarism-avoidance" software. This branding matters enormously for how students perceive the ethics of their use, a point developed at length by Ahluwalia in his analysis of the theory and practice of producing "plagiarism-free" content, which distinguishes between genuine paraphrastic synthesis (understanding a source and re-expressing its ideas in one's own analytical framework) and mechanical word substitution that never engages with the source's meaning at all [6]. Ahluwalia's companion study of detection techniques and tools further demonstrates that most first-generation plagiarism checkers were architected around exactly the failure mode that synonym substitution exploits: they compare surface strings, not underlying propositions, and are therefore structurally blind to well-executed lexical substitution [7].

This paper treats synonym-substitution and sentence-restructuring software — hereafter referred to collectively as "spinning tools," following the terminology established in the computational plagiarism-detection literature — as a **new frontier of contract cheating**: not because they involve a paid human contractor, but because they perform the same functional substitution of authorial labour, at a lower cost and with a lower perceived moral weight, while still producing text that is not genuinely the submitting student's own analytical work when applied to material the student did not themselves author. The paper's specific objectives are to:

- Develop a conceptual taxonomy of spinning technologies, from early Markov-chain-based "article spinners" through neural paraphrasers to LLM-mediated rewriting;
- Synthesise published, citable empirical evidence on the detectability of each category of tool;

- Analyse the phenomenon from four stakeholder perspectives — student, educator/institution, technologist, and policymaker;
- Situate the discussion within a concrete regulatory case (India's UGC 2018 Regulations) to illustrate the gap between existing anti-plagiarism law and AI-mediated obfuscation; and
- Propose a layered set of pedagogical, technological, and policy responses.

2. Literature Review

2.1 From Essay Mills to Text Spinners: A Widening Definition of Contract Cheating

The classical contract-cheating literature has historically bracketed synonym substitution as a "minor" or "grey area" offence, distinct from the wholesale outsourcing that defines the essay-mill economy. Yet several threads in the literature push toward treating it as a first-class member of the same family.

First, definitionally, Clarke and Lancaster's original formulation was never restricted to human ghost-writers; it referred to "the process of offering the completion of an assignment out to tender" [2], a framing broad enough to include any mechanism — human or automated — by which the actual composing labour is displaced away from the student. Second, empirically, Amigud and Lancaster's work on the essay-mill economy (discussed in Draper et al. [4]) shows that the price students are willing to pay for outsourced writing falls steeply as perceived detection risk falls; free tools that promise "plagiarism-proof" output occupy exactly the price-and-risk niche predicted by that economic logic, effectively functioning as a zero-cost essay mill substitute. Third, and most importantly, the *function* performed by a spinning tool is identical to that performed by a ghost-writer: it inserts a layer between the source of the ideas (another author, another assignment, or the student's own previous unoriginal draft) and the words submitted for assessment, and it does so specifically to defeat the institution's verification mechanism.

Kannangara's early empirical study of "word-spinning software" is instructive here because it was conducted on genuine student submissions at a New Zealand polytechnic rather than on synthetic examples [8]. Examining assignments generated with commercial spinning software, Kannangara documented "significant quality degradation" in the spun text — incoherent synonym choices, broken idiom, syntactic anomalies — alongside near-total evasion of the plagiarism checks then in use, and argued that the absence of a "legal" plagiarism finding (because the checker registered low similarity) did not remove the underlying ethical violation, since the student had neither written nor genuinely understood the submitted content [8]. This is one of the earliest peer-reviewed demonstrations that automation could decouple *detectability* from *authenticity* — precisely the property that makes spinning tools dangerous as a contract-cheating vector: they weaponise the gap between what a similarity score measures and what academic integrity actually requires.

2.2 Taxonomy of Obfuscation: Word Spinning, Sentence Restructuring, and Translingual Spinning

The plagiarism-detection literature has developed a fairly stable taxonomy of "obfuscation" strategies used to disguise reused text, and it is useful to map spinning tools onto that taxonomy rather than treating "paraphrasing tools" as a monolith.

Lexical substitution (word spinning). The earliest and crudest tools operate by replacing individual words with thesaurus-derived synonyms, sometimes using nested "spintax" syntax (e.g., `{quick|fast|rapid}`) that allows a single template to generate thousands of superficially distinct outputs for search-engine content-farming purposes before this technique migrated into the student-cheating ecosystem. Commercial tools of this type (SpinBot and similar products are repeatedly named in the empirical literature as representative examples [9]) typically leave sentence structure, argument order, and paragraph organisation completely untouched.

Syntactic/sentence-level restructuring. A more sophisticated category changes clause order, converts active to passive voice, merges or splits sentences, and substitutes synonymous phrases rather than single words, while still preserving the original semantic content almost exactly. Products such as SpinnerChief and, later, machine-learning-based paraphrasers (e.g., consumer tools built on sequence-to-sequence or transformer architectures) fall into this category [9].

Translingual ("back-translation") spinning. Akbari's study of what he terms "spinning-translation" documents a further technique in which a source text is machine-translated into an intermediate language and then translated back into the original language, producing text that is semantically equivalent to the source but lexically and syntactically quite different, without the user ever engaging a dedicated "paraphrasing" product at all [10]. Akbari's analysis is particularly important for this paper's argument because it explicitly frames the technique as a *resistance strategy against plagiarism detection* used by students, and shows that ordinary, freely available translation software — not any purpose-built cheating tool — is sufficient to produce contract-cheating-grade obfuscation.

AI/LLM-mediated rewriting. The newest and most consequential category uses large generative language models, either through dedicated "AI paraphraser" products or through direct prompting of general-purpose chatbots ("rewrite this in different words," "make this sound less like it was copied"), to perform simultaneous lexical, syntactic, and even rhetorical-register transformation. Ansorge, Ansorgeová and Sixsmith's case study of a real plagiarised scientific article demonstrates the forensic signature of this category particularly well: they show that a paraphrasing tool applied to a plagiarised passage introduces characteristic errors — terminological inconsistency, semantically garbled sentences, and even corrupted references and missing tables — that would not occur in ordinary human paraphrase, and that these artefacts can themselves become a *detection signal* precisely because the automated rewriting process is imperfect [11]. Cabanac, Labbé and Magazinov's identification of "tortured phrases" — bizarre, over-specific synonym substitutions such as "counterfeit consciousness" appearing in place of "artificial intelligence" — provides a parallel and now well-known forensic marker of automated synonym substitution operating on technical vocabulary in the scientific-publishing context, illustrating that the problem extends well beyond student assignments into peer-reviewed research itself [12].

2.3 Detection Science: What the Evidence Actually Shows

A central claim of this paper is that the *detectability* of spinning-tool output is not a fixed property but a variable that depends heavily on the sophistication of the obfuscation technique — a claim well supported by the computational plagiarism-detection literature.

Foltýnek, Meuschke and Gipp's systematic literature review of 239 papers on academic plagiarism detection remains the most comprehensive synthesis of detection methods available [13]. Their review classifies detection approaches into character-based, syntax-based, semantic-based, cross-lingual, and citation-based methods, and concludes that while character- and syntax-based methods (the basis of almost all commercial similarity checkers) perform well against verbatim and lightly obfuscated copying, they degrade substantially against paraphrased and "idea" plagiarism, whereas citation-pattern-based and semantic-similarity methods retain more of their detection power under heavy obfuscation, albeit at much higher computational cost and with far less mainstream institutional deployment [13].

The most directly relevant empirical benchmark, however, comes from Wahle, Ruas, Kirstein and Gipp's 2022 study, which is the first to systematically compare detection performance across both classical spinning tools and neural, LLM-based paraphrase generation [9]. Working with real product output from SpinBot and SpinnerChief alongside paraphrases generated by T5 and GPT-3, the authors evaluated six automated detection approaches plus one commercial plagiarism-detection tool, and additionally ran a human study with 105 participants. Three findings from this study are especially important for the argument advanced here:

- Detection performance is highly technique-dependent: the best-performing automated detector (a fine-tuned Longformer model) achieved an F1-score of 99.68% against SpinBot-generated text but only 71.64% against SpinnerChief-generated text, despite both tools belonging to the same broad "spinning software" category [9].
- Human evaluators performed *worse* than the best automated detector against both tools (F1 = 78.4% for SpinBot, 65.6% for SpinnerChief), showing that reliance on instructor judgement alone is not a reliable backstop [9].
- Against LLM-generated paraphrases (GPT-3), mean human detection accuracy fell to just 53%, essentially coin-flip performance, while GPT-3-authored paraphrases were independently rated by human experts as equal in clarity (4.0/5) and close in fluency (4.2/5) and coherence (3.8/5) to the original, non-plagiarised texts [9].

Table 1 synthesises this evidence into a comparative taxonomy.

Table 1. Taxonomy of synonym-substitution and restructuring tools, mechanism, and reported detectability

Category	Representative mechanism	Typical output artefacts	Reported detection performance (best automated method)	Reported human detection performance	Key source
Word spinning (lexical)	Dictionary/thesaurus word swap, sometimes nested "spintax"	Broken idiom, wrong word sense, structure unchanged	F1 = 99.68% (Longformer vs. SpinBot)	F1 = 78.4%	Wahle et al. 2022 [9]
Syntactic restructuring	Rule-based or shallow ML clause reordering, active↔passive conversion	Awkward phrasing, structure altered, meaning preserved	F1 = 71.64% (Longformer vs. SpinnerChief)	F1 = 65.6%	Wahle et al. 2022 [9]
Translingual ("spinning-translation")	Machine translation to intermediate language and back	Idiom loss, register shift, occasional mistranslation	Not systematically benchmarked; reported as evasive in case analyses	Not systematically benchmarked	Akbari 2021 [10]
Neural/LLM paraphrasing (T5, GPT-3-class)	Transformer-based sequence generation conditioned on source text	Fluent, near-human prose; few surface artefacts	Best model accuracy ≈ 66% (F1)	Mean accuracy 53%	Wahle et al. 2022 [9]
AI-assisted "humanising" of AI text	LLM paraphrase applied specifically to disguise AI-generated drafts	Fluent but occasionally semantically diluted; may retain statistical "rhythm" of	Increasingly targeted by combined similarity + AI-writing classifiers	Highly variable, tool- and rater-dependent	Pal et al. 2023 [14]; Ahluwalia 2025b [15]

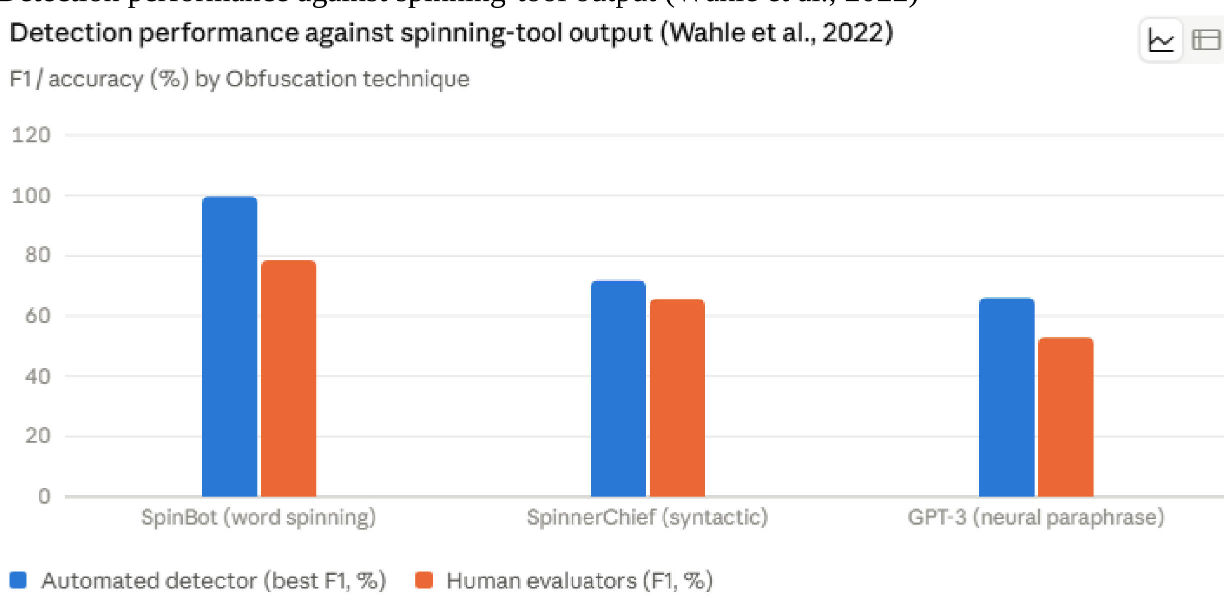
		source AI model			

Note: Figures for word spinning and syntactic restructuring rows are drawn directly from the SpinBot/SpinnerChief comparison reported in Wahle et al. (2022) [9]; figures for neural/LLM paraphrasing are drawn from the same study's GPT-3 experiments. This table is a synthesis for expository purposes and should be read alongside the original studies for full methodological detail.

The pattern in Table 1 is striking and, for the purposes of this paper's argument, decisive: as spinning technology has moved from crude lexical substitution toward neural and LLM-based rewriting, detection accuracy for both machines and humans has fallen sharply, while the fluency and "naturalness" of the output — the two qualities that determine whether a marker will *suspect* anything at all — has risen to match genuine human writing. In other words, the technological trajectory of spinning tools is moving in exactly the direction that maximises their usefulness for contract-cheating-style evasion.

To make this comparison visually explicit, Figure 1 plots the F1-scores reported by Wahle et al. [9] for the best-performing automated detector against each tool category.

Detection performance against spinning-tool output (Wahle et al., 2022)



Detection performance against spinning-tool output (Wahle et al., 2022)
F1 / accuracy (%) by Obfuscation technique

Obfuscation technique	Automated detector (best F1, %)	Human evaluators (F1, %)
SpinBot (word spinning)	99.68	78.4
SpinnerChief (syntactic)	71.64	65.6
GPT-3 (neural paraphrase)	66	53

Figure 1. As obfuscation sophistication increases from simple word spinning to neural paraphrase generation, both automated and human detection performance decline steeply — with human accuracy falling to near-chance levels against GPT-3-class output. Data reproduced from Wahle, Ruas, Kirstein and Gipp (2022) [9].

2.4 The AI Inflection Point

Pal, Bhattacharya, Islam and Chakraborty's editorial commentary on ChatGPT and scientific writing crystallises why the emergence of general-purpose LLMs represents a qualitative rather than merely incremental shift in the spinning-tool landscape [14]. They note that AI-generated content has "exacerbated" plagiarism concerns specifically by lowering the effort required to disguise reused material, and cite earlier work examining the ethical ramifications of AI technologies capable of generating text that evades conventional plagiarism detectors [14]. Ahluwalia's analysis of artificial intelligence's role in combating plagiarism in scholarly publishing extends this observation into a genuinely double-edged framework: the same generative and semantic-analysis capabilities that allow an LLM to produce a highly convincing paraphrase are, in principle, the capabilities needed to build a next-generation *detector* capable of recognising semantic equivalence rather than lexical overlap [15]. Ahluwalia argues that AI-assisted detection — embedding-based semantic similarity, stylometric consistency analysis, and machine-generated-text classifiers — represents the most promising counter-technology, but cautions that detection tools available to most institutions still lag several technology generations behind the generation tools students and researchers can access for free [15].

This asymmetry — powerful, freely available generation technology versus institutionally-constrained, frequently paywalled or under-resourced detection technology — is, this paper argues, the structural condition that has turned synonym substitution and sentence restructuring from a minor stylistic nuisance into a genuine contract-cheating vector at scale.

2.5 Institutional and Regulatory Response: The UGC 2018 Case

A useful case study of the regulatory lag described above is India's University Grants Commission (Promotion of Academic Integrity and Prevention of Plagiarism in Higher Educational Institutions) Regulations, 2018 [16]. The Regulations were a genuinely significant policy achievement: they created a graded, four-level penalty structure tied to measured textual similarity, mandated that every higher-education institution establish a Departmental Academic Integrity Panel and an Institutional Academic Integrity Panel, required training of academic staff in the use of plagiarism-detection tools, and obliged institutions to maintain dedicated detection infrastructure [16]. Crucially, however, the Regulations — like almost every national anti-plagiarism policy drafted before 2020 — operationalise plagiarism almost entirely in terms of a *percentage textual-similarity score* against an indexed corpus. This design choice, sensible for the copy-paste plagiarism the Regulations were written to address, creates exactly the structural blind spot that spinning tools are built to exploit: text that has been sufficiently reworded returns a low similarity score and is treated, for regulatory purposes, as if it were original, irrespective of whether any genuine authorial transformation occurred. The Regulations contain no explicit provision addressing AI-generated or AI-paraphrased text, a gap that is now widely noted in secondary commentary on the policy [16] and that mirrors the situation in most other jurisdictions' academic-integrity frameworks.

3. Conceptual Framework

Building on the literature reviewed above, this paper proposes a simple two-dimensional conceptual framework for classifying contract-cheating-adjacent text-transformation behaviour (Figure 2, described narratively below since it synthesises the argument rather than reporting external data).

The two dimensions are:

- **Provenance of ideas** — ranging from fully original authorial thought, through synthesised understanding of multiple sources (legitimate paraphrase/citation), to verbatim appropriation of a single uncited source.
- **Degree of mechanical transformation** — ranging from no transformation (verbatim copy), through cosmetic lexical/syntactic transformation (spinning), to genuine cognitive re-expression (true paraphrase, which requires comprehension and re-derivation of the argument in the writer's own analytical voice).

Classical plagiarism sits at the intersection of low idea-originality and low transformation (verbatim copying of an uncited source). Classical contract cheating sits at low idea-originality but *high apparent* transformation, because a human ghost-writer produces genuinely original prose, even though the ideas and effort are not the submitting student's own. Synonym-substitution and sentence-restructuring tools occupy a previously under-theorised third quadrant: low idea-originality **and** low-to-moderate *genuine* transformation, but an output profile that is engineered to statistically resemble the high-transformation quadrant closely enough to defeat both algorithmic and human observers. It is this specific property — mimicking the *signature* of legitimate transformation without performing the *cognitive act* that legitimate transformation requires — that justifies treating spinning tools as a distinct and named category rather than folding them indiscriminately into either "plagiarism" or "contract cheating" as traditionally defined.

4. Methodology

This paper adopts a **narrative-critical review methodology**, appropriate to a fast-moving, cross-disciplinary phenomenon that spans computer science (plagiarism-detection algorithms), higher-education policy (academic-integrity regulation), and applied linguistics (paraphrase and translation studies), for which a single narrowly-scoped systematic review would risk omitting relevant strands of evidence. Sources were identified through targeted search of the peer-reviewed and grey literature on contract cheating, academic plagiarism detection, machine/AI paraphrasing, and Indian academic-integrity regulation, supplemented by the author-nominated works of Ahluwalia on plagiarism-free content practice, detection techniques, and AI's role in combating plagiarism [6, 7, 15]. Priority was given, wherever possible, to sources reporting quantifiable detection-performance data (accuracy, F1-score, precision/recall) rather than anecdotal or purely descriptive claims, since a central objective of the paper is to move the discussion of "AI paraphrasing detectability" beyond folk claims (of the kind common in student-facing blog content) toward citable, peer-reviewed evidence. Where the available literature reports only descriptive or case-study findings (as with the Kannangara and Ansorge et al. case studies), these are presented explicitly as qualitative evidence rather than generalised statistically. No new human-subjects data collection, tool testing, or plagiarism-detection experiment was conducted for this paper; all quantitative figures reported here are drawn from, and cited to, the original published studies. This methodological choice is itself an ethical one: a paper on plagiarism and contract cheating should model the practice of transparent, verifiable citation of others' original empirical work rather than manufacture apparent primary data.

5. Analysis and Discussion

5.1 Perspective One: The Student

From the perspective of the student user, the appeal of spinning and paraphrasing tools rests on a set of psychological and situational factors well documented in the broader academic-integrity literature and consistent with Sinha, Singh and Kumar's observation that minor, "editing-level" plagiarism is normalised precisely because it does not feel like the theft that wholesale copying does [1]. Four factors appear particularly salient:

- **Moral distancing.** Because the tool, not the student, performs the substitution, students can attribute the resulting text to "using a writing tool" rather than "copying," even when the underlying propositional content is entirely unoriginal. Ahluwalia's distinction between genuine paraphrase (which requires comprehension) and mechanical substitution (which does not) is precisely the distinction that this moral distancing elides [6].
- **Time and cost asymmetry versus essay mills.** A spinning or AI-paraphrasing tool is free or low-cost and near-instantaneous, in contrast to the days-long turnaround and non-trivial expense associated with commissioning an essay-mill writer, as documented in the essay-mill business-model literature [4]. This dramatically lowers the threshold at which a time-pressured student will choose to cheat.
- **Perceived low detection risk**, which, as Table 1 shows, is often *accurate* rather than merely perceived, particularly for neural and LLM-based tools [9].
- **Legitimate use overlap.** Many spinning and paraphrasing tools are marketed — and genuinely used by a great many students — for legitimate purposes: improving grammar, adjusting register for a target audience, or assisting English-as-an-additional-language writers in producing fluent prose from their own draft ideas. This legitimate-use overlap makes outright prohibition difficult to justify or enforce and is a recurring theme in survey-based research on paraphrasing-tool awareness among research scholars.

5.2 Perspective Two: The Educator and Institution

From the institutional perspective, the central problem exposed by this literature is a *false sense of security* generated by similarity-score-based detection infrastructure. An instructor or integrity panel that treats a low Turnitin-style similarity percentage as equivalent to "no plagiarism" is, on the evidence synthesised in Section 2.3, making an inference that the underlying detection technology does not support once spinning or AI-paraphrasing tools are in play — a point implicit throughout Foltýnek, Meuschke and Gipp's systematic review, which explicitly separates "text-matching" performance from genuine "idea plagiarism" detection performance and finds the former far more mature than the latter [13]. Kannangara's finding that word-spun assignments could achieve near-zero similarity scores despite serious quality and originality problems is a direct, empirically grounded illustration of this false-security effect operating in a real assessment context [8].

This creates a difficult institutional dilemma. Overreliance on any single automated similarity score risks both false negatives (spun plagiarism passing undetected) and, especially with the newer generation of "AI-writing detectors," false positives (flagging legitimate non-native-English writing, heavily edited drafts, or genuinely original writing that happens to share statistical properties with AI output). Ahluwalia's proposed synthesis — combining semantic-similarity analysis, stylometric consistency checking across a student's writing history, and disclosure-based policy rather than relying on any single detection technology — reflects an emerging institutional consensus that no purely technological fix is currently sufficient on its own [7, 15].

5.3 Perspective Three: The Technologist

From a computational perspective, the arms-race dynamic documented by Wahle et al. is the central analytic fact [9]. Detection systems built around character- and syntax-level matching (the dominant commercial paradigm, exemplified by classical Turnitin-style architecture) are close to a local performance ceiling against naïve word-spinning (F1 ≈ 99.7%) precisely because such spinning preserves sentence structure almost perfectly, leaving strong syntactic fingerprints even after lexical substitution. Performance degrades meaningfully against tools that also alter syntax (F1 ≈ 71.6% for SpinnerChief) and degrades further still — to levels not meaningfully better than chance for human raters — against neural paraphrase generation that can simultaneously vary vocabulary, syntax, and even rhetorical register while preserving meaning [9]. Foltýnek, Meuschke and Gipp's broader review situates this specific finding within a wider pattern: citation-pattern-based detection (analysing whether a document's pattern of in-text citations matches a suspected source, independent of the exact wording used) and cross-lingual semantic-similarity methods retain substantially more robustness against heavy obfuscation than lexical matching does, but both are computationally expensive, harder to deploy at the scale of routine undergraduate coursework checking, and much less widely adopted in commercial detection products [13]. The practical consequence is a persistent, non-trivial gap between the state of the art in academic plagiarism-detection *research* and the detection capability actually deployed in the median university's assessment pipeline — a gap the technologist's perspective identifies as the single most tractable lever for near-term improvement, since it requires deployment and integration effort rather than fundamentally new algorithmic breakthroughs.

5.4 Perspective Four: The Policymaker

The UGC 2018 case study illustrates a broader structural problem in policy design: regulations calibrated to a specific detection technology (percentage textual similarity) become obsolete as the underlying threat model shifts, without any change to the regulation's text [16]. This is not a criticism specific to Indian policy — comparable percentage-similarity-anchored frameworks are common internationally — but the UGC Regulations are a particularly clear illustration because their four-tier penalty structure is explicitly defined by similarity-score bands, meaning that a spun or AI-paraphrased submission that returns a low score is, by the letter of the regulation, classified as non-plagiarised, regardless of the reviewing panel's independent judgement. Policymakers therefore face a choice between three broad responses: (a) redefine plagiarism and contract cheating in output-agnostic, intent- and process-based terms (e.g., requiring disclosure of any tool used to transform text, regardless of the resulting similarity score); (b) mandate continuous reinvestment in detection technology tied to a rolling definition of "current best-available method"; or (c) shift assessment design itself away from formats (unsupervised, take-home written essays) that are maximally vulnerable to any form of contract cheating, spun or otherwise, toward formats — oral defence, in-class writing, iterative drafting with visible version history, viva-style questioning — that are structurally resistant to post-hoc textual substitution regardless of its sophistication. The literature on contract cheating prevention increasingly converges on option (c) as the most durable response, precisely because it is invariant to whichever specific spinning or generation technology becomes dominant next.

5.5 Synthesis: Why Spinning Tools Deserve Classification as Contract Cheating

Pulling the four perspectives together, this paper argues for explicitly classifying synonym-substitution and sentence-restructuring tool use — when applied to material the user did not themselves originate — as a form of contract cheating, on four grounds:

- **Functional equivalence.** As argued in Section 2.1, the tool performs the same substitution of authorial labour for the appropriated ideas that a human ghost-writer performs in classical contract cheating; only the mechanism (software rather than a paid person) differs.
- **Deliberate evasion architecture.** The commercial marketing and design of word-spinning and many "AI humaniser" tools is explicitly organised around defeating institutional detection, a feature shared with essay mills' "plagiarism-free guarantee" marketing documented by Medway et al. [5] and directly evidenced by Kannangara's finding of near-total detection evasion despite serious quality degradation [8].
- **Escalating accessibility mirrors essay-mill growth dynamics.** Just as Newton's data show contract cheating prevalence rising as essay-mill accessibility increased [3], the near-zero cost and instant availability of spinning and LLM-paraphrasing tools plausibly represents a further, larger step-change in accessibility, consistent with Pal et al.'s and Ahluwalia's warnings about AI's exacerbating effect on plagiarism risk [14, 15].
- **Empirically demonstrated evasion capability.** Unlike much discourse on this topic, which relies on anecdote, the Wahle et al. benchmark provides direct, quantified evidence that sophisticated spinning and paraphrasing genuinely defeats both algorithmic and human detection at rates far above chance for the most advanced tools [9] — precisely the evidentiary standard that should be required before classifying any activity as a serious integrity threat rather than a stylistic curiosity.

6. Table: Multi-Perspective Summary

Table 2. Stakeholder perspectives on synonym-substitution and restructuring tools

Perspective	Primary concern	Key evidence	Implication
Student	Moral distancing; time/cost pressure;	Normalisation of "editing-level"	Requires literacy-based, non-punitive first-line intervention

	overlap with legitimate use	plagiarism [1]; free/instant tool access	
Educator/Institution	False security from low similarity scores	Near-zero similarity despite quality degradation [8]; text-matching vs. idea-plagiarism gap [13]	Requires multi-signal assessment of authenticity, not single-score reliance
Technologist	Detection accuracy collapses as obfuscation sophistication rises	F1 99.68%→71.64%→66% across spinning tiers [9]	Requires investment in semantic/stylometric, not purely lexical, detection
Policymaker	Regulations anchored to similarity-score bands become obsolete	UGC 2018 four-tier similarity structure with no AI provision [16]	Requires output-agnostic, process/disclosure-based regulation and assessment redesign

7. Recommendations

Drawing on the analysis above, the paper proposes a layered response operating simultaneously at four levels.

7.1 Assessment design. Institutions should reduce reliance on unsupervised take-home written assignments as the sole or dominant assessment vehicle in courses at high risk of contract cheating (a risk profile documented, for related outsourcing behaviour, in the discipline-level contract-cheating literature). Complementary formats — in-class writing, oral defence of submitted work, staged submission with visible drafting history — are structurally resistant to any form of post-hoc text substitution, spun or otherwise, because they require real-time demonstration of authorial understanding.

7.2 Detection technology. Consistent with Foltýnek, Meuschke and Gipp's conclusion that integrated, multi-method detection (combining lexical, semantic, and citation-pattern analysis) outperforms any single method against sophisticated obfuscation [13], institutions should prioritise detection tools that move beyond pure string-matching, while remaining realistic — following Ahluwalia's caution [15] — that detection technology will likely always lag one generation behind generation technology, and should therefore never be treated as a sufficient response on its own.

7.3 Policy and disclosure. Regulatory frameworks such as the UGC 2018 Regulations should be updated to define permissible and impermissible use of text-transformation tools explicitly and to require disclosure of any such tool's use, independent of the resulting similarity score, following the logic developed in Section 5.4. A disclosure-based model shifts the compliance burden toward transparency rather than toward an arms race with detection software.

7.4 Digital and academic literacy education. Perhaps the most durable intervention, and the one most directly supported by Ahluwalia's practice-oriented work on producing genuinely plagiarism-free content [6], is direct instruction distinguishing mechanical synonym substitution from genuine scholarly paraphrase — teaching students *why* the cognitive act of comprehending and re-deriving an argument constitutes the ethically and pedagogically meaningful boundary, rather than teaching to a detection-score threshold that a spinning tool can trivially satisfy without that cognitive act ever occurring.

8. Limitations

This review has several limitations that should qualify its conclusions. First, as a narrative-critical rather than fully systematic review, it does not claim exhaustive coverage of every published study on machine paraphrasing or contract cheating, and some relevant literature — particularly rapidly emerging work on the very latest generation of consumer AI-humaniser tools launched after 2024 — may postdate the sources synthesised here. Second, the quantitative detection-performance figures reported in Table 1 and Figure 1 are drawn from a small number of benchmark studies (principally Wahle et al. [9]) conducted on specific corpora (arXiv articles, student theses, Wikipedia) and specific tool versions; detection accuracy for any given tool changes over time as both generation and detection technology are updated, so the absolute figures should be read as illustrative of a *pattern* (declining detectability with increasing obfuscation sophistication) rather than as current, tool-specific performance guarantees. Third, this paper synthesises existing evidence and does not present new experimental data of its own, a deliberate methodological choice explained in Section 4 but one that means the analysis cannot speak directly to detection performance against the most recent generation of tools without further dedicated empirical study.

9. Conclusion

Synonym-substitution and sentence-restructuring tools have moved from a marginal curiosity in the plagiarism-detection literature to a central and rapidly evolving threat to assessment validity. The evidence reviewed here shows that this is not merely an impressionistic concern: empirical benchmarks demonstrate that as spinning technology has progressed from crude lexical substitution to neural and large-language-model-based paraphrase generation, both automated and human detection performance have collapsed toward — and in the case of human raters against GPT-3-class output, essentially reached — chance levels [9], even as the underlying content remains entirely unoriginal to the submitting author. This paper has argued that this pattern justifies classifying sophisticated use of such tools as a genuine, if technologically mediated, form of contract cheating: functionally equivalent to commissioning a ghost-writer, deliberately architected to evade institutional verification, and now dramatically more accessible than the essay-mill economy it increasingly displaces. Existing regulatory frameworks, illustrated here through the case of India's UGC 2018 Regulations, remain anchored to a similarity-score paradigm that these tools are specifically engineered to defeat, and existing detection infrastructure — despite genuine sophistication in the research literature — remains unevenly deployed at the institutional level. Addressing this new frontier will require a coordinated response across assessment design, detection technology investment, disclosure-based policy reform, and, above all, sustained investment in the kind of academic-writing literacy that teaches students to distinguish the mechanical rearrangement of someone else's words from the genuine, effortful, and ultimately far more valuable act of understanding an idea well enough to express it, honestly, in their own.

References

1. Sinha R, Singh G, Kumar C. Plagiarism and unethical practices in literature. Indian J Ophthalmol. 2009 Nov-Dec;57(6):481-5.
2. Clarke R, Lancaster T. Eliminating the successor to plagiarism? Identifying the usage of contract cheating sites. In: Proceedings of the 2nd International Plagiarism Conference; 2006 Jun 19-21; Newcastle upon Tyne, UK.
3. Newton PM. How common is commercial contract cheating in higher education and is it increasing? A systematic review. Front Educ. 2018;3:67.
4. Draper MJ, Lancaster T, Dann S, et al. Essay mills and other contract cheating services: to buy or not to buy and the consequences of students changing their minds. Int J Educ Integr. 2021;17:13.
5. Medway D, et al. Contract cheating in UK higher education: a covert investigation of essay mills. Br Educ Res J. 2018. doi:10.1002/berj.3335.
6. Ahluwalia PS. How we make plagiarism-free content: theory and practices. Shodh Prakashan J Humanit Soc Sci. 2025:46-53.
7. Ahluwalia PS. Plagiarism: detection techniques and tools. Siddhanta's Int J Adv Res Arts Humanit. 2024 Sep:1-11.
8. Kannangara DN. Quality, ethics and plagiarism issues in documents generated using word spinning software. MIER J Educ Stud Trends Pract. 2017;7(1):24-32.
9. Wahle JP, Ruas T, Kirstein F, Gipp B. How large language models are transforming machine-paraphrased plagiarism. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP); 2022 Dec 7-11; Abu Dhabi, UAE. p.952-963.
10. Akbari A. Spinning-translation and the act of plagiarising: how to avoid and resist. J Furth High Educ. 2021;45(1):49-64.
11. Ansorge L, Ansorgeová K, Sixsmith M. Plagiarism through paraphrasing tools—the story of one plagiarized text. Publications. 2021;9(4):48.
12. Cabanac G, Labbé C, Magazinov A. Tortured phrases: a dubious writing style emerging in science. Evidence of critical issues affecting established journals. arXiv preprint arXiv:2107.06751. 2021.
13. Foltýnek T, Meuschke N, Gipp B. Academic plagiarism detection: a systematic literature review. ACM Comput Surv. 2019;52(6):112.
14. Pal S, Bhattacharya M, Islam MA, Chakraborty C. AI-enabled ChatGPT or LLM: a new algorithm is required for plagiarism-free scientific writing. Int J Surg. 2023 Nov 23. doi:10.1097/JS9.0000000000000939.
15. Ahluwalia PS. The role of artificial intelligence in combating plagiarism in scholarly publishing. Shodh Prakashan J Humanit Soc Sci. 2025:26-30.
16. University Grants Commission. Promotion of Academic Integrity and Prevention of Plagiarism in Higher Educational Institutions Regulations, 2018. New Delhi: UGC; 2018.