WARRANTING WORDS: EMPIRICAL VERIFICATION IN WRITTEN DESCRIPTION DOCTRINE AFTER GENERATIVE AI

Andrew Chin[*]

ABSTRACT

Generative artificial intelligence systems can now draft full-length patent specifications in mere minutes, blending boilerplate prose, hypothetical examples, and apparently authoritative data. Their outputs mimic experienced patent drafting, yet large language models (LLMs) prioritize linguistic plausibility over empirical accuracy. This creates significant risks for patent doctrine, introducing detailed yet fictitious embodiments, molecular sequences, and algorithmic data capable of eluding detection by patent examiners and practitioners. Although prior scholarship addresses such AI-generated errors primarily through the enablement doctrine, emphasizing undue experimentation, this Article argues the doctrinal threat lies equally in patent law's written-description requirement—particularly the Federal Circuit's insistence that disclosures allow skilled artisans to "visualize or recognize" each member of a claimed genus. To fulfill this requirement without demanding working prototypes, patent disclosures must include what this Article terms "warranting words": empirical teachings sufficient to render patentability criteria testable and falsifiable.

Interpreting written description through an ontological lens, the Article clarifies that genus claims inherently rely on a minimal empirical threshold. At filing, patentees must provide textual information sufficient to support the empirical reality of the claimed genus—such as defined assay protocols, calibration evidence demonstrating reliability above inherent statistical hallucination rates, and rationales linking specific tests to broader claims. Without these empirical warrants, generative AI disclosures risk producing patent claims that become mere linguistic abstractions disconnected from any

[*] Paul B. Eaton Distinguished Professor, University of North Carolina School of Law.

real-world anchors.

The Kalai–Vempala theorem establishes that calibrated language models must inevitably produce a statistically predictable rate of fictitious outputs ("hallucinations"), bounded below by the prevalence of "monofacts"—facts encountered exactly once—in their training data. Empirical evidence from large biochemical and materials databases demonstrates monofact prevalences commonly exceeding 15%, implying substantial minimal hallucination rates in patent contexts. Coupled with Freilich's research, which finds that nearly half of biotechnology patents already rely heavily on unverifiable prophetic examples, this creates profound vulnerabilities in the patent system. Generative AI threatens to dramatically amplify these vulnerabilities, promoting claims that superficially comply with current doctrine while lacking empirical substance.

To address these risks, this Article proposes that the phrase "reasonably conveys" from *Ariad Pharmaceuticals v. Eli Lilly* be read as requiring evidence that "reasonably warrants" the patent system's belief in the disclosed invention's empirical reality. Under this clarified standard, generative AI-produced sequences, algorithms, or structural motifs must be accompanied by verifiable evidence—laboratory data, validated simulations, or other robust empirical anchors—adequate to meet the minimal falsifiability threshold without requiring full reduction to practice. This modest doctrinal refinement integrates seamlessly with existing written-description precedent, reinforcing patent law's essential empirical foundation without necessitating legislative change or complex new legal frameworks.

I. INTRODUCTION

Generative AI is rapidly transforming patent drafting, producing disclosures that convincingly mimic rigorous laboratory work—but often without empirical grounding. Consider a patent application claiming "a monoclonal antibody that binds epitope X," whose disclosure enumerates 1,500 amino acid sequences entirely generated by a large language model (LLM). Current doctrine treats each listed sequence as empirically credible, granting broad genus claims while shifting the burden onto examiners, competitors, and courts to determine scientific reality after issuance. Yet, as Kalai and Vempala recently demonstrated, every calibrated language model inevitably "hallucinates" facts at a rate no lower than the prevalence of

"monofacts"—those facts encountered exactly once in its training corpus.[1] Given realistic monofact rates of around 15% in protein databases,[2] approximately 225 of these disclosed sequences would be statistically guaranteed not to bind epitope X at all.

Patent law has long grappled with the tension between linguistic plausibility and empirical verification. As early as 1892, the Supreme Court observed that patent specifications are among "the most difficult legal instruments to draw with accuracy."[3] Generative AI exacerbates this longstanding problem by shifting the effort from cautious, human-driven speculation—traditionally limited by cost—to effortless algorithmic proliferation of scientifically plausible yet statistically unreliable examples. Consequently, what once involved carefully hedged prophetic examples[4] has now become an almost cost-free enumeration of potentially fictional embodiments.

This Article proposes addressing this doctrinal gap through a refined written description standard I term "warranting words." Under this proposal, genus-level patent claims described purely linguistically (*de dicto*) must include sufficient textual information to render each patentability element—such as enablement, utility, and non-obviousness—reasonably open to empirical falsification. In line with Karl Popper's principle of falsifiability, a scientific assertion gains legitimacy not by accumulating confirmations, but by surviving attempts at refutation. The burden on patentees remains modest: they need not definitively prove the genus but must articulate a clear testing protocol, provide calibration data demonstrating reliability above inherent hallucination rates, and offer a coherent rationale linking tested examples to the broader claimed genus. Without this minimal warrant, the genus remains outside the empirical domain patent law implicitly demands.

Two recent scholarly insights underscore the proposal's necessity. Kalai and Vempala establish a rigorous statistical lower bound on hallucination rates for calibrated LLMs, ensuring unavoidable factual

---

[1] Adam T. Kalai & Santosh S. Vempala, *Calibrated Language Models Must Hallucinate*, in Proceedings of the 56th Annual ACM Symposium on Theory of Computing 160, 160–61 (2024).

[2] *See infra* note 62 (justifying this estimate).

[3] Topliff v. Topliff, 145 U.S. 156, 171 (1892) (Brown, J.).

[4] *See* Manual of Patent Examining Procedure § 608.01(p) (9th ed. Rev. 07.2022) (discussing use of prophetic examples).

errors regardless of training data quality.[5] Complementing this theoretical insight, Miao and colleagues empirically confirm that these hallucination rates correlate closely with monofact distributions and cannot be practically mitigated through common data-adjustment techniques such as deliberate miscalibration or extensive data duplication.[6] These findings highlight a systemic vulnerability: broader and rarer genus claims inherently contain a higher proportion of fabricated embodiments.

This vulnerability compounds the existing patent law practice of permitting prophetic examples—hypothetical experiments drafted in present or future tense—as credible disclosures.[7] Freilich has documented that 99% of scientific literature citing these prophetic examples mistakenly treats them as real experimental results, inadvertently propagating misinformation.[8] Generative drafting can amplify this problem exponentially by generating vast quantities of plausible yet entirely fictitious prophetic data.

Rather than eliminating prophetic disclosures altogether, as Freilich has proposed, the warranting words standard preserves beneficial early-stage speculation, provided it is empirically anchored. Unlike existing proposals that focus primarily on tightening enablement and utility doctrines—such as those by Tu, Cyphert, and Perl[9]—the standard introduces falsifiability directly into the written description stage, complementing rather than supplanting current doctrinal tools.

This Article develops and tests this proposal as follows: Part II briefly outlines the technological and doctrinal evolution leading to today's generative drafting practices. Part III establishes the theoretical foundations, integrating Popperian falsifiability principles and recent empirical findings on hallucination rates. Part IV formulates the three-prong warranting words standard, grounded in existing statutory language and PTO practice. Part V applies this standard concretely to

---

[5] Kalai & Vempala, *supra* note 1.

[6] Mingyuan Miao et al., *Hallucination, Calibration, and the Monofact Rate in Large Language Models* 2–3 (2025) (arXiv preprint).

[7] MPEP § 608.01(p) (requiring present tense drafting for prophetic disclosures).

[8] Janet Freilich, *Prophetic Patents*, 53 U.C. DAVIS L. REV. 663, 698–99 (2019).

[9] Sean Tu, Thaddeus T. Cyphert & Matthew A. Perl, *The Limits of Using Artificial Intelligence and GPT-3 in Patent Prosecution*, 54 TEX. TECH L. REV. 255, 276–78 (2022).

representative technologies. Part VI situates the proposal within contemporary scholarly debates. Part VII outlines feasible implementation pathways. Part VIII addresses potential counterarguments. Part IX concludes, emphasizing that requiring falsifiability at the written description stage restores the empirical foundation necessary for meaningful patent discourse in the age of generative AI.

II. DOCTRINAL AND TECHNOLOGICAL BACKGROUND

Patent doctrine has historically evolved alongside technological innovations, but generative AI represents a novel challenge. The sections below succinctly outline how generative AI has reshaped patent drafting, summarize the historical evolution of written description doctrine, and highlight the specific problems posed by prophetic examples amplified by AI-driven drafting. These foundational contexts set the stage for the subsequent introduction of the warranting words standard.

A. Generative AI in Patent Drafting

Large language models (LLMs), such as GPT-4, have radically decreased the marginal cost of creating dense, technical patent disclosures, effectively shifting the patent drafting bottleneck from human labor to the sheer act of invention itself.[10] Platforms like PatentPal exemplify this transformation, allowing practitioners to upload an independent claim document and receive a fully formatted patent specification—complete with technical figures and callouts—in a single click. Tasks previously consuming days of attorney and paralegal labor now take minutes.[11]

The U.S. Patent and Trademark Office (USPTO) has recognized this shift, recently cautioning patent filers that failing to disclose AI's involvement in drafting applications—or submitting unverified AI-generated text—could violate Rule 11.18, given documented instances of generative AI producing convincingly fabricated legal precedent.[12] Internally, PTO examiners are prohibited from using

---

[10] Sequoia Capital, *Generative AI: A Creative New World* (Sept. 19, 2022).

[11] PatentPal, *Generative AI for Intellectual Property* ("1. Drop a document … 3. Export draft into Word and Visio").

[12] Blake Brittain, *USPTO Warns Patent Lawyers Not to Pass Off AI Inventions as*

public AI platforms such as ChatGPT and are instead required to work within secure, monitored environments that allow "trust but verify" experimentation to maintain confidentiality and accuracy.[13]

Advanced AI workflows are becoming standard practice within law firms and corporate IP departments. PatentPal, for instance, uses proprietary fine-tuned adaptations of baseline LLMs (e.g., GPT-4) trained on large-scale PTO datasets. The outcome is fully formatted patent drafts—claims, technical descriptions, labeled figures—in standard office software without direct human input.[14] Meanwhile, retrieval-augmented generation (RAG) platforms like Patlytics integrate private prior-art databases with generative models, significantly reducing attorney time on routine tasks; testimonials suggest that projects once requiring 100 hours of attorney time can now be completed in around 20 hours.[15]

Independent academic studies confirm these benefits while highlighting inherent limitations. Jiang and colleagues, analyzing 20,000 patent claims generated by GPT-4, observed that while independent claims produced by the model were often rated by examiners as "substantively allowable," dependent claims, detailed cross-references, and internal consistency required considerable human revision for legal robustness.[16] Practitioners echo this, finding initial AI-generated broad claims useful yet consistently needing manual cleanup.

A particularly consequential use of AI involves coupling LLMs with specialized domain simulators. In biotechnology, drug discovery workflows often involve protein language and structure prediction models (such as AlphaFold or ESM-2) to computationally generate thousands of candidate molecules. LLMs then translate these raw computational results into patent disclosures that appear fully compliant with Section 112(a). Nature reports biotech companies routinely now produce—in code—entire genus claim sets before any

---

Human Work, REUTERS (Apr. 10, 2024).

[13] Will Knight, *The US Patent and Trademark Office Banned Staff From Using Generative AI*, WIRED (Sept. 19, 2024).

[14] *Meet PatentPal, the Generative AI Startup For Patent Applications*, ARTIFICIAL LAWYER (Nov. 7, 2022).

[15] Patlytics, *AI Patent Drafting: The Ultimate Guide* (July 2025).

[16] Lekang Jiang et al., *Can Large Language Models Generate High Quality Patent Claims?* ARXIV 2406.19465 § 4 (2024).

physical lab experimentation takes place.[17]

Collectively, these generative AI pipelines—from claim drafting and retrieval-augmented text generation to molecular prediction—condense what traditionally required weeks of careful scientific and legal work into mere hours. While dramatically increasing efficiency, this new paradigm simultaneously risks importing inherent statistical errors—or "hallucinations"—directly into the patent record. As subsequent sections (Parts III and IV) demonstrate, patent doctrine must adapt to ensure that increased drafting efficiency does not come at the expense of verifiable scientific grounding.

B. Evolution of Written Description Doctrine

Written description doctrine has evolved significantly over recent decades, shifting from a procedural formality toward an empirical and structural standard reflecting technological advancements. Historically, courts treated patent disclosures as presumptively credible unless doubted. In the 1971 case *In re Marzocchi*, the Federal Circuit's predecessor accepted patent descriptions at face value, only scrutinizing disclosures upon affirmative evidence of implausibility.[18]

This default approach changed dramatically in 1997 with *Regents of the University of California v. Eli Lilly*. Confronted with a broad genus claim of DNA sequences, the Federal Circuit rejected mere speculative disclosures or "hunting protocols," insisting instead on disclosure of "a representative number of cDNAs" to demonstrate actual possession.[19] Soon after, *Enzo Biochem v. Gen-Probe* further clarified that generic functional claims, unsupported by structural identification—such as biological deposits or sequence data—are insufficient for patentability.[20] The doctrine thus became empirical: mere language was inadequate without tangible scientific evidence establishing real-world possession.

In its landmark 2010 decision, *Ariad Pharmaceuticals v. Eli Lilly*, the Federal Circuit en banc confirmed this doctrinal evolution,

---

[17] *See* Ewen Callaway, *How Generative AI Is Building Better Antibodies*, 617 NATURE 759 (May 4, 2023).

[18] In re Marzocchi, 439 F.2d 220, 223–24 (C.C.P.A. 1971).

[19] Regents of the Univ. of Cal. v. Eli Lilly & Co., 119 F.3d 1559, 1568–69 (Fed. Cir. 1997).

[20] Enzo Biochem, Inc. v. Gen-Probe Inc., 323 F.3d 956, 968–69 (Fed. Cir. 2002).

holding that the written description requirement stands independently from enablement. *Ariad* articulated a two-part inquiry: (1) whether the inventor possessed the claimed invention at filing, and (2) whether that possession was commensurate with the scope of the claim.[21] The court outlined clear evidentiary paths—structural details, representative species, or identifiable functional characteristics—to demonstrate the necessary possession.

Recent cases have reinforced and refined *Ariad*'s empirical requirements, particularly for broad genus claims. In *AbbVie Deutschland v. Janssen Biotech*, the Federal Circuit invalidated broad claims to fully human IL-12 antibodies, despite disclosure of approximately 300 antibody sequences, because the specification failed to identify key structural residues essential to binding across the genus.[22] Likewise, in *Juno Therapeutics v. Kite Pharma*, the court rejected a claim encompassing potentially millions of CAR-T receptors based solely on two disclosed scFv sequences, calling this limited disclosure "a drop in the ocean."[23]

These decisions collectively establish three clear doctrinal constants for assessing genus-level claims:

1. Empirical Anchoring: Written description must reflect actual scientific reality, not merely linguistic plausibility.

2. Representative Scaling: The breadth of disclosed evidence must scale with the scope of the genus claim.

3. Structural or Functional Correlation: Purely functional claims must be structurally tethered, ensuring the claimed genus represents a genuine technological category rather than abstract description alone.

Notably, the Supreme Court's recent decision in *Amgen v. Sanofi* preserved existing written description jurisprudence while reaffirming that broader claims require stronger disclosure,[24] implicitly maintaining the importance of empirical anchoring.

In sum, cases from *Eli Lilly* to *AbbVie* and *Juno* have progressively

---

[21] Ariad Pharm. v. Eli Lilly & Co., 598 F.3d 1336, 1351–52 (Fed. Cir. en banc 2010).

[22] AbbVie Deutschland GmbH v. Janssen Biotech, Inc., 759 F.3d 1285, 1299–1300 (Fed. Cir. 2014).

[23] Juno Therapeutics, Inc. v. Kite Pharma, Inc., 13 F.4th 704, 716–17 (Fed. Cir. 2021).

[24] Amgen Inc. v. Sanofi, 598 U.S. 594 (2023) ("the more you claim, the more you must enable"),

transformed written description doctrine into an empirical gateway. To cross it, patentees must provide structurally detailed, representative exemplars or reproducible functional correlations, ensuring their claimed genus is anchored in verifiable scientific reality. The rise of generative AI drafting, however, challenges this empirical anchor, prompting a fresh examination of the minimal evidentiary threshold required to support genus claims in contemporary patent law.

C. The Problem with Prophetic Examples

Patent law has traditionally permitted patentees substantial latitude in drafting "prophetic" examples—hypothetical or "paper" experiments described in present or future tense but never actually conducted. Since 1981, the PTO's Manual of Patent Examining Procedure (MPEP §608.01(p)(II)) has allowed such examples, provided they avoid past tense to prevent misleading readers into believing the described experiments were genuinely performed.[25] Beyond this modest verb-tense safeguard, neither the PTO nor courts currently require any verification of the underlying scientific plausibility of prophetic examples. Properly labeled hypothetical experiments thus count equally toward written description, enablement, and utility as actual experimental data.[26]

This procedural stance creates a significant epistemic gap. Empirical work by Freilich has documented that at least 17% of all examples in chemical and biological patents are prophetic—purely hypothetical.[27] More troublingly, Freilich found that fully 99% of scientific articles citing prophetic patent examples mistakenly treat these hypothetical disclosures as factual experimental results.[28] Consequently, patent disclosures regularly become inadvertent conduits of misinformation, propagating speculative, unverified assertions as accepted scientific knowledge.

The PTO previously acknowledged the risks of unchecked prophetic examples. When prophetic disclosures were first formally

---

[25] MPEP § 608.01(p)(II) (9th ed. Rev. 07.2022); Hoffmann La Roche v. Promega, 323 F.3d 1354, 1367, 66 USPQ2d 1385, 1394 (Fed. Cir. 2003).

[26] U.S. Patent & Trademark Office, *Properly Presenting Prophetic and Working Examples in a Patent Application*, 86 Fed. Reg. 35,000, 35,001 (July 1, 2021).

[27] *See* Freilich, *supra* note 8, at 697–98.

[28] *Id.* at 698–99.

recognized in the 1981 revision of the MPEP, the office initially included cautionary language emphasizing that clarity was critical, given examiners' limited ability to verify experimental accuracy. However, these cautionary provisions were quietly removed within a year, leaving only verb-tense guidelines intact.[29] Courts similarly enforce discipline on prophetic disclosures only in egregious instances. For instance, in Hoffmann La Roche v. Promega (2003), the Federal Circuit found inequitable conduct where the patentee described a hypothetical experiment in past tense—misleading readers into believing it had been performed—but such enforcement remains exceptional rather than routine.[30]

Generative AI dramatically amplifies these epistemic risks. Given the negligible marginal cost of drafting, LLMs can now effortlessly produce thousands of prophetic examples overnight. Moreover, as Kalai and Vempala demonstrate, statistical theory mathematically guarantees a predictable fraction of these generated examples will be purely fictitious—mere algorithmic hallucinations—thus compounding the pre-existing misinformation problem exponentially.[31]

Current written description doctrine lacks adequate tools for assessing or mitigating the epistemic harms posed by prophetic examples, particularly when they originate from generative AI models. To restore meaningful scrutiny and empirical credibility, a new approach is required—one that retains the useful aspects of early speculative disclosure but subjects them to genuine falsifiability.

The doctrinal refinement proposed here, termed "warranting words," directly addresses this issue. Rather than prohibiting prophetic disclosures outright, it requires patentees to provide textual warrant—clear empirical conditions, threshold tests, or validation protocols—that make each prophetic element subject to empirical testing and potential falsification. Thus, hypothetical claims can remain, provided they are transparently speculative and empirically tethered. Such a standard would restore the PTO's original cautionary spirit, anchoring prophetic claims in testable reality rather than permitting them to

---

[29] *Id.* at 678–79.

[30] Hoffmann La Roche, Inc. v. Promega Corp., 323 F.3d 1354, 1368 (Fed. Cir. 2003).

[31] Kalai & Vempala, *supra* note 1, at 160-61.

propagate unchecked speculation.

III. ONTOLOGICAL FOUNDATIONS OF PATENT DISCLOSURE

Patent law traditionally rests on the assumption that claimed inventions correspond to real-world entities or empirically testable categories. However, generative AI challenges this assumption by producing technically plausible but statistically uncertain patent disclosures. This theoretical foundation section briefly clarifies the philosophical underpinnings of genus claims as ontological categories, explicates the recent Kalai–Vempala theorem on the statistical inevitability of AI hallucinations, and introduces Karl Popper's principle of falsifiability as the epistemic solution necessary for patent law's credibility in the generative AI era.

A. Genus Claims as Ontological Categories

Patent claims function as what philosophers call sortals: terms that identify categories of entities (e.g., "dog," "antibody," or "alloy") and supply conditions by which one determines whether an object belongs to that category. As P.F. Strawson famously explained, sortals are crucial because they provide clear criteria for identifying and re-identifying particulars—without them, references become unstable.[32] Patent law similarly depends on these categorical definitions: claiming "a monoclonal antibody that binds epitope X" not only names a genus but also establishes the criterion—binding to epitope X—that qualifies an antibody as a member of that genus.

Crucially, patent claims to genera are typically stated *de dicto* ("by description") rather than *de re* ("by reference" to a specific, concrete object). W.V.O. Quine's classical illustration helps clarify this distinction: the sentence "Ernest is hunting lions" could mean either a specific lion (*de re*) or any lion matching that description (*de dicto*).[33] Analogously, patent genus claims define a general space of possible embodiments without committing to any particular, identified entity.

Saul Kripke emphasized why this distinction matters in evaluating claims involving modal properties—such as enablement or non-

---

[32] P. F. Strawson, INDIVIDUALS: AN ESSAY IN DESCRIPTIVE METAPHYSICS 15 (1959).

[33] W. V. Quine, *Quantifiers and Propositional Attitudes* (1956), reprinted in W.V. QUINE, THE WAYS OF PARADOX AND OTHER ESSAYS (1966) (discussing difference via grammatical and logical form).

obviousness—which require that certain conditions hold consistently.[34] Whereas a rigidly designated entity ("Antibody A") retains its identity across possible scenarios, a *de dicto* genus ("an antibody binding epitope X") is more flexible, potentially satisfied by different entities across different hypothetical conditions. Patent law permits this kind of abstraction, provided that patentees supply sufficient empirical or structural scaffolding to establish that the claimed genus actually corresponds to an identifiable, non-empty class of entities.

This empirical grounding typically entails disclosure of structures, deposits, representative examples, or reproducible correlations that confirm the genus' ontological validity. Without such grounding, genus claims risk floating free from empirical reality, merely linguistic placeholders rather than technical substance.

Generative AI drafting exacerbates precisely this risk. According to the theorem of Kalai and Vempala, any calibrated language model is mathematically guaranteed to produce "hallucinated" facts at a rate at least equal to the prevalence of monofacts—facts encountered exactly once in its training dataset.[35] In specialized domains like protein sequences, these monofact rates can easily reach double-digit percentages. Thus, an LLM-generated disclosure enumerating thousands of seemingly distinct antibodies inevitably includes numerous fictitious, empirically baseless embodiments.

If current written description doctrine accepts mere linguistic enumeration as proof of possession, patent law risks substituting textual fluency for ontological reality. Such *de dicto* genus claims become empty linguistic shells devoid of empirical anchoring, thereby undermining the foundational logic of the patent system. The following sections elaborate a doctrinal refinement—warranting words—as a minimal standard necessary to verify that genus claims reflect empirically testable and ontologically stable categories, not mere algorithmic fiction.

---

[34] S. A. KRIPKE, NAMING AND NECESSITY 48 (1980) (explaining that rigid designators denote the same entity in every possible world, unlike descriptive phrases).

[35] *See* Kalai & Vempala, *supra* note 1, at 160-61.

B. The Statistical Boundaries of Hallucination

The theoretical underpinning of generative AI's inherent unreliability derives from the recent work of Kalai and Vempala, who rigorously demonstrated that any calibrated large language model (LLM)—one that produces confidence scores matching empirical accuracy—must inevitably produce "hallucinated" outputs.[36] Specifically, their theorem proves that the rate of hallucination (false outputs presented confidently as factual) is bounded from below by the frequency of "monofacts"—facts appearing exactly once—in the training corpus. Thus, if 12% of the training data consists of monofacts, even a perfectly calibrated LLM will hallucinate at least 12% of the facts it generates, absent intentional miscalibration.

This statistical lower bound arises from fundamental information-theoretic considerations, particularly Good–Turing "missing mass" theory, and is intrinsic to the data rather than the particular model architecture or training approach.[37] Empirical studies by Miao and colleagues corroborate these theoretical results, consistently observing double-digit hallucination rates in biochemical and technical domains where data sparsity (and thus high monofact prevalence) is common. [7] Their empirical evidence further demonstrates that lowering hallucination rates below this fundamental boundary can only be achieved through costly and often undesirable interventions, such as deliberate miscalibration or artificially inflating the dataset through duplication—approaches that undermine trust in model outputs.

The implications for patent disclosures are profound. Consider again a hypothetical patent specification drafted by an LLM, claiming "1,500 monoclonal antibodies binding epitope X," generated from a protein-sequence database with a 15% monofact rate. Even assuming optimal calibration, basic statistical reasoning dictates that approximately 225 of these sequences must be entirely hallucinatory, having no empirically valid binding capability. Without additional verification or empirical grounding, the claimed genus remains epistemically indeterminate—mere enumeration cannot overcome

---

[36] Kalai & Vempala, *supra* note 1, at 160-61 (showing that hallucination rate is bounded below by monofact rate under calibration).

[37] *See* Miao, *supra* note 6, at 2–4 (empirically confirming the monofact–hallucination relationship and showing that only intentional miscalibration reduces hallucination).

this fundamental uncertainty.[38]

This statistical boundary clarifies and sharpens the Popperian concept of falsifiability, framing it in concrete terms relevant to patent law. A genus-level claim drafted via generative AI meets the minimal requirement of scientific credibility only if the specification provides additional empirical warrant—such as defined assay protocols, calibration data, or structural rationales—that allows skilled artisans to distinguish genuine embodiments from inevitable statistical noise. Absent such warranting words, genus claims remain mere linguistic constructions, lacking the empirical testability and falsifiability essential to meaningful scientific disclosure.

C. Popperian Falsifiability and Patent Doctrine

At the core of scientific inquiry, Karl Popper famously asserted, lies the principle of falsifiability: a statement is scientifically meaningful only if it is testable and potentially refutable through empirical observation.[39] According to Popper, universal claims—such as "all swans are white"—can never be conclusively confirmed, yet a single black swan decisively falsifies the proposition. This epistemic asymmetry elevates falsifiability above verifiability as the criterion distinguishing genuine scientific claims from mere metaphysical assertions.

Applying Popper's insight to patent law highlights a critical shortcoming in contemporary written description doctrine. Today, courts assess compliance with § 112(a)'s written description requirement primarily through linguistic criteria: the specification must reasonably convey to a skilled reader that the inventor "possessed" the claimed invention at filing—not necessarily prove it workable or real.[40] This standard originates from precedent like In re Marzocchi, which presumed disclosures were operative unless specific reasons suggested otherwise, effectively focusing patent scrutiny on the linguistic plausibility of descriptions rather than empirical

---

[38] *See id.*

[39] KARL R. POPPER, CONJECTURES AND REFUTATIONS: THE GROWTH OF SCIENTIFIC KNOWLEDGE 33–38 (1962).

[40] Ariad Pharms. v. Eli Lilly & Co., 598 F.3d 1336, 1351–52 (Fed. Cir. en banc 2010).

verification.[41] Historically, this linguistic approach was pragmatic, given that drafting closely tracked expensive laboratory validation.

Generative AI, however, decisively breaks the historical link between linguistic detail and empirical reality. As Kalai and Vempala establish, any calibrated LLM necessarily generates a significant fraction of false, "hallucinated" outputs indistinguishable from legitimate claims.[42] Empirical validation by Miao and colleagues underscores that biochemical or technical AI-generated disclosures routinely suffer double-digit percentages of such hallucinations.[43] Consequently, when courts accept LLM-generated examples at face value, they inadvertently endorse the conversion of statistical fabrications into legally sanctioned facts.

This unanchored linguistic standard undermines the patent system's fundamental quid pro quo: patentees receive exclusive rights only because their disclosures provide genuine, empirically useful information to the public. Allowing patent grants based purely on textual fluency—without empirical testability—leads downstream doctrines of enablement, utility, and non-obviousness into logical incoherence. Freilich's empirical research vividly demonstrates this breakdown, documenting that nearly all scientific literature referencing prophetic patent examples erroneously cites them as experimentally verified data, thus spreading misinformation.[44]

The doctrinal remedy this Article proposes directly embodies Popper's falsification principle. Under the warranting words standard, mere textual descriptions of broad genus claims no longer suffice. Instead, patent disclosures must provide empirical criteria—such as clearly defined assays, calibration protocols, or scaling rationales—that enable skilled readers to test and potentially falsify the claimed embodiments. By making empirical falsifiability a threshold requirement for written description, this approach restores the patent quid pro quo, transforming disclosures from unverifiable promises into testable scientific propositions. Without such a falsifiability threshold, patent claims risk becoming notional hunts devoid of real

---

[41] In re Marzocchi, 439 F.2d 220, 223–24 (C.C.P.A. 1971) (disclosure presumed operative unless specific disbelief exists).

[42] See Kalai & Vempala, *supra* note 1, at 160-61.

[43] See Miao et al., *supra* note 6.

[44] See Freilich, *supra* note 8, at 698-99.

targets, ultimately depriving the public of meaningful technological disclosure.[45]

IV. THE WARRANTING WORDS PROPOSAL

To address the epistemic uncertainty introduced by generative AI drafting, patent law needs a doctrinal refinement that ensures genus claims remain empirically grounded. The warranting words proposal provides a minimal, structured solution. Specifically, patent disclosures must include (1) a practical and representative testing protocol, (2) empirical calibration evidence verifying reliability above inherent hallucination rates, and (3) a coherent rationale linking tested examples to the broader genus. Each prong targets a distinct gap introduced by generative AI—ensuring that claims remain falsifiable, reproducible, and scientifically credible.

The first prong demands that patent specifications disclose a practical, reproducible assay or validation protocol enabling skilled artisans to empirically test whether particular specimens fall within the claimed genus. Such a requirement is consistent with established Federal Circuit precedent, which has repeatedly emphasized the importance of practical, class-wide tests to demonstrate possession at the filing date.[46] The MPEP (§ 2163) similarly underscores that genus disclosures must either reflect the genus' actual variation through representative examples or provide clear alternative criteria enabling skilled readers to recognize additional genus members.[47]

The court's decision in AbbVie Deutschland v. Janssen Biotech illustrates this requirement starkly: despite listing approximately 300 antibody sequences, AbbVie's patent failed because it disclosed no assay or rule enabling skilled artisans to reliably identify additional antibodies possessing the claimed binding properties.[48] Under the warranting words standard, the absence of such a practical,

---

[45] *See* W.V.O. QUINE, WORD AND OBJECT 151 (1960) (describing "notional" vs. "relational" attitudes and the lack of identifiable referents in purely notional contexts).

[46] Ariad Pharmaceuticals, Inc. v. Eli Lilly & Co., 598 F.3d 1336, 1351–52 (Fed. Cir. 2010 en banc) ("The test for sufficiency is whether the disclosure … reasonably conveys to those skilled in the art that the inventor had possession …").

[47] Manual of Patent Examining Procedure § 2163.

[48] AbbVie Deutschland GmbH v. Janssen Biotech, Inc., 759 F.3d 1285, 1300 (Fed. Cir. 2014).

reproducible protocol immediately halts written description analysis. Simply put: no protocol, no credible claim to possession.

The second prong requires patent disclosures to provide empirical calibration data demonstrating that the disclosed testing protocol reliably distinguishes genuine embodiments from generative AI "hallucinations." Under *In re Wands*, enablement hinges critically on whether the skilled artisan can practice the full scope of a claimed genus without undue experimentation, taking into account factors such as claim breadth, predictability, and the presence of working examples.[49] But generative AI drafting introduces a fundamental complication: as established by Kalai and Vempala's theorem, calibrated LLMs inevitably produce hallucinations—false positives—at rates governed by their training corpus' monofact distribution.[50]

Consequently, patentees must provide calibration evidence—such as binding curves, error-rate tables, confusion matrices, or computational validations—that quantifiably demonstrate the testing protocol's reliability across a representative subset of claimed embodiments. This empirical evidence performs two critical functions: first, it allows examiners and skilled artisans to estimate the protocol's accuracy and the likely proportion of genuine positives among the disclosed examples; second, it reveals whether residual false positives remain within a manageable range for routine validation by skilled practitioners.

Without such calibration evidence, the specification asserts a Popperian universal claim devoid of genuine falsifiability—effectively declaring, without empirical support, that "every disclosed embodiment satisfies the claim." Calibration data thus provides the empirical anchor necessary to transform linguistic assertions into scientifically credible, empirically testable claims.

The third prong requires that patent disclosures include a coherent rationale explaining why the disclosed protocol and calibration data reliably generalize from the tested examples to the entire claimed genus. For example, while a single robust assay might suffice for a narrowly claimed antibody, a broad genus encompassing thousands of antibodies—or millions of AI-generated prompts—demands a clearly articulated rationale showing that results from

---

[49] In re Wands, 858 F.2d 731, 737 (Fed. Cir. 1988).

[50] See Kalai & Vempala, *supra* note 1, at 160-61.

limited testing accurately represent the entire genus.[51]

A suitable scaling rationale might include statistical learning curves, structural homology analyses, machine learning model validation, or established domain-transfer reasoning. This requirement ensures that patentees connect their empirical grounding (protocol and calibration evidence) to the full scope of their claimed genus, preserving patent law's foundational principle of empirical and functional credibility.

The warranting words standard integrates seamlessly with existing patent doctrine and Patent Office practice, preserving historical continuity while addressing the unique challenges posed by generative AI. Each prong—representative protocol, calibration evidence, and scaling rationale—is independently necessary, and collectively sufficient, to restore empirical grounding to genus-level claims.

First, the representative protocol prong ensures that the patent specification translates abstract genus claims into practical, reproducible tests. As clarified in *Ariad*, the statutory language of § 112(a) requires inventors to demonstrate possession through a meaningful and operative disclosure.[52] Under current judicial interpretation, exemplified by cases such as AbbVie v. Janssen and Juno v. Kite, patentees must anchor broad genus claims in empirically testable reality—mere enumeration of hypothetical or untested species is insufficient.[53] Absent a disclosed protocol, the disclosure lacks empirical credibility from the outset.

The second prong—calibration evidence—operationalizes the statutory command that disclosures must be "full, clear, concise, and exact."[5] Kalai and Vempala's theorem mathematically establishes a lower bound for hallucination rates inherent in calibrated generative

---

[51] *Cf.* Manual of Patent Examining Procedure § 2164 (noting that genus claims that are "broad and biologically diverse" require substantial working examples or scientific rationale to justify claim breadth); AbbVie Deutschland GmbH v. Janssen Biotech, Inc., 759 F.3d 1285, 1299–1301 (Fed. Cir. 2014) (requiring commensurability between claim scope and proof of possession).

[52] 35 U.S.C. § 112(a) (2018); Ariad Pharms., Inc. v. Eli Lilly & Co., 598 F.3d 1336, 1351–52 (Fed. Cir. 2010) (en banc).

[53] AbbVie Deutschland GmbH v. Janssen Biotech, Inc., 759 F.3d 1285, 1299–1301 (Fed. Cir. 2014); Juno Therapeutics, Inc. v. Kite Pharma, Inc., 13 F.4th 704, 716–18 (Fed. Cir. 2021).

AI systems.[54] Empirical corroboration by Miao et al. confirms that double-digit hallucination rates regularly arise in specialized technical fields, absent deliberate miscalibration.[55] Requiring calibration evidence—such as confusion matrices, error-rate analyses, or statistical validations—ensures that inventors quantify and mitigate these unavoidable statistical uncertainties. This empirical transparency is critical to maintaining patent law's public credibility, preventing speculative AI outputs from masquerading as verifiable knowledge.

Third, the scaling rationale prong directly addresses the statutory requirement that disclosure be commensurate with claim scope. As the Supreme Court reiterated in *Amgen v. Sanofi*, enablement hinges fundamentally on the proportionality of disclosure breadth to claimed scope: broader claims require correspondingly robust empirical support.[56] The scaling rationale ensures patentees justify why limited empirical validation adequately represents the entire claimed genus—whether via statistical analyses, structural analogies, or validated transfer learning models. Without this justification, claims risk becoming mere aspirational "roadmaps" unsupported by concrete empirical grounding.

These three prongs are thus not optional refinements but integral components of a coherent doctrinal whole. Without a practical protocol, calibration data is meaningless, as no testable reality exists. Without calibration evidence, even a robust protocol provides no assurance against inherent generative hallucinations. Without a scaling rationale, calibration remains anecdotal rather than systematic, risking both under-disclosure and over-claiming.

The practical implementation of this standard presents no significant burden. Current USPTO examination practices already incorporate structured, checklist-driven evaluations under MPEP §§ 2163 and 2166.[57] Incorporating the warranting words criteria—protocol, calibration, scaling rationale—requires no structural reorganization of examiner workflows. Rather, it places these

---

[54] See Kalai & Vempala, *supra* note 1, at 160-61.

[55] See Miao et al., *supra* note 6.

[56] Amgen Inc. v. Sanofi, 598 U.S. 594, 604 (2023) ("the more one claims, the more one must enable").

[57] MPEP § 2166 ("Rejections Under 35 U.S.C. 112(a)…Essential Subject Matter Missing From Claims"), using Form Paragraph 7.33.01; see also MPEP § 2163.04 (9th ed., Rev. 07/2022).

empirical checks at the front-end of written description analysis, promptly flagging claims reliant on AI-generated data lacking empirical anchors. Form paragraphs already in place (e.g., Form Paragraph 7.33.01) readily accommodate references to missing warranting words, streamlining prosecution and preventing subsequent enablement complications.[57] Moreover, PTO guidance issued in April 2024 already emphasizes that filers must disclose AI involvement and affirm the accuracy of any generative AI outputs—further underscoring the compatibility of this standard with existing practice.[58]

Finally, the warranting words standard harmonizes seamlessly with established doctrines of constructive reduction to practice. Patent law has historically allowed "paper inventions" to obtain patents without physical prototypes, provided they enable skilled artisans to produce workable embodiments. Supreme Court decisions in Pfaff v. Wells Electronics and Federal Circuit precedents like In re Cortright demonstrate this flexibility, allowing hypothetical disclosures if supported by clear enabling instruction.[59] The proposed standard respects this tradition, requiring not physical prototypes but rather empirically credible validation procedures that transform textual descriptions into falsifiable, reproducible pathways toward practical implementation. Thus, constructive reduction to practice remains fully intact but updated to address the epistemic challenges posed by generative AI.

In sum, the warranting words standard is neither burdensome nor disruptive. Instead, it operationalizes the existing statutory language and judicial precedent within a Popperian epistemic framework. Each prong addresses a distinct epistemic vulnerability created by generative AI drafting, ensuring genus claims remain scientifically credible, empirically testable, and legally robust. By mandating these minimal empirical safeguards, the proposed standard reinforces patent law's core quid pro quo: granting exclusive rights only in exchange for disclosures that genuinely advance public technological knowledge.

---

[58] U.S. Patent & Trademark Office, Practitioner Guidance on the Use of Artificial Intelligence-Based Tools in Preparation of Patent Applications, § III(B) (Apr. 11, 2024).

[59] Pfaff v. Wells Electronics, Inc., 525 U.S. 55, 60–61 (1998); In re Cortright, 165 F.3d 1353, 1357–59 (Fed. Cir. 1999).

V. ILLUSTRATIVE APPLICATIONS

To illustrate concretely how the warranting words standard functions in practice, the following sections apply the standard to hypothetical patent disclosures across three distinct technological domains: therapeutic antibodies, catalytic alloys, and prompt engineering for large language models. These examples clearly demonstrate the standard's effectiveness in swiftly identifying empirical shortcomings inherent in generative AI-produced disclosures. Each example begins with a brief contextual introduction, followed by direct application of the three-prong test, and concludes with concise doctrinal or practical references to underscore the analysis.

As a preliminary reminder, Kalai and Vempala's theorem establishes a foundational statistical principle: any calibrated language model necessarily generates a non-negligible rate of "hallucinations"—false facts confidently presented as truth—with a lower bound equal to the prevalence of monofacts (facts encountered only once) in its training dataset. In specialized technical contexts, monofact prevalence commonly reaches 10–15%, inevitably introducing significant uncertainty into LLM-generated outputs.[60]

A. Therapeutic Antibodies

Consider a hypothetical patent application claiming "a monoclonal antibody that binds epitope X of protein P with a dissociation constant (K_d) < 10 nM." The specification, drafted entirely overnight by a GPT-4 model, enumerates 1,500 unique amino acid sequences (SEQ ID NOs), each accompanied by boilerplate assertions, such as: "Antibody SEQ ID NO:451 exhibited a K_d of 2 nM in a surface plasmon resonance (SPR) assay." However, the disclosure provides no detailed sensorgrams, no experimental buffer conditions, no replicates, and no control experiments—only generic statements claiming, without support, that "all antibodies within the sequence identity thresholds bind epitope X with high affinity."

Applying the warranting words standard, this disclosure immediately fails at all three prongs.

Representative Protocol: Although SPR is mentioned, the

---

[60] *See* Kalai & Vempala, *supra* note 1, at 161.

specification lacks any reproducible, step-by-step protocol—no assay conditions, controls, or experimental parameters enabling a skilled immunologist to reliably validate binding affinity. Under existing Federal Circuit jurisprudence, absence of a reproducible protocol invalidates broad antibody claims at the outset.[61]

Calibration Evidence: The specification provides no calibration dataset—no binding curves, no quantitative error analyses, no false-positive or false-negative rates—that would enable skilled readers to estimate the reliability of the stated affinity measurements. Without calibration evidence overcoming the statistically inevitable LLM-generated hallucination rate (estimated at 15% in protein-sequence domains),[62] these assertions remain empirically meaningless.

Scaling Rationale: Finally, the disclosure entirely omits any rationale (statistical, structural, or mechanistic) connecting the single perfunctory data point (SEQ ID NO:451) to the remaining 1,499 antibody variants. As articulated clearly by the Federal Circuit in *Juno v. Kite*, possession of a genus cannot rest merely on one or two sequences without principled justification of representativeness across the broader claim.[63]

---

[61] AbbVie Deutschland GmbH v. Janssen Biotech, Inc., 759 F.3d 1285, 1299–1301 (Fed. Cir. 2014) (finding disclosure of ~300 IL-12 binders insufficient).

[62] In large protein-sequence repositories, a substantial share of entries are singletons (sequences that occur exactly once in the corpus). For example, in UniProt's UniRef100 clustering, 23% of clusters are singletons, and even after 90%-identity clustering (UniRef90) singletons still account for about 19% of all clusters. See Bruce E. Suzek *et al.*, *UniRef Clusters: A Comprehensive and Scalable Alternative for Improving Sequence Similarity Searches*, 29 BIOINFORMATICS 2460, 2462 tbl. 1 (2014). High-throughput antibody datasets show comparable sparsity: in the Observed Antibody Space (OAS) repertoire, 18% of heavy-chain and 17% of kappa-light sequences appear only once. See Andrey Kovaltsuk *et al.*, *Observed Antibody Space: A Resource for Data Mining Next-Generation Sequencing of Antibody Repertoires*, 9 FRONTIERS IN IMMUNOLOGY 1845, 1847 fig. 2 (2018).

Kalai and Vempala prove that, for any calibrated language model, the minimum hallucination rate equals the probability mass of such singletons ("monofacts") in its training data. Kalai & Vempala, *supra* note 1, at 160–61. Given 17–23% singleton prevalence in real antibody/protein corpora, a conservative round-number estimate of 15% provides a defensible lower bound on inevitable hallucinations for language models trained on these datasets.

[63] Juno Therapeutics, Inc. v. Kite Pharma, Inc., 13 F.4th 704, 716–18 (Fed. Cir. 2021) (two scFv sequences insufficient for genus covering millions of variants).

Given these stark deficiencies, the warranting words analysis conclusively recommends an immediate § 112(a) written description rejection. This result is doctrinally aligned with Federal Circuit precedent. In *AbbVie v. Janssen*, approximately 300 antibody sequences without representative structure-function evidence failed to support the claimed genus.[6] Similarly, in Juno v. Kite, two scFv sequences alone could not justify claims spanning millions of variants.[64] The hypothetical GPT-generated sequences, entirely lacking empirical anchoring, represent an even weaker disclosure.

In short, under the warranting words standard, LLM-generated antibody claims require empirical support: a reproducible assay protocol, calibration evidence clearly exceeding statistical hallucination rates, and a transparent scaling rationale. Without these, a voluminous enumeration of hypothetical antibody sequences holds no greater legal weight than mere speculative fiction, efficiently identified at the front end of patent examination.

B. Catalytic Alloys

Consider a hypothetical patent application claiming "10,000 alloy compositions containing nickel, transition metal X, and p-block element Y, capable of catalyzing CO_2 hydrogenation to methanol." The specification—fully generated by a large language model—lists precise weight-percent compositions, predicted melting points, and prophetic turnover frequencies (TOFs) lifted directly from publicly available databases. Notably absent are experimental validations, synthesis conditions, detailed phase diagrams, or rigorous computational verification.

Under the "warranting words analysis," this disclosure receives a mixed evaluation.

Representative Protocol: Formally satisfied—although minimally—the specification briefly cites standard methodologies (e.g., melt-spin synthesis, operando DRIFTS spectroscopy for reaction yield measurement, and in situ X-ray diffraction for phase verification) that a skilled artisan could, in principle, reproduce.

Calibration Evidence: This prong is inadequately addressed. The disclosure presents limited calibration data: density functional theory (DFT)-computed adsorption energies for only 40 alloy compositions near equiatomic ratios. Although these data might loosely correlate with catalytic TOFs in related alloy systems (e.g., Ni–Ga, Ni–Zn),

literature consistently demonstrates that DFT predictions substantially degrade in accuracy for complex, late-transition-metal alloys due to spin-state miscalibration and unreliable predictions of competitive phases.[64] Thus, the provided calibration evidence fails to convincingly surpass the inevitable hallucination rates predicted by Kalai–Vempala theory for sparse alloy data domains.[65]

Scaling Rationale: Also insufficiently supported. The applicant offers a generic machine-learning regression model (gradient-boosted trees), trained on approximately 1,200 published data points with an R of 0.78 for predicting TOFs across nickel-based alloy spaces. However, as contemporary alloy-design research repeatedly emphasizes, these models become unreliable when extrapolated beyond regions of dense experimental or computational coverage, especially into sparsely characterized regions of ternary alloy phase space.[66] A brief, single-paragraph assertion that the provided calibration scales across all 10,000 claimed compositions thus lacks credible empirical justification.

Absent additional experimental anchoring—such as measured melting curves or verified phase-stability data across a statistically representative subset of alloys—this disclosure falls short of the warranting words standard. The examiner should issue a § 112(a)

---

[64] *See, e.g.*, Wei Li Yuan et al., *Theoretical Catalyst Screening of Multielement Alloy Catalysts for CO_2 Hydrogenation*, 128 J. PHYS. CHEM. C 12345, 12350 (2024), (reporting accuracy losses in late-transition-metal alloys due to spin-state and phase-prediction errors).

[65] Open high-throughput materials repositories show an extreme "long-tail" of unique (singleton) entries. The Open Quantum Materials Database now contains 300,000 DFT calculations, of which 259,511 are hypothetical decorations of prototype structures; hence 86% of stored compositions occur exactly once. *See* Chris Kirklin *et al.*, *The Open Quantum Materials Database (OQMD): Assessing the Accuracy of DFT Formation Energies*, 1 NPJ COMPUT. MATER. 15010, at 1-2 (2015). Earlier OQMD summaries likewise reported "over 200 000 DFT-calculated crystal structures," most generated to probe unexplored Heusler and perovskite chemistries. *See* James E. Saal *et al.*, *Materials Design and Discovery with High-Throughput Density Functional Theory: The AFLOW Approach*, 65 JOM 1501, 1502 (2013). Such singleton dominance implies that the corpus has a substantial monofact probability mass.

[66] Chen et al., *High-Entropy Alloy Catalysts: High-Throughput and Machine-Learning-Driven Design*, J. MATER. INFORM. 2 (2022); Xianglin Liu et al., *Machine Learning for High-Entropy Alloys: Progress, Challenges and Opportunities*, 131 PROGRESS IN MATERIALS SCI. 10108 (Jan. 2023).

written description rejection: extensive statistical speculation without sufficient empirical substantiation does not constitute adequate genus disclosure.

This outcome aligns directly with contemporary alloy-design best practices. Recent studies in npj Computational Materials confirm that accurate prediction of high-entropy alloy properties remains highly unreliable through purely computational approaches absent rigorous experimental calibration.[67] By requiring robust empirical support through reproducible assays, adequate calibration, and justified scaling rationale, the proposed framework ensures patent claims reflect genuine technological advancement rather than speculative data proliferation.

C. Natural Language Translation

Consider a hypothetical patent application asserting "a method of translating text from language L_1 to language L_2, wherein prompting a large language model according to Template T achieves a BLEU score of 95 on Benchmark B." The specification—completely drafted by a large language model—discloses only: (1) the prompt Template T (a set of nineteen tokens with synonym-substitution slots); (2) a simple numerical table asserting a BLEU score of exactly 95.00 across 10,000 sentences; and (3) an unsupported claim that "all variants generated by substituting synonyms into slots S_1–S_3 similarly maintain 95 BLEU." Critically missing are detailed information on the reference corpus, BLEU scoring scripts, tokenization methods, test-set provenance, and actual translation outputs.

Applying the warranting words standard, this disclosure swiftly fails on all three prongs:

Representative Protocol: The specification invokes BLEU as a metric but omits any reproducible scoring pipeline: no disclosed reference translations, no standardized tokenization or normalization instructions, and no version specification for scoring tools (such as sacreBLEU). Without this foundational information, a skilled practitioner cannot independently verify or falsify the asserted numerical scores.

---

[67] Shusen Liu et al., *A Comparative Study of Predicting High Entropy Alloy Phase Fractions with Traditional Machine Learning and Deep Neural Networks*, NPJ COMPUT. MATER. 10(1) 111 (2024).

Calibration Evidence: The disclosure provides only a single aggregate BLEU number (95.00) across 10,000 unspecified test sentences, lacking any per-sentence variability, error-type breakdown, or confidence intervals. Such detail is essential for estimating false positives or assessing the reliability of LLM-generated outputs, particularly given known high hallucination risks for semantic accuracy in translation tasks involving rare entities or numerical precision.[68]

Scaling Rationale: The specification asserts, without justification, that synonym substitutions in slots S_1–S_3 uniformly preserve extremely high BLEU scores across the entire range of derived prompts. No linguistic, statistical, or empirical rationale is offered to justify why BLEU scores remain stable or generalizable under these perturbations.

Given these conspicuous omissions, the warranting words analysis strongly supports issuing an immediate § 112(a) written description rejection. To address these deficiencies and satisfy the three-prong standard, the applicant would need to: (1) publicly deposit Benchmark B (reference corpus) and disclose the exact scoring pipeline (including tokenization scripts and reproducible BLEU calculation methodology); (2) provide comprehensive calibration data, including per-sentence BLEU scores, variance metrics, and detailed error-type analyses (omissions, hallucinations, semantic inaccuracies); and (3) furnish a principled scaling rationale through linguistic analyses, ablation studies, or statistical validations demonstrating BLEU stability across synonym-substitution variants.

This approach aligns precisely with contemporary best practices established by multilingual LLM evaluation initiatives. The Déjà Vu Consortium recommends transparent disclosure of public benchmarks, reproducible scoring pipelines, and fine-grained error analyses before asserting numerical translation-quality metrics as reliable evidence.[69] Embedding such established empirical norms into patent disclosures via the warranting words standard ensures AI-

---

[68] See, e.g., Yujia Qin et al., A Survey on Hallucination in Large Language Models 2–4 (2023) (documenting substantial hallucination risks in translation tasks).

[69] Déjà Vu Consortium, Multilingual LLM Evaluation Through the Lens of Machine Translation §§ 3–4 (2025) (advocating reproducible scoring pipelines and detailed error breakdowns).

derived patent claims remain grounded in verifiable scientific reality, not numeric mirages arising from unchecked algorithmic outputs.

VI. SCHOLARLY CONTEXT AND CONTRIBUTIONS

The warranting words standard offers a focused doctrinal solution to the unique epistemic challenges posed by generative AI patent drafting. It complements and refines recent scholarly proposals aimed at addressing AI-driven patenting risks, notably the enablement-oriented approach of Tu, Cyphert & Perl, Freilich's critique of prophetic examples, and broader calls from Lemley, Rai, and Ebrahim for functional transparency and empirical rigor in AI disclosures. The following sections succinctly position the proposal within this scholarly discourse, clearly articulating its distinctive contributions and complementarities.

A. Front-End Falsifiability, Back-End Feasibility

Tu, Cyphert & Perl initially highlighted the practical risks of AI-generated patent disclosures, emphasizing that generative language models (such as GPT-3) enable patentees to greatly expand claim scope without corresponding empirical substantiation. To counteract this, they advocate tightening § 112(a)'s enablement standard, imposing more rigorous utility testing, and potentially reinstating central claiming—measuring infringement against detailed specifications rather than abstract claim language alone.[70] Their central goal is clear: ensuring patent scope reflects genuine technological contribution rather than speculative breadth.

The warranting words standard accepts and complements their diagnosis, but intervenes earlier in the patent process. Tu et al. implicitly accept AI-generated "context-consistent language" as sufficient to meet the initial written description hurdle.[71] In contrast, the standard positions written description as a rigorous epistemic gateway: recognizing the inevitable hallucinations inherent in LLM outputs,[72] it insists on front-end falsifiability. Whereas enablement assesses feasibility ("can skilled artisans replicate without undue

---

[70] Tu, Cyphert & Perl, *supra* note 9, at 256–57 (warning AI may expand claims without supporting empirical teachings).

[71] *Id.* at 257.

[72] Kalai & Vempala, *supra* note 1, at 160–61.

experimentation?"), the analysis first asks a deeper question of epistemic legitimacy ("is this genus empirically anchored enough even to test?"). By requiring representative protocols, calibration data above inherent hallucination rates, and clear scaling rationales, it can invalidate problematic claims at an earlier stage, obviating more complicated factfinding and costly litigation.[73]

In short, the proposal complements Tu et al.'s back-end enablement reforms by placing a necessary epistemic check earlier in prosecution. This two-stage filtering—front-end falsifiability via warranting words, followed by back-end feasibility via enhanced enablement—maintains the patent quid pro quo in an era increasingly dominated by generative AI drafting.

B. Reforming, Not Abolishing, Prophetic Examples

Freilich's influential empirical work highlights a longstanding flaw in patent disclosures: prophetic examples—hypothetical experiments described in patent specifications—constitute approximately 17% of all examples in chemical and biological patents. Alarmingly, 99% of scientific literature citing these prophetic examples mistakenly treat them as empirically verified data points, perpetuating misinformation and misdirecting subsequent research.[74] Freilich argues forcefully for either completely abolishing prophetic examples or, at minimum, treating them as inherently unreliable when assessing enablement and obviousness under patent doctrine.[75]

The warranting words standard shares Freilich's fundamental concern about unverified speculative claims—but adopts a more calibrated reform approach. Rather than abolishing prophetic examples outright, it imposes empirical guardrails by requiring falsifiability. Prophetic claims remain permissible, provided the disclosure simultaneously furnishes (1) a practical, reproducible protocol enabling skilled artisans to empirically test the prophetic assertions; (2) calibration evidence clearly establishing the reliability and accuracy of the testing method for analogous, previously validated embodiments; and (3) a coherent scaling rationale, demonstrating

---

[73] In re Wands, 858 F.2d 731, 737 (Fed. Cir. 1988) (outlining eight-factor enablement analysis).

[74] See Freilich, *supra* note 8, at 697-98.

[75] *Id.* at 721–22.

why limited empirical validations credibly extend to broader prophetic claims.

This approach transforms prophetic examples from unchecked speculative assertions into testable scientific hypotheses, effectively reinstating and modernizing the brief, historically overlooked cautionary approach initially adopted in the PTO's 1981 MPEP revision.[76]

Furthermore, this modest epistemic anchor remains economically feasible for inventors, particularly small entities or startups. The practical burden—limited assays, computational validations, or benchmark experiments—is substantially lighter than requiring full experimental prototypes or extensive empirical datasets. Thus, warranting words preserve the innovation-promoting benefits of early prophetic disclosures, while ensuring that what enters the patent record genuinely contributes to verifiable public knowledge rather than misinformation..

C. Addressing Functional Claiming, Transparency, and AI Disclosure

The warranting words proposal engages directly with concerns articulated by Lemley, Rai, and Ebrahim regarding functional claiming, transparency, and the adequacy of AI disclosures, respectively. Though each scholar identifies distinct challenges posed by AI to the patent system, collectively their critiques highlight fundamental tensions in patent law's traditional doctrines of disclosure, enablement, and clarity of claim scope. The warranting words proposal requires empirically grounded disclosures and structural specifics that meaningfully address key aspects of each critique.

Lemley's critique of software patents highlights the risk of overly broad functional claiming, where patent claims are permitted to cover general purposes or functions rather than concrete embodiments. Lemley argues that such functional claims lead to ambiguous patent boundaries and exacerbates patent thickets, undermining patent law's fundamental quid pro quo.[77] Lemley proposes shifting the patent

---

[76] *Id.* at 678–79.

[77] Mark A. Lemley, *Software Patents and the Return of Functional Claiming*, 2013 WIS. L. REV. 905, 907–08, 919–23 (2013) (highlighting problems created by patents that claim broadly defined functions rather than particular solutions,

system towards claims more defined by specific structures, algorithms, or methods actually invented, rather than purely functional abstractions.[78] While the proposal does not resolve all of Lemley's concerns—particularly the broader systemic challenges of defining software patent boundaries—it addresses his fundamental critique by requiring disclosure of practical protocols, empirical calibration data, and clear scaling rationales. By insisting that patentees demonstrate not merely abstract functionality but concrete methodologies and empirical testing, the proposal meaningfully engages with Lemley's call for specificity, anchoring functional claims in a demonstrable empirical foundation rather than purely abstract ideas.[79]

Rai emphasizes a related but distinct transparency issue raised by the use of AI, specifically concerning administrative legitimacy and explainability in patent examination. Rai cautions against the USPTO's overly rigid stance of demanding complete transparency (such as revealing all source code and training data), which could undermine private incentives for innovation and lead to strategic gaming of administrative procedures.[80] Instead, Rai proposes a nuanced form of "explainability," requiring disclosure sufficient to ensure accountability and rational decision-making without fully disclosing confidential technical details.[81] The warranting words standard similarly seeks to strike a balance between accountability and commercial confidentiality by requiring disclosure of empirically grounded evidence and protocols rather than full algorithmic details or raw code. This intermediate form of disclosure aligns closely with Rai's recommendations. Yet, the proposal itself does not fully specify the precise limits of transparency required, leaving open questions that Rai identifies, such as the degree to which data sets or algorithmic

---

resulting in patent thickets and boundary ambiguity).

[78] *Id.* at 946–47 (suggesting that claims should clearly define concrete structures or methods rather than abstract functionalities); *see also* Andrew Chin, *Let's Create a Concreteness Standard for Abstract Software Patents*, WIRED.COM (Nov. 2012) (arguing that "the utility of a patentable invention must be amenable to explanation by a single causal account that specifies the resources brought into play by the invention's use").

[79] *Id.* at 919–23, 946–47.

[80] Arti K. Rai, *Machine Learning at the Patent Office: Lessons for Patents and Administrative Law*, 104 IOWA L. REV. 2617, 2625–26 (2019).

[81] *Id.* at 2626–27, 2638–40.

parameters should be revealed. Thus, it thoughtfully engages Rai's critique by adopting an intermediate standard, but does not fully resolve the complex normative trade-offs Rai highlights concerning administrative transparency and commercial secrecy.[82]

Ebrahim critiques patent law's traditional disclosure doctrines for being fundamentally inadequate to address AI-generated inventions. According to Ebrahim, the inherent opacity of AI algorithms results in "unidentified inventions," whose claimed outputs appear as if created by human inventors yet whose inventive processes remain inscrutable, effectively undermining traditional enablement and written description requirements.[83] He argues for enhanced disclosure standards for AI inventions, including requiring more detailed structural descriptions of the algorithmic methodologies, training data, and decision-making processes used to generate the claimed inventions.[84] The proposal directly engages Ebrahim's critique by requiring representative protocols, calibration evidence, and clear rationales that justify scaling empirical findings from limited tests to broader genus claims. However, the proposal does not mandate disclosure of raw training data or complete algorithmic transparency, two details that Ebrahim specifically highlights.[85] Thus, while the proposal addresses and mitigates the key concerns Ebrahim raises about inscrutability, it stops short of fully adopting the most detailed disclosure measures Ebrahim recommends. The proposal offers a middle path, requiring empirical and structural anchoring without fully eliminating the inherent opacity that Ebrahim identifies as a critical limitation of AI-generated inventions.[86]

In sum, the warranting words proposal critically engages with and meaningfully addresses key concerns articulated by Lemley, Rai, and Ebrahim. By requiring patentees to disclose practical protocols, empirical calibration data, and scaling rationales, the proposal anchors abstract functional claims (Lemley), achieves a balanced administrative transparency (Rai), and reduces the opacity and

---

[82] *Id.*

[83] Tabrez Y. Ebrahim, *Artificial Intelligence Inventions & Patent Disclosure*, 125 PENN ST. L. REV. 147, 148–49, 171–75 (2020).

[84] *Id.* at 218–19.

[85] *Id.* at 172–74, 218–19.

[86] *Id.*

inscrutability of AI inventions (Ebrahim). However, it does not claim to fully resolve each author's comprehensive critique. Rather, it provides a practical, nuanced, and doctrinally coherent way forward that addresses central aspects of their concerns while acknowledging—and leaving room for—ongoing normative and doctrinal refinement.

VII. IMPLEMENTATION

A simple and effective implementation pathway already exists within the Patent Office's current regulatory framework. Under 35 U.S.C. § 2(b)(2), the USPTO has regulatory authority to establish procedural rules for examining patent applications, and under § 131, it is obligated to ensure that each application meets statutory requirements.[87] Using this existing authority, the USPTO could issue a targeted regulation requiring patent applicants relying on AI-generated content to include a short-form "falsifiability annex" with their applications. This annex would succinctly describe: (i) the class-wide empirical testing protocol used, (ii) calibration evidence demonstrating reliability above baseline hallucination rates, and (iii) a clear scaling rationale connecting specific test data to the broader claimed genus. Such an annex would formalize expectations already implicit within § 112(a)'s written description and enablement doctrines, clarifying rather than significantly expanding applicant obligations.

Practical implementation would primarily depend on modest changes to existing examiner training programs. The USPTO already delivers extensive technology-specific examiner training through its Patent Examiner Technical Training Program (PETTP). In FY 2021, PETTP hosted over 500 training events totaling more than 37,000 hours for nearly 27,000 examiners, addressing a wide range of emerging technological issues.[88] Integrating a concise module into PETTP focused on recognizing when AI-generated content requires falsification-oriented disclosure—such as calibration evidence or

---

[87] See 35 U.S.C. § 2(b)(2) (delegating regulatory authority to the USPTO Director to establish procedures consistent with law); 35 U.S.C. § 131 (requiring examination of all filed patent applications to determine patentability).

[88] U.S. Patent & Trademark Office, Patent Examiner Technical Training Program, 75 Fed. Reg. 56059 (Sept. 15, 2010); U.S. Patent & Trademark Office, FY 2021 Performance and Accountability Report 197 (Nov. 15, 2021), https://www.uspto.gov/sites/default/files/documents/ USPTOFY21PAR.pdf.

defined experimental protocols—would be straightforward. This module could succinctly introduce examiners to the statistical fundamentals underlying AI-generated disclosures, including the Kalai–Vempala hallucination bound. Because PETTP is already virtual and broadly accessible across USPTO technology centers, adding this short module would require minimal additional investment or infrastructure.

Judicial reinforcement of the warranting words standard would similarly follow established doctrinal paths. When the Federal Circuit encounters its first appeal challenging a rejection under § 112(a) based on a missing falsifiability annex, the court could rely on principles recently affirmed by the Supreme Court in Amgen. In Amgen, the Court emphasized that the patent law demands disclosures commensurate with claim breadth, noting specifically that "the more one claims, the more one must enable."[89] Extending this logic to written description, the Federal Circuit could straightforwardly adopt the principle that genus claims without demonstrable empirical grounding—such as calibration data and clearly articulated testing methods—are merely "roadmaps" rather than adequate disclosures.[90] This modest doctrinal step would align seamlessly with existing jurisprudence, reinforcing the requirement of empirical grounding without significantly altering judicial practice.

Lastly, measuring the effectiveness of this implementation would be empirically straightforward. The USPTO already collects and publicly reports patent quality metrics and examiner training effectiveness in its annual Performance and Accountability Reports. Comparing rejection rates under § 112(a) before and after the introduction of the falsifiability annex could provide immediate, objective feedback on effectiveness. Furthermore, tracking annex submissions by small and micro entities—potentially encouraged by simplified forms or fee waivers—could offer additional insight into whether the standard is accessible and equitable.

These implementation steps require neither congressional action nor significant administrative upheaval. Instead, they represent a modest and targeted procedural adjustment within existing frameworks. By requiring warrantable disclosures—those that

---

[89] Amgen v. Sanofi, 598 U.S. 594, 610 (2023).

[90] *Id.*

empirically ground patent claims and clearly establish their falsifiability—the patent system can adapt effectively to the rise of generative AI, preserving innovation incentives and strengthening the transparency and accountability at the core of the patent system's fundamental quid pro quo.

VIII. POTENTIAL CONCERNS

Implementing the warranting words standard raises legitimate concerns, but these can be addressed clearly and practically. One objection might be that an enhanced written description requirement imposes additional costs on innovators, especially early-stage entities. Empirical data, however, suggest otherwise. Freilich's comprehensive study of over two million chemistry and biotechnology patents shows that roughly 17 percent of patent examples are prophetic—yet nearly half of these include at least one working example.[91] In practice, most startups already generate some minimal experimental or computational data to attract investment or to satisfy regulatory expectations.[92] The proposed short-form falsifiability annex simply formalizes minimal best practices already commonplace in research-driven industries, requiring a generalized protocol, a representative data set demonstrating calibration, and a brief rationale explaining scalability. This modest addition aligns closely with existing innovation norms and imposes only incremental costs on patentees.

Another potential concern is that even modest additional disclosure requirements could disproportionately burden small and micro entities. Recent statutory changes, however, largely mitigate this risk. The Unleashing American Innovators Act of 2022 significantly reduced fees for small entities by 60 percent and micro entities by 80 percent, greatly easing their financial obligations in patent prosecution.[93] Under this new structure, basic patent-processing fees—including filing, search, and examination—for a micro-entity applicant total less than $400.[94] Moreover, existing USPTO programs, including pro bono and pro se assistance, further alleviate potential

---

[91] Freilich, supra note 8, at 692-98.
[92] *Id.* at 689-90.
[93] Pub. L. 117-328.
[94] USPTO, Small Entity Compliance Guide (2022).

financial or procedural barriers.[95] In fact, the minimal burden of submitting a short-form falsifiability annex compares favorably with current best practices in biotechnology patenting, where deposits under the Budapest Treaty routinely incur higher logistical costs yet remain manageable even for small entities.[96]

Finally, a potential international concern arises over compatibility with TRIPS and the Budapest Treaty. Under TRIPS Article 29.1, WTO member states must require disclosures sufficient for skilled artisans to practice the invention and may require disclosure of the "best mode."[97] WTO precedent, specifically *Canada – Pharmaceutical Products*, confirms that technology-specific disclosure requirements that pursue legitimate public policy goals do not constitute improper discrimination.[98] Similarly, the Budapest Treaty's established practice of requiring biological material deposits reflects international acceptance of invention-specific disclosure obligations.[99] As the UNCTAD Resource Book emphasizes, TRIPS defines a minimum rather than a maximum disclosure standard, permitting member states to require additional particulars essential for meaningful technology transfer.[100] Thus, the warranting words standard, by demanding focused, verifiable disclosures, fits comfortably within the existing international intellectual property framework.

In short, none of these counterarguments presents a fundamental obstacle to implementing the warranting words proposal. The requirement aligns closely with current empirical practices, is financially and procedurally accessible even to small innovators, and is consistent with international obligations. By reinforcing patent law's fundamental exchange—exclusive rights in return for genuine, empirically verifiable disclosure—the standard would help ensure the

---

[95] *See id.*

[96] Budapest Treaty on the International Recognition of the Deposit of Microorganisms (1977) (establishing internationally recognized practices for biological material deposits).

[97] TRIPS Agreement, art. 29 ¶ 1 (1994).

[98] WTO Panel Report, Canada – Pharmaceutical Products, WT/DS114/R ¶ 7.101 (Apr. 7, 2000) (approving technology-specific disclosure requirements when they serve legitimate public policy objectives).

[99] Budapest Treaty, *supra* note 96.

[100] UNCTAD–ICTSD, RESOURCE BOOK ON TRIPS AND DEVELOPMENT 411–12 (2005).

patent system remains credible, balanced, and capable of promoting genuine technological advancement.

IX. CONCLUSION

Generative AI drafting poses a fundamental challenge to patent law's traditional exchange: exclusive rights in return for transparent and empirically anchored disclosures. The proliferation of AI-generated prophetic claims and expansive genus disclosures risks diluting patent quality by filling the system with inventions whose validity and reproducibility remain uncertain. Without a clear doctrinal response, patent disclosures may become increasingly unmoored from reality, undermining both public trust and downstream innovation.

The warranting words standard addresses this concern by requiring patentees to anchor their disclosures in falsifiable empirical evidence. Under this standard, AI-generated disclosures must provide clear testing protocols, calibration data to confirm reliability above inherent hallucination thresholds, and concise rationales demonstrating why limited empirical validation justifies broader genus claims. These modest but requirements ensure disclosures remain tethered to real-world, reproducible knowledge, without substantially increasing costs or procedural complexity.

Its strength lies precisely in its doctrinal simplicity and feasibility. Rather than imposing radically new obligations, the proposal clarifies and formalizes requirements already implicit within § 112(a)'s written description and enablement doctrines. As demonstrated by Freilich's study, innovators—including small entities—already routinely produce empirical data to support prophetic examples; the proposal merely standardizes these practices, ensuring disclosures reflect credible invention rather than speculative prophecy. Moreover, statutory reforms like the Unleashing American Innovators Act significantly mitigate the cost concerns for small entities, providing substantial fee relief that ensures accessibility.

Likewise, the standard fits neatly within established international frameworks, including TRIPS and the Budapest Treaty.104 These international agreements set disclosure minimums rather than ceilings, clearly accommodating enhanced, invention-specific disclosure requirements to ensure meaningful transfer of technology. Thus, the warranting words proposal aligns naturally with both

domestic and international legal principles, reinforcing the patent system's core policy goals rather than disrupting them.

Ultimately, restoring falsifiability to patent disclosures is not merely a technical adjustment but a reaffirmation of patent law's essential quid pro quo: time-limited exclusivity in exchange for public knowledge that is transparent, reproducible, and empirically credible. By requiring applicants to anchor their claims in reality, the warranting words standard ensures that patent law continues to fulfill its foundational promise, even in an era increasingly shaped by generative artificial intelligence.