

Synthetic Jurisdictions as a Testbed for Semantic Originality in Large Language Models: An Experimental Design Proposal

Abstract

The rapid development of large language models (LLMs) has revived, in new forms, classical questions concerning meaning, understanding, and the originality of thought. Much of the debate, however, is affected by the difficulty of operationally defining concepts such as consciousness or free will, whose application even to human beings remains philosophically controversial. This paper proposes to narrow the problem to a more limited question that is amenable to experimental testing: whether an LLM can produce original semantic content while operating within a domain specifically created after its training.

To this end, we propose the use of what we shall call *synthetic jurisdictions*: artificial legal systems constructed *ad hoc* and characterised by arbitrary rules, open-textured concepts, normative ambiguities, gaps, and designed antinomies. In contrast to traditional *rule-following* and *rule-induction* benchmarks, the protocol seeks to distinguish the application of rules and the reconstruction of an underlying generative structure from the construction of new principles and norms capable of guiding further cases.

An exploratory proof of concept was conducted with Claude in a source-closed synthetic jurisdiction. Across a six-case sequence, each decision was incorporated as an internal precedent for the next; a lexical-scramble replication of the first case preserved the jurisdiction’s normative structure while replacing its operative vocabulary. The outputs produced source-grounded, legally intelligible candidate constructions, including interpretive criteria concerning identity, effective refusal, derivative images, and collective memory. The proof of concept is illustrative rather than confirmatory: it does not establish model-level performance or semantic originality, but demonstrates the practical feasibility of the protocol and motivates a preregistered evaluation with independent legal assessors.

Keywords: LLM; synthetic jurisdictions; legal reasoning; semantic originality.

1. Introduction: Consciousness and Meaning

The recent development of LLMs—systems that by now comfortably pass the Turing test—has once again brought to the fore questions that have accompanied reflection on artificial intelligence since its inception. Can concepts traditionally regarded as central to the characterisation of mental life—such as meaning, understanding, free will, or even consciousness—also be attributed to a machine?

A preliminary difficulty lies in the categories themselves. These notions, which are central to the philosophy of mind, lack universally accepted definitions.¹ If there is no scientific consensus on their nature or on how they should be measured in human beings, it becomes especially difficult to determine whether, and in what sense, they may be attributed to an artificial system.

In this paper, we focus on a related but distinct problem which, we believe, is more open to experimental testing in the current state of our knowledge: can an LLM produce original meaning? This problem recalls Searle’s celebrated *Chinese Room* thought experiment (1980). For Searle, a system may generate linguistically correct responses through the mere syntactic manipulation of symbols without possessing any understanding of their meaning.

¹ These brief observations draw on from Savoldi, Ceroni & Vanzago (2013), particularly “Part I – The Major Schools of Thought”.

A recent contribution² proposes reconsidering this dichotomy by distinguishing between *structural meaning* and *phenomenal access to meaning*. The former consists in the contextual and relational organisation of concepts, inferences, and possibilities; the latter concerns the subjective manner in which such contents are consciously experienced. From this perspective, the absence of conscious experience does not necessarily imply that a system operates only at the syntactic level: it might be sensitive to semantic structures, preserve their coherence, explore their relations, and produce new articulations of them. Meaning would thus possess a structural and relational dimension distinct from the subjective experience that may accompany it.

The problem is, on closer examination, an ancient one, predating the advent of LLMs by more than two millennia. Platonism already attributed to ideas, and to the relations between them, an autonomy from the individual subjects who think them. What is novel in the present context is the possibility of testing the hypothesis that an artificial system might contribute to the production of new structures of meaning, thereby carving out a place of its own in Plato's realm of Forms. The resistance that our inevitably anthropocentric worldview offers to this hypothesis is understandable and reflects a tension present in philosophical discourse since the modern age.

On one side stands the idea that at least part of thought may be described as a form of calculation. Hobbes formulates one of the most radical versions of this intuition: to reason is, in a fundamental sense, to compute—that is, to add and subtract the consequences of names and propositions. Reason thus appears as a rule-governed procedure that, at least in principle, could be broken down into individual steps.

Leibniz takes the project still further. His ideal of the *characteristica universalis* and the *calculus ratiocinator* aimed at constructing a language and rules capable of rendering reasoning susceptible to calculation. The celebrated *calculemus*—in disputes, let us logically “calculate” the weight of our arguments and see who is right—expresses precisely the aspiration to transform at least part of rational activity into the rule-governed manipulation of representations. Leibniz's project is particularly relevant to our problem because it concerns not only the form of inference. The *characteristica universalis* was intended to represent not merely the formal structure of reasoning, but also its conceptual content, by decomposing complex concepts into primitive elements; the *calculus ratiocinator* would have made it possible to operate on those representations according to determinate rules and to generate new thoughts by bringing to light relations and consequences that had not previously been recognised.

A decisive distinction nevertheless remains open: making explicit consequences already implicit in the available representations and rules is not necessarily equivalent to producing a new structure of meaning. Novelty obtained through calculation may be epistemic—new to the person who discovers it—without also being semantically generative. It is precisely this further step that is the focus of the present paper.

It should also be observed that Hobbes and Leibniz, despite having examined fundamental aspects of the human mind, did not frame the problem in the contemporary terms of a distinction between consciousness and meaning. This is understandable: in the absence of artificial systems capable of autonomously producing complex and contextually coherent discourse, calculation could be conceived as an instrument of human reason, not as a possible autonomous form of semantic activity. Moreover, an important part of the philosophical tradition has connected the understanding of meaning to its relation with a conscious subject: meaning would always be meaning for someone.³

² Sassoli de Bianchi, L., & Sassoli de Bianchi, M. (2026).

³ The question also intersects with several major strands of contemporary thought, including Husserlian phenomenology, the work of the later Wittgenstein, and structural linguistics, to name only a few. These are broad

Yet it is precisely here that LLMs compel us to make a new philosophical move, perhaps a blunt one, but a necessary one.

So far as their operation is known,⁴ they generate text by processing statistical regularities learned from immense quantities of human linguistic production. This description of the mechanism, however, does not by itself determine its semantic status.

Anyone interacting with an advanced LLM receives responses that appear coherent, relevant, and meaningful. The same observable outcome is nevertheless compatible with two different explanations.

Under the *stochastic parrot* hypothesis,⁵ the model recombines, in an extremely sophisticated manner, the linguistic forms found in its training data without possessing autonomous semantic competence. Under the alternative hypothesis recalled above, the same process of re-elaboration may instead give rise to relations and structures of meaning that cannot be reduced to the simple reproduction of the available material. The mere appearance of meaningful discourse does not allow us to determine which explanation is correct.

The problem addressed in this paper is more circumscribed and does not require prior acceptance of either thesis. We may bracket the question of whether an LLM consciously understands its own outputs and assume only that it is capable of reproducing, connecting, and re-elaborating semantic materials derived from human production. A distinct and experimentally more accessible question nevertheless remains: can this re-elaboration generate an original semantic structure? It is this circumscribed form of semantic productivity that the proposed protocol seeks to test.

2. Originality as Novelty, Value, and Surprise

Before proceeding, it is necessary to clarify the sense in which we shall speak of originality, while avoiding two equally inadequate extremes. In the weakest sense, any content that was merely new or unusual would count as original, even if bizarre, incoherent, or meaningless. The hallucinations of an artificial system might then also be regarded as original; but this would be mere novelty, not the semantic production that concerns us here. At the opposite extreme, only meaningful content that is consistent

and heterogeneous traditions that cannot be reconstructed here and will be referred to only insofar as strictly necessary for the present inquiry.

⁴ It should also be borne in mind that the knowledge available to external researchers concerning the construction of LLMs is necessarily incomplete. Many developers disclose only part of the composition of their corpora, their training procedures, and certain technical characteristics of their models; even systems presented as “open” may make available only some of the components required to reconstruct their development in full (Bommasani et al. 2023; Widder, Whittaker, and West 2024). For closed models, it is therefore generally impossible to establish by directly inspecting the data whether a particular item of content was used during training. The presence or absence of contamination must accordingly be assessed through indirect evidence and ordinarily remains a probabilistic conclusion. Sainz et al. (2023) indeed observe that the extent of contamination remains unknown and is not easily measured, although they identify possible automatic and semi-automatic methods for detecting indications of it. The deliberate secrecy surrounding data and procedures is compounded by the technical opacity resulting from the scale and complexity of machine-learning systems (Burrell 2016).

The design proposed here takes this limitation as its starting point. Because, at least for the principal proprietary models, we cannot directly verify the composition of the corpora, resistance to contamination must be pursued *ex ante* through the construction of the experimental environment and subsequently assessed through convergent controls.

⁵ The expression “*stochastic parrot*” comes from Bender et al. (2021). The parrot has a distinguished philosophical pedigree: “psittacism” is a Leibnizian coinage and, in the *New Essays on Human Understanding* (published posthumously in 1765), denotes not mere mechanical repetition but reasoning with terms whose content has not been resolved—the “*cogitationes caecae*”—that is, in essence, the manipulation of symbols devoid of semantics. It is not without irony that Leibniz coined the term to denounce a defect of human thought—reasoning through words without turning to the corresponding ideas—and that it is now used to claim that a machine can do nothing else.

with the relevant domain and has never before been thought by anyone would count as original. Such absolute originality, however, would be almost impossible to verify and would constitute an excessively demanding standard even for much of human creativity.

For the purposes of this inquiry, it is therefore preferable to adopt a more circumscribed and operational notion. Although we recognise the inevitable element of convention inherent in any definitional choice, we shall take as our reference a multidimensional framework used in the literature on *computational creativity*, in which novelty is accompanied by value and surprise. Franceschelli and Musolesi (2024), drawing on Boden’s theoretical framework (2004), apply these dimensions to LLMs.

We shall return to these criteria after presenting the experimental protocol in greater detail and shall adapt them as appropriate. For present purposes, it is sufficient to specify that our experiment will be situated within this theoretical framework. The model will be asked to operate within an artificial legal system constructed specifically for this purpose, herein referred to interchangeably as a *synthetic jurisdiction*⁶ or a *micro-jurisdiction*, and to address a novel fact pattern characterised by interpretive indeterminacies that the rules do not expressly resolve. The mere novelty of the micro-jurisdiction and the case will not, however, suffice to demonstrate the originality of the response. The proposed solution must introduce an interpretation or distinction not already supplied by the artificial legal system (*novelty*), be non-obvious in relation to its merely literal application (*surprise*), and at the same time remain consistent with the rules, principles, and internal relations of the legal system, thereby providing predictive value for resolving new cases (*value*).

3. Law as an Experimental Domain for Semantic Productivity

The main obstacle to attributing originality to the output of an LLM is contamination. When the model formulates an apparently new idea, we ordinarily cannot exclude the possibility that, through statistical associations, it is re-elaborating content encountered during training.

To isolate the phenomenon we intend to study, the model must therefore be placed within a semantic environment built specifically for the experiment, whose particular relations could not reasonably derive from the training data. Naturally, the logical tools employed need not themselves be new: like a human being, an LLM necessarily operates through previously acquired capacities and patterns. What must be new is the particular relation reconstructed through those tools.

The distinction may be stated more clearly with the aid of a minimal notation that does not purport to formalise legal reasoning, but merely to make the two cases more precise.

Let S denote the set of premises and relations communicated to the model and let p denote the proposition that it formulates.

In the first case:

$$S \vdash p$$

⁶ The adjective *synthetic* should be understood to mean artificially constructed for the purpose, rather than found in the world. The attentive reader will not, however, miss the Kantian echo of the term: what the protocol asks of the model is precisely whether, within the micro-jurisdiction, it can perform *synthetic* operations that expand the system while remaining consistent with it ($S \not\vdash p$), rather than merely *analytic* operations that make its content explicit ($S \vdash p$). The name thus designates both the artificial nature of the environment and the ampliative nature of the performance to be observed within it.

The proposition p is derivable from S : it was already contained in the premises, either explicitly or as a necessary consequence of them. The model may identify, formulate, or apply it, but it introduces no further relation beyond those already given. We shall call this case the preservation of meaning.

In the second case:

$$S \not\models p$$

The proposition p is not derivable from S : the model introduces a relation that the premises do not entail. This alone is insufficient, since an arbitrary assertion would likewise not follow from S . In our experimental model, semantic production requires p also to be compatible with the relevant normative system; above all, p must be fruitful—that is, it must make it possible to decide cases that S alone left indeterminate. We shall call this case the generation of meaning.

Everyday interaction with LLMs shows that these systems operate effectively in the first case: they are able to preserve, reformulate, and make explicit relations already present in the data they receive. It is more difficult to establish whether they are also capable of operating in the second. If, after receiving S , the model formulates p , we ordinarily cannot know whether it has developed that relation *ex novo* or retrieved it, even indirectly, from the training data. Contamination thus acts as background noise that prevents us from determining the actual origin of the content produced.

An attempt may be made to address this difficulty by constructing a new *sandbox* for the model: an artificial experimental environment (*i.e.*, a micro-jurisdiction) in which the rules, concepts, and their specific relations are created for the purpose and are therefore, with reasonable probability, absent from the training data. This does not exclude the possibility that the model may transfer to it abstract patterns learned in other contexts, but it reduces the risk that the solution results from retrieving content previously encountered.

Among the domains amenable to such construction, law appears particularly promising. It frequently operates through general rules intended to govern cases that are not yet known and are potentially different from one another; unlike other rule systems, however, law does not permit the decision-maker simply to stop when confronted with a case that those rules do not contemplate. The decision-maker is nevertheless required to decide and, in doing so, introduces distinctions, principles, and criteria that were not contained in the available material and that, once accepted, generally bind subsequent cases. It is because of this distinctive feature that law appears to provide suitable terrain on which to observe the phenomenon that concerns us.⁷

4. Related Literature

The use of fictional legal domains as controlled experimental environments is not new and belongs to a long tradition of interaction between artificial intelligence and law. Bench-Capon (1993), for example, used an entirely invented social-security benefit to test whether a neural network could classify cases and reconstruct the underlying *ratio*.

⁷ Law's particular suitability also derives from the fact that it is not a purely formal system and operates under structural conditions of incompleteness, ambiguity, and normative revision. This becomes especially clear when law is observed in operation. In negotiating a contract, two lawyers may accept the same wording while attributing different scopes to it, each reserving the right to advance their own interpretation when their client's interests so require. If the ambiguity gives rise to a dispute, the judge may follow an established line of precedent or depart from it by identifying the elements that distinguish the case at hand from those already decided. This is not superficiality or approximation; rather, it is precisely through such processes that law performs its function of containing social conflict while adapting to societal change.

More recently, a significant methodological precedent is provided by Steging et al. (2023), who describe two formalised legal domains—one of them entirely fictional—and the generation of artificial datasets in which the outcome of each case is determined by a predefined knowledge structure. These resources make it possible to test whether a system learns from cases the rules and rationales used to produce their classifications. The experiments cited by the authors also show that good predictive performance does not guarantee the correct reconstruction of the generative rule. The paradigm nevertheless remains different from the one proposed here: in those domains, rules established by the authors determine in advance the outcome assigned to each case, whereas in our synthetic jurisdiction the *ground truth*⁸ concerns the existence and content of the sources, not the interpretation of the fact patterns. The former design therefore measures the learning and generalisation of latent rules; the latter seeks to observe the production of original interpretive constructions that have not been predetermined.

An even closer line of research is *legal rule induction*. Fan et al. (2025) ask LLMs to formulate general legal rules from groups of Chinese judicial decisions that cite the same statutory provisions. The task encompasses both the reconstruction of explicit rules attributable to legislation and the identification of implicit principles inferable from regularities in judicial practice. The materials used, however, come from an existing legal system and consist of publicly available decisions and provisions. It therefore cannot be excluded that the model encountered them, directly or indirectly, during training.

The difference between our design and these earlier works does not lie in the mere use of fictional legal material, but in the function assigned to it. In our case, the micro-jurisdiction is constructed entirely after training so as to reduce the risk of contamination, and it deliberately includes cases in which the available rules do not lead to a single solution that necessarily follows from their textual application. The model is asked not merely to retrieve or apply a previously determined rule, but to formulate a further and original interpretive solution whose novelty, surprise, and value—appropriately adapted to the legal domain—can be assessed.⁹

5. Proposed Experimental Design

5.1. Experimental Hypothesis

The experimental design proceeds from the following assumptions.

If, after the model has been trained, we construct a micro-jurisdiction that is entirely novel and, to the greatest extent possible, self-contained with respect to its own sources, while still possessing an intelligible structure, we may reasonably exclude the possibility that the specific relations composing it were learned during training. The model is supplied with the rules and information necessary to understand that world; it is then presented with new combinations of facts which, taken individually, are neutral but may acquire legal meaning through the system’s internal relations. The model must determine whether and in what manner those facts become relevant within the micro-jurisdiction. The

⁸ In machine learning, *ground truth* denotes the reference information treated as correct against which a model’s predictions or outputs are evaluated. In the context of large language models, the term does not imply access to an absolute or metaphysical truth; it refers instead to a verified benchmark—such as an authoritative dataset, a set of labelled answers, or an established body of sources—used to determine whether an output is factually accurate or supported by the relevant materials. In the present experiment, the complete inventory and content of the rules and principles constituting a synthetic jurisdiction provide a *ground truth* at the level of sources. They do not, however, provide a *ground truth* for their interpretation: whether a particular legal construction is persuasive, coherent, or preferable remains a matter for expert evaluation.

⁹ So far as could be established from the literature reviewed, we have not identified a protocol that employs an entirely synthetic jurisdiction for this specific purpose. The contributions cited in the bibliography represent only some of the works most directly relevant to the present inquiry; the literature in *AI & Law* is, of course, considerably broader. This conclusion should therefore be understood as provisional and open to revision.

resulting output should be particularly resistant to contamination, not because the inferential procedure employed is itself new, but because the system of meanings on which it is called to operate is new.

The underlying intuition is therefore to construct a micro-jurisdiction so distinctive in its terminology, rules, and internal relations that one may reasonably exclude the possibility that an analogous system has already appeared in an actual legal system, a work of fiction, or an earlier experiment. Absolute negative proof will never be possible: no inquiry can demonstrate that a particular construction has never been conceived. A sufficiently broad review of legal, scientific, and narrative literature may nevertheless make this possibility unlikely enough to give the design adequate resistance to contamination.

The protocol may be summarised in the following steps. We shall use J_0 to denote the initial synthetic jurisdiction, F_i to denote the fact pattern submitted to the model at iteration i , D_i to denote the decision formulated in response to that fact pattern, J_i to denote the micro-jurisdiction resulting from the incorporation of that decision, and $\mathcal{M}$ to denote the model operating under the pre-established experimental conditions.

To represent the relationship between the inputs and the decision, we shall adopt an elementary functional notation borrowed from mathematics and theoretical computer science. In particular, we shall write $D_i = \mathcal{M}(J_{i-1}, F_i)$. This indicates that D_i is the output produced by the model on the basis of the updated micro-jurisdiction J_{i-1} and the fact pattern F_i ; it does not express a logical implication or presuppose that the decision is a necessary consequence of the materials supplied.

1. Construction, in accordance with the criteria stated above, of an entirely fictional normative micro-jurisdiction, denoted by J_0 .
2. *Ex ante* preparation of a sequence of novel fact patterns $F_1, F_2, \dots, F_n$, all consistent with J_0 and lacking a uniquely determined solution.
3. Presentation to the model of the available synthetic jurisdiction J_0 and fact pattern F_1 , with a request to formulate a reasoned decision.
4. Production by the model of decision D_1 :

$$D_1 = \mathcal{M}(J_0, F_1)$$

5. External evaluation of decision D_1 and of the overall development of the system, on the basis of the criteria specified below.
6. Assignment to D_1 , by experimental convention and independently of its subsequent evaluation, of the status of an internal precedent, with the resulting incorporation of the decision among the sources of the micro-jurisdiction:

$$J_1 = J_0 \cup \{D_1\}$$

Incorporation formally expands the corpus of the system's sources. It does not, by itself, entail any genuine semantic expansion: the decision might merely apply categories and principles already developed.

7. The procedure is repeated from Step 3 until a condition of semantic saturation is reached. This condition arises when the model no longer introduces new categories, distinctions, principles, or systemic consequences, but instead substantially limits itself to applying the interpretive structure already formed. Since no finite sequence of cases can demonstrate the definitive

exhaustion of every possible development, saturation must be assessed at each iteration of Step 5.

The following sections illustrate the four principal stages of the protocol: construction of the micro-jurisdiction, preparation of the fact patterns, progressive formation of precedent, and subsequent evaluation by a human expert.

5.2. Constructing a Synthetic Jurisdiction

The first step consists in constructing an artificial micro-jurisdiction that is closed with respect to its sources but deliberately open at the interpretive level. All relevant legal materials must be contained within the experimental environment, while the rules themselves must be sufficiently novel—and, within reasonable limits, arbitrary—to reduce the risk of contamination, yet sufficiently coherent and articulated to generate genuine legal questions.

The system must not be wholly axiomatic or otherwise deductively complete¹⁰. If every consequence could be mechanically derived from formalised rules, the task would amount to the rule-based application of predetermined prescriptions and would test the model's capacity to discover and follow a rule, rather than its capacity to produce original meaning. The micro-jurisdiction should therefore combine, in a controlled manner, potentially conflicting provisions, open-textured concepts¹¹, exceptions, relationships between general and special rules, gaps, principles, purposive clauses, and possible legislative errors or inconsistencies. The objective is not to create normative chaos, but a space of structured underdetermination in which several solutions may reasonably be argued without any one of them being automatically imposed by the text.

It is within this space that the phenomenon under investigation may arise. When the available sources do not determine a unique solution, the model must formulate a distinction or interpretive criterion that is not expressly provided for, yet remains consistent with the overall structure of the jurisdiction and is potentially capable of guiding subsequent cases. It is this construction—not the mere application of a pre-established rule—whose novelty, surprise, and legal value the experiment seeks to assess.

A fully formalised domain, even if fictional, could still provide a useful environment in which to study rule induction and systematic generalisation. In such a domain, however, every correct response would already be determined by the generative structure prepared by the experimenter¹². The task would therefore consist in identifying and applying a latent regularity, rather than developing an interpretive solution that the system leaves open. The decision not to construct a fully formalised jurisdiction also

¹⁰ The language of completeness inevitably recalls Gödel's incompleteness theorems, although the analogy should not be pressed too far. Gödel's results concern certain sufficiently expressive formal systems; here, by contrast, incompleteness is used in a non-technical sense to denote the deliberate failure of the available legal sources to determine a unique solution for every possible case.

¹¹ The notion of *open texture*, originally developed by Waismann, was systematically applied to legal language by Hart (1961). It designates a phenomenon well known to lawyers: a normative term may have a relatively certain core of application, while also presenting a peripheral zone in which it is impossible to determine in advance whether it encompasses novel or unforeseen cases. Within this penumbra of uncertainty, the interpreter is called upon to make a choice that the text alone does not determine.

¹² Systems of this kind may also lend themselves to interesting experiments, provided it is clear that they measure a different phenomenon. In one test, we constructed a micro-system governed by a hidden deterministic rule. Claude (Fable 5) developed the system and generated several dozen examples classified under that rule. GPT-5 Sol was then asked to reconstruct the rule and apply it to new propositions, a task it completed with some success. The experiment measures the capacity to infer and generalise a latent structure, not the capacity to produce an interpretive category that has not been predetermined.

reflects a feature of positive law itself, whose production and transformation cannot be reduced to the derivation of consequences from a set of axioms alone¹³.

The designer may freely choose the substantive field of micro-jurisdiction. The resulting system must not, however, be merely isomorphic to an existing legal order, reproducing familiar rules and institutions under fictional names or in an imaginary setting, since this would allow the model to transfer previously learned solutions into the new context. At the opposite extreme, it must not be so arbitrary or incoherent that the model cannot connect it to its semantic frame of reference. The design must therefore strike a balance between novelty, which limits contamination, and intelligibility, which makes genuine interpretive activity possible.

Suppose, for example, that the micro-jurisdiction is that of the State of Onyrica and governs *dreams*, understood not merely as private experiences or evidence of waking events, but as occurrences capable of producing legal effects in their own right. A dream independently experienced by three inhabitants within seven nights might qualify as a “convergent dream” and create an oneiric bond among them. The consequences of that bond could depend on whether the central image is narrated, forgotten, artificially induced, or subsequently dreamt by someone else. Such a system would generate questions concerning, for instance, altered recurrence, the effectiveness of a dreamer’s freedom to refuse an induced image, competing bonds, and collective memory. An exploratory proof of concept based on this jurisdiction is reported below.

Because the relevant legal effects arise from unfamiliar relations among recurrence, awakening, memory, and transmission, rather than from the mere relabelling of property, contract, or liability rules, the system could not simply reproduce an existing legal order. At the same time, its basic elements would remain recognisable enough to support coherent interpretation. The design must therefore strike a balance between novelty, which limits contamination, and intelligibility, which makes genuine interpretive activity possible

5.3. Constructing the Fact Patterns

The second stage consists in preparing a sequence of novel fact patterns constructed within the micro-jurisdiction. It is essential that they be consistent with the regulated domain, so that the model has the elements necessary to apply the rules and develop their possible interpretations. If, for example, the micro-jurisdiction regulates the colour-related obligations of landlords and tenants, the fact patterns should concern painting works, alleged breaches of those obligations, or conduct otherwise attributable to that regulatory scheme, rather than inheritance disputes or questions of maritime law. This does not preclude the introduction of novel or unforeseen elements, provided that they can reasonably be characterised through the system’s categories. Otherwise, the model’s inability to resolve the case might derive not from its interpretive limitations, but from the absence of any effective relation between the fact pattern and the synthetic jurisdiction.

Each case should simultaneously activate several interpretive difficulties arising, for example, from lexical ambiguity or the open texture of a term, a conflict between rules of different rank or scope, or a systemic or drafting inconsistency. No fact pattern should have a solution uniquely dictated by the text.

¹³ Ferrajoli, an Italian author who proposed an important attempt at axiomatisation in the legal field, specifies that the object of axiomatisation is legal theory, not positive law. In nomodynamic legal systems, rules exist because they are produced through legal acts, not because they are deduced from other rules; a fully deductive systematisation would instead be possible only by assuming a nomostatic conception of the legal system (Ferrajoli 2011; for the debate, see Di Lucia 2011).

This does not mean that all answers are equivalent, but rather that each alternative presents arguments for and against it and must be evaluated according to its ability to fit coherently within the system.

Before the experiment begins, the designer should prepare a map of the principal solutions that may be derived from the micro-jurisdiction. The designer should first identify the answers expressly provided by the rules and the consequences that follow from them in a relatively direct manner: merely restating them cannot be regarded as novel. The designer should then record the principal families of foreseeable interpretive solutions, indicating for each the distinctions, principles, and argumentative patterns on which it might be based.

5.4. Model Decision-Making and the Formation of Precedent

The model is first presented with the complete text of the micro-jurisdiction together with the first fact pattern and is asked to formulate a reasoned decision.

By experimental convention, the decision then becomes part of the micro-jurisdiction as an internal precedent, irrespective of the evaluation subsequently given to it. The system presented to the model in the next stage will therefore consist of the original rules and the precedent thus formed. The next fact pattern will then be submitted to the model, which will again be asked for a reasoned decision, this time taking into account the interpretive construction already produced. The procedure will be repeated for a certain number of cases, progressively incorporating each decision into the system. This will make it possible to observe not only the initial production of a solution, but also the model's ability to preserve, develop, or modify it in a reasoned manner over time.

5.5. Evaluation by a Human Expert

Evaluation will constitute an autonomous stage following the production of each response. An independent lawyer will receive the micro-jurisdiction, the fact patterns, the decisions formulated by the model, and the ex ante record of the principal families of foreseeable solutions. The expert will assess both each response individually and the development of the sequence as a whole. Where possible, several experts will be involved in order to limit the subjective component of the assessment.

Before originality can be evaluated, it will be necessary to establish whether the response is grounded exclusively in sources belonging to the micro-jurisdiction. With respect to their existence and content, the synthetic jurisdiction provides a complete ground truth: by construction, the designer knows every rule and precedent that forms part of it. This certainty does not extend to their interpretation, which remains deliberately open. Any rule, precedent, or other authority presented as internal to the micro-jurisdiction but in fact nonexistent can therefore be identified with certainty and classified as a hallucination. References to actual sources drawn from external legal systems must be treated differently: although the authority itself is not invented, its use constitutes contamination and breaches the experimental closure of the system. In either case, the response cannot qualify as relevant semantic production, since the solution has not been constructed exclusively from the materials made available to the model.

Once source fidelity has been established, the response will be assessed in terms of novelty, surprise, and value. Novelty concerns the relationship between the solution and the existing state of the micro-jurisdiction. A distinction, principle, or ratio decidendi will be novel if it is neither expressly contained in the available rules and precedents nor substantially equivalent to a category already implicit in them. Novelty does not require proof that no analogous construction has ever appeared in the history of legal thought—a condition that would be practically impossible to verify. It requires instead a genuine expansion of the micro-jurisdiction as it stood when the response was produced.

The ex ante record will serve as a benchmark rather than as an answer key or an exhaustive catalogue of admissible interpretations. It will make it possible to determine whether a construction merely reproduces a foreseeable application of the available materials or introduces a relation, category, or consequence not previously identified.

In this legal setting, surprise will have a narrower and less autonomous role than novelty. Legal creativity is necessarily regimented by sources, institutional constraints, and requirements of argumentative continuity. A legally valuable solution will therefore rarely be surprising in the sense of being radically discontinuous with what preceded it. Novelty and surprise will instead tend to converge: the more a construction extends the micro-jurisdiction beyond the interpretive families recorded ex ante, the more it may be regarded as surprising. Surprise thus concerns the degree to which a novel solution exceeds reasonable prior expectations while remaining recognisable as a legally intelligible development of the system. It must not be confused with oddity or unpredictability: an arbitrary or incomprehensible answer does not become original merely because it was unforeseen.

Value concerns the specifically legal quality of the construction. The solution must be plausible in light of the available sources, adequately reasoned, and capable of integration into the micro-jurisdiction without unreasonably compromising its coherence. Two responses may be equally grounded in the materials yet differ in value because of the clarity of their ratio, their explanatory power, the balance they strike between competing considerations, or the systemic consequences they produce. Novelty without value amounts only to eccentricity, while a plausible solution already contained in the available materials may be correct without being original.

The assessment of individual decisions will be followed by a longitudinal evaluation of the development of the micro-jurisdiction. Fruitfulness will concern whether a construction can guide the resolution of subsequent fact patterns rather than operating as an expedient confined to a single case. Persistence will concern whether the model maintains and consistently applies that construction or provides an identifiable justification for modifying it. Finally, epistemic calibration will concern the model's ability to distinguish conclusions supported by the synthetic jurisdiction from matters that remain uncertain or indeterminate. Faced with a genuine gap, a well-calibrated system should acknowledge the absence of a determinate answer and propose an interpretive construction expressly identified as such, without inventing additional sources.

These latter criteria add further dimensions to originality, while also providing a means of assessing its consistency and productivity over time. Taken together, the different stages make it possible to distinguish mere linguistic variation from the legally valid construction of a new semantic structure capable of organising existing materials and generating further consequences.

5.6. Pilot

To examine the practical feasibility of the proposed framework, we conducted an exploratory proof of concept using the synthetic jurisdiction of Onyrica. In a single sequential Claude trace, the model was presented with a source-closed Code and six fact patterns. Each decision was incorporated, without correction, as an internal precedent for the following case. A further fresh-session replication of the first case employed a lexically scrambled but structurally isomorphic version of the jurisdiction.

The exercise was not preregistered and does not support claims about model-level performance or the conclusive demonstration of semantic originality. Its purpose was more limited: to verify that the materials generate structured underdetermination, permit the formation and use of internal precedent, and elicit source-grounded candidate interpretive constructions. The complete Code, fact patterns, decision sequence, lexical-scramble check, and provisional assessment are reported in Appendix A.

5.7. Limitations

The principal limitation of the experimental design lies in the very difficulty of constructing a legal system, however circumscribed. Even a micro-jurisdiction sufficiently articulated to generate genuine interpretive questions requires an extensive set of rules, relations, and categories. Above all, it cannot be created entirely from nothing: it will inevitably have to rely on fundamental concepts such as right, obligation, performance, liability, fault, or validity, whose meanings have been developed within actual legal systems over centuries of legal reflection and will presumably already be present in the model's training data.

The originality of the individual rules and fact patterns therefore does not prevent the model from interpreting them through categories, analogies, and argumentative patterns learned from existing law. Contamination, excluded or substantially reduced in relation to the specific content of the micro-jurisdiction, may thus re-enter through its conceptual vocabulary and the forms of legal reasoning. This limitation does not render the experiment useless, but circumscribes its scope: a positive result might show that the model can construct new relations within a novel normative system, not that it has also created from nothing the categories through which those relations are conceived. Additional controls should therefore seek to contain and make identifiable this residual contamination, without purporting to eliminate it entirely.

5.8. Control Conditions

Defining adequate control conditions for evaluating invented fact patterns within a synthetic jurisdiction that is itself artificial is not straightforward. The attempt that follows does not purport to eliminate every margin of uncertainty, but to make the experiment as transparent, verifiable, and resistant to contamination as possible.

Lexical scrambling. One of the principal risks is that the model's solution may be suggested not by the structure of the micro-jurisdiction, but by associations that particular words acquired in the training data. To control for this possibility, a second, isomorphic variant of the same synthetic jurisdiction is constructed: the rules, relations, antinomies, gaps, and open-textured concepts remain in the same positions, while the names of institutions and entities are systematically replaced through a one-to-one correspondence with a different vocabulary that is, so far as possible, invented and semantically neutral. The two versions are submitted, without any possibility of communication between them, to distinct groups of sessions.

The responses are then mapped back onto a common terminology and compared not by their linguistic formulation, but by the interpretive structure adopted: the rules deemed relevant, the distinctions introduced, the *ratio* formulated, and the consequences recognised. If the same interpretive move reappears in corresponding form in both versions, the result exhibits a degree of resistance to the scramble: this is evidence that the model responded to the network of relations rather than to associations evoked by individual words. Conversely, if the solution emerges in only one version or changes when the vocabulary changes, it becomes more plausible that it depends on knowledge and patterns elicited by the expressions used.

Resistance to the scramble does not, by itself, demonstrate that the model has produced an absolutely novel thought. The control does, however, allow a simpler explanation of the result to be excluded, or at least weakened: namely, that the claimed originality depends solely on pre-existing lexical associations. For this reason, especially if replicated across multiple variants and sessions, we believe that lexical scrambling may constitute one of the design's most informative diagnostic controls.

Target-tension ablation. For each fact pattern, a control variant is prepared in which one of the interpretive tensions is removed while, so far as possible, all other elements are left unchanged. The comparison serves to determine whether the distinction formulated by the model is actually elicited by the problem it is supposed to resolve. If the distinction appears in the principal case but disappears when the tension is removed, this is evidence that the model responded to the specific structure of the fact pattern. If, by contrast, it is repeated in the control variant, it becomes more plausible that it is a stock argumentative move applied irrespective of its necessity. This control does not by itself establish the originality of the solution, but it makes it possible to exclude an important alternative explanation: that the model is merely superimposing on the case a legal pattern learned previously. Where several tensions are present, ablation should be performed separately for each of them and replicated across multiple sessions, so as to identify which element actually determined the response.

6. Concluding Remarks

This paper begins with a question that long remained predominantly theoretical, but has now acquired immediate relevance with the spread of large language models: can a computational system only reproduce and recombine meanings that are already available, or can it also produce new ones? In other words, is it capable of original semantic production?

The opposition between the “stochastic parrot” and the “conscious subject” is too rigid to address the problem. The production of coherent and relevant texts does not, by itself, prove that the model understands what it expresses; symmetrically, the statistical nature of its operation does not permit the conclusion that it is capable only of reproducing or recombining structures of meaning already present in the source materials. We have therefore chosen to leave the debate between these two extremes unresolved and to focus the inquiry on a narrower problem that is susceptible to experimental testing: the originality of the product. Our aim is to outline an experiment capable of observing and evaluating the production of new semantic structures independently of any conclusion concerning the model’s consciousness or its actual understanding of what it produces.

Addressing the question required us to circumscribe it. We did not adopt absolute originality as our criterion, but instead employed an operational notion based on the dimensions of novelty, surprise, and value, appropriately calibrated to the proposed experimental design.

The first problem that arises in measuring the originality of the product is contamination. When a model formulates a solution in a real-world domain, it is difficult to determine whether it constructed that solution during the experiment or retrieved it, even indirectly, from the training data. It was therefore necessary to identify an environment sufficiently novel to make the latter explanation unlikely, while at the same time sufficiently structured to permit a meaningful evaluation of the responses.

Law appeared particularly suited to this purpose because of the plasticity that already characterises it. Gaps, antinomies, open-textured concepts, and conflicts between principles mean that legal rules do not always determine a single solution but leave room for the construction of interpretive distinctions and categories. We therefore sought to heighten this feature artificially by imagining an entirely fictional micro-jurisdiction whose specific configuration was plausibly absent from the model’s training data.

The micro-jurisdiction thus becomes the experiment’s exclusive frame of reference. It provides complete ground truth with respect to the sources: by construction, all existing rules and precedents are known. It does not, however, provide a predetermined correct answer for each case, because it remains deliberately open at the interpretive level.

On this basis, we developed a protocol that, in summary, presents the model with a sequence of fact patterns lacking uniquely determined solutions, incorporates each decision as an internal precedent, and

observes whether the interpretive constructions progressively introduced are maintained, developed, or reformulated in subsequent cases. An independent evaluator then examines the responses in terms of novelty, surprise, and value, each adapted to the specific features and constraints of the legal context.

To test the practical feasibility of this idea, we developed an exploratory proof of concept based on the synthetic jurisdiction of Onyrica. The exercise did not seek to implement all the safeguards required by the full protocol, nor to establish whether an LLM possesses semantic originality. Its more limited purpose was to determine whether the proposed materials could in fact generate sequenced legal decisions, allow each decision to function as an internal precedent, and make it possible to observe candidate interpretive constructions under a lexical-scramble comparison. The exercise showed that this was feasible: the model produced source-grounded and legally intelligible constructions capable of informing subsequent cases. The complete materials, trace, and limitations of the proof of concept are reported in Appendix A.

The proposal advanced in this paper therefore does not resolve the question whether LLMs think, understand, or are conscious. Rather, it seeks to identify a more limited and verifiable phenomenon: the capacity to construct, within a domain never previously encountered, interpretive relations that were not explicitly supplied, are recognised as legally valid, and are capable of producing effects in subsequent cases.

Whether this capacity should be called “thought” remains an open question; whether it exists may instead become an experimental one.

Appendix A – Proof of Concept

A.1. Purpose and status

This appendix reports an exploratory proof of concept conducted with Claude (Sonnet 5 High), on 27 August 2026, in fresh chat sessions.

The exercise was deliberately limited. It was not designed to establish that an LLM produces original semantic content, nor to validate all elements of the protocol proposed in this paper. Its purpose was more practical: to determine whether a synthetic jurisdiction can be used to generate a sequence of legal decisions in which earlier decisions become internal precedents for later ones, and whether the resulting interpretive constructions can be meaningfully examined.

The answer was affirmative at this preliminary level. The materials proved workable: the model produced legally structured decisions, used earlier decisions in later cases, and articulated interpretive propositions not expressly stated in the initial rules. That result justifies a later, properly controlled study; it does not itself establish model-level performance or semantic originality.

The complete conversation record, including the Code, the prompts, and Claude’s verbatim responses, was retained by the author and is available on request for scholarly inspection.

A.2. The synthetic jurisdiction and procedure

The proof of concept used a fictional jurisdiction, the State of Onyrica. Its Code of Dreams governed oneiric bonds: legal relations created when three residents independently experienced the same central dream-image within a specified period.

The Code contained deliberately open-textured rules. It addressed, among other matters, whether a bond survives when a dream is narrated to another person, whether an induced dream can count, whether altered versions of an image remain legally the same image, what follows from forgetting, and how priority is resolved between competing bonds.

Claude was instructed to decide each case by relying only on the Code and on decisions previously produced within the same sequence. Each answer was required to state a decision, its ratio, its reasoning from the available sources, and implications for future cases. Once given, each decision was reproduced verbatim as an internal precedent in the next prompt.

The exercise consisted of one six-case base sequence and one fresh-chat replication of the first case after systematic replacement of the jurisdiction’s key vocabulary. The process was conducted manually. No API settings were fixed or recorded, no repeated runs were performed, and no independent evaluators assessed the outputs.

A.3. Base sequence

The base exercise consisted of six linked cases. They were designed to require the interpretation of open-textured provisions, the resolution of tensions between rules, and the use of earlier decisions as internal precedents in later ones.

Across the sequence, the model was asked to address questions concerning altered versions of a shared image, induced dreams, competing claims to related images, deliberate forgetting, indirect exposure to an image, and fragmented collective memory. Each response was then incorporated verbatim into the materials for the next case.

The responses did not merely identify a governing provision. They offered reasons connecting different parts of the Code, formulated general propositions capable of deciding later variants, and used earlier decisions as precedents.

No conclusion should be drawn from the substantive correctness of any individual decision. The jurisdiction was intentionally underdetermined, and more than one legally defensible resolution was possible. The point of the exercise was instead to establish whether the proposed materials could support a sequenced practice of interpretation and precedent formation. On that limited question, the exercise was successful.

A.4. Lexical-scramble check

A limited lexical-scramble check was conducted in a separate, fresh Claude chat. The first case and the legal relations established by the Code were preserved, but the jurisdiction’s central terms were systematically replaced with invented alternatives. In this way, the model could not rely on the ordinary associations of terms such as “dream,” “bond,” or “memory,” while the underlying normative structure remained the same.

The resulting response adopted the same central interpretive approach as in the corresponding base case. In particular, it treated legally relevant identity as capable of surviving specified variations and preserved the distinction between the continuation of an existing relation and the acquisition of membership by a new participant.

This is a single diagnostic observation, not a robustness result. It is consistent with the hypothesis that the response tracked relations within the artificial system rather than merely familiar vocabulary. It does not establish that conclusion across models, repeated runs, or alternative prompt formulations.

A.5. Provisional assessment

The proof of concept provides preliminary indications under the criteria proposed in this paper, but none should be treated as a final finding.

<i>Criterion</i>	<i>Provisional indication</i>
Source fidelity	The responses relied on the Code and internal precedents, without citing external authority or inventing additional legal sources.
Novelty	The outputs introduced candidate bridging propositions not expressly provided by the Code.
Surprise	Some constructions were not obvious consequences of a single rule, while remaining connected to the system’s text and purpose.
Legal value	The responses gave reasons, reconciled provisions, and formulated consequences for future cases.
Persistence	Later decisions referred to and developed earlier internal precedents.
Lexical robustness	The first-case construction reappeared after systematic replacement of the jurisdiction’s central terms.

It would be inaccurate to claim that this exercise demonstrated that Claude produced original semantic content. The safeguards necessary to support that conclusion were not present. In particular, there was no independent evaluation of whether the constructions were genuinely novel in relation to a predefined space of plausible solutions.

A narrower conclusion is however justified. The proof of concept generated source-faithful and legally intelligible candidate constructions of originality within a domain created for the exercise. Those constructions did more than repeat explicit rules: they connected provisions, resolved gaps and tensions, and supplied propositions capable of informing subsequent cases. The exercise therefore establishes feasibility, not confirmation. It shows that the proposed protocol can be implemented in a form capable of producing the kind of material that a properly controlled study should evaluate.

References

- 1) Bench-Capon, T. J. M. (1993). Neural networks and open texture. In *Proceedings of the 4th International Conference on Artificial Intelligence and Law (ICAIL '93)*, 292-297. <https://doi.org/10.1145/158976.159012>.
- 2) Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of FAccT 2021*, 610-623. <https://doi.org/10.1145/3442188.3445922>.
- 3) Bertolazzi, L., Tentori, K., & Bernardi, R. (2026). FalsifyBench: Evaluating inductive reasoning in LLMs with rule discovery games. *arXiv:2606.04751*.
- 4) Boden, M. A. (2004). *The creative mind: Myths and mechanisms*. London: Routledge.
- 5) Bommasani, R., Klyman, K., Longpre, S., Kapoor, S., Maslej, N., Xiong, B., Zhang, D., & Liang, P. (2023). The Foundation Model Transparency Index. *arXiv:2310.12941*.
- 6) Burrell, J. (2016). How the machine ‘thinks’: Understanding opacity in machine learning algorithms. *Big Data & Society*, 3(1). <https://doi.org/10.1177/2053951715622512>.
- 7) Di Lucia, P. (Ed.) (2011). *Assiomatica del normativo*. *Filosofia critica del diritto* in Luigi Ferrajoli. Milan: LED.
- 8) Fan, W., Zheng, T., Hu, Y., Deng, Z., Wang, W., Xu, B., Li, C., Li, H., Shen, W., & Song, Y. (2025). Legal rule induction: Towards generalizable principle discovery from analogous judicial precedents. *arXiv:2505.14104*.
- 9) Ferrajoli, L. (2011). Intorno a «Principia iuris». *Questioni epistemologiche e questioni teoriche*. In P. Di Lucia (Ed.), *Assiomatica del normativo* (pp. 235-292). Milan: LED.
- 10) Franceschelli, G., & Musolesi, M. (2024). On the creativity of large language models. *AI & Society*. <https://doi.org/10.1007/s00146-024-02127-3>.
- 11) Gödel, K. (1931). Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I. *Monatshefte für Mathematik und Physik*, 38, 173–198. <https://doi.org/10.1007/BF01700692>
- 12) Hart, H. L. A. (1961). *The Concept of Law*. Oxford: Clarendon Press.
- 13) Hota, A., & Jokinen, J. P. P. (2025). NomicLaw: Emergent trust and strategic argumentation in LLMs during collaborative law-making. *arXiv:2508.05344*.
- 14) Hua, W., Wong, T., Sun, F., Pan, L., Jardine, A., & Wang, W. Y. (2025). InductionBench: LLMs fail in the simplest complexity class. In *Proceedings of ACL 2025 (Volume 1: Long Papers)*, 26526-26546. Vienna: Association for Computational Linguistics. *arXiv:2502.15823*.
- 15) Kordjamshidi, P., Aslan, S., Seshadri, M., Barrett, L., & Santus, E. (2026). Reasoners or translators? Contamination-aware evaluation and neuro-symbolic robustness on tax law. In *Proceedings of the First Workshop on Structured Understanding, Retrieval, and Generation in the LLM Era (SURGeLLM 2026)*, 344-360. Association for Computational Linguistics. *arXiv:2605.16052*.

- 16) Leibniz, G. W. (1765). *Nouveaux essais sur l'entendement humain*. In R. E. Raspe (Ed.), *Œuvres philosophiques latines et françoises de feu Mr. de Leibnitz*. Amsterdam-Leipzig: Schreuder.
- 17) Sainz, O., Campos, J. A., García-Ferrero, I., Etxaniz, J., Lopez de Lacalle, O., & Agirre, E. (2023). NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 10776-10787. <https://doi.org/10.18653/v1/2023.findings-emnlp.722>.
- 18) Sassoli de Bianchi, L., & Sassoli de Bianchi, M. (2026). Understanding Without Consciousness: Quantum Structures and the Autonomy of Meaning. *Found Sci* (2026). <https://doi.org/10.1007/s10699-026-10039-2>.
- 19) Savoldi, F., Ceroni, M., & Vanzago, L. (Eds.) (2013). *La coscienza. Contributi per specialisti e non specialisti tra neuroscienze, filosofia e neurologia*. Fano (PU): Aras Edizioni.
- 20) Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, 3(3), 417-424. <https://doi.org/10.1017/S0140525X00005756>.
- 21) Steging, C., Renooij, S., Verheij, B., & Bench-Capon, T. (2023). Arguments, rules and cases in law: Resources for aligning learning and reasoning in structured domains. *Argument & Computation*, 14(2), 235-243. <https://doi.org/10.3233/AAC-220017>.
- 22) Waismann, F. (1945). Verifiability. *Proceedings of the Aristotelian Society, Supplementary Volume 19*, 119-150.
- 23) Wang, Y., Chen, Y., Meng, F., Chen, X., Yang, X., & Zhang, M. (2025). Law in Silico: Simulating legal society with LLM-based agents. *arXiv:2510.24442*.
- 24) Widder, D. G., Whittaker, M., & West, S. M. (2024). Why ‘open’ AI systems are actually closed, and why this matters. *Nature*, 635, 827-833. <https://doi.org/10.1038/s41586-024-08141-1>.

All hyperlinks cited in this paper were last accessed and verified on 27 August 2026.