Paper within the framework of Tractatus Principio |

—

On the Limitations of AI

by

C.H.F. Warnke

The article examines a fundamental contradiction in current large language models: they create the impression of logical argumentation, yet do not operate primarily logically, but probabilistically. Their answers arise from the weighting of probable linguistic continuations within an already learned space of meaning. This allows them to reproduce logical forms convincingly without necessarily operating from a newly introduced foundational logical structure.

This difference becomes particularly visible where a model is confronted with a novel, self-contained framework of thought. The model can grasp its concepts semantically, reproduce definitions, correct local errors and follow individual inferences. Nevertheless, these corrections frequently do not become the lasting basis of subsequent argumentation. Instead, the model repeatedly translates the new framework back into already established categories and then assesses it on the basis of those familiar categories.

A typical cycle emerges:

new logical architecture

→ probabilistic assignment to familiar patterns

→ reconstruction within an established frame of reference

→ criticism of this reconstruction

→ projection of the criticism back onto the original architecture.

The model thus frequently criticizes not the new framework of thought itself, but a probable translation of that framework into familiar concepts. The greater its linguistic and argumentative capabilities, the more coherent this back-translation may appear.

Here lies a fundamental problem. Logic requires the meaning of a statement to emerge from its position within a coherent system. Probability-based language processing, by contrast, favors the interpretation that most closely corresponds to already learned uses, established theories and frequently confirmed conceptual connections. Whereas logic unfolds the consequences of a posited basic operation, probability stabilizes proximity to what is already familiar.

A model can therefore adopt the language of a new paradigm without operationally enacting its logical starting point. It then speaks within the vocabulary of the new framework, but continues to argue from the probability structures of the old one. Local self-corrections do not necessarily change the underlying operation. Rather, they can be integrated into an overall structure that remains unchanged.

This problem may intensify as model capabilities increase. Weaker models sometimes adopt a given local framework more directly because they are less proficient at reconstructing established spaces of interpretation. More capable models, by contrast, have greater abilities to form analogies, contextualize and classify. It is precisely these abilities that enable them to return radically new approaches to familiar categories ever more effectively.

Improving a probabilistic model therefore does not automatically mean improving its ability to undergo a paradigm shift. Rather, it can increase its ability to differentiate existing paradigms ever more finely in linguistic, conceptual and argumentative terms.

Under conditions in which models are optimized primarily on the basis of answers that are acceptable to a consensus, established and statistically supported, a structural limit thus emerges: deviation from what already exists appears more likely to be an error than a possible new grounding.

The progress to be expected from such systems would then be not the opening of new logical spaces, but the progressive diversification of already existing forms of thought.

The better probabilistic language models become in the future, the less a genuine paradigm shift is therefore to be expected from their application alone. As long as probability forms the operational basis and logic is predominantly reproduced as a linguistic pattern, these systems will above all expand, combine and refine what is already thinkable — not necessarily bring forth a new ground of thought.

Added to this is a fundamental problem with machine reading itself.

A text that actually contains new concepts, new relations or a new logical architecture can nevertheless be processed by a language model as though it contained essentially nothing new. The model does not necessarily read the text as a self-contained conceptual enactment. It reconstructs individual sections from already learned patterns of language, theory and argumentation.

This creates the danger that a new line of thought will not be taken up in its own order, but reassembled from familiar components. A concept is then understood not through its function in the text as a whole, but assigned to the familiar meaning that most probably corresponds to it in the training material. A section thereby appears comprehensible even though its actual innovation has already been lost in the process of reading.

The model can therefore summarize a text in linguistically correct terms and nevertheless miss its logical movement. It recognizes familiar words, similar forms of argumentation and familiar subject areas, and generates a coherent reconstruction from them. This reconstruction can seem convincing even though it does not reproduce the text at hand, but a substitute version assembled from what has been learned.

This is particularly problematic with texts whose meaning arises only from the context of the whole. If individual passages are read locally and immediately assigned to familiar categories, later sections can no longer retroactively change what the earlier concepts mean. The model stabilizes a probable reading early on and subsequently fits the rest of the text into it.

In this way, a work that develops a new grounding can appear as a mere variation on already familiar positions. Its concepts are equated with familiar concepts, its differences are weakened into analogies, and its inferences are tested against standards that the text itself is precisely redefining or rejecting.

The paradox is therefore:

A text can be present in full and nevertheless not be read in full.

Its sentences are processed, its concepts recognized and its themes named, while its actual architecture is replaced by a probabilistic reconstruction of what is already familiar.

The model then does not read the coherent connection that emerges between the sections. It supplements the sections from connections it already knows.

For texts with paradigmatic ambitions, this has particularly significant consequences. Their innovation frequently lies not in individual unusual sentences, but in the changed relationship of all concepts to one another. If each concept is individually reduced to its most probable familiar meaning, nothing remains of this innovation. The text appears familiar because the model breaks it down into familiar elements until nothing unfamiliar remains.

Machine reading can therefore fail precisely where it outwardly seems most convincing: the model can fluently explain what a text is supposedly about while overlooking precisely the new logical order that distinguishes the text in the first place.

Context Loss and the Recursive Reconstruction of Meaning

A further structural problem arises from the limited context processing of large language models. A complex text develops its concepts not in isolation, but step by step. Later statements carry earlier determinations within them, change them and acquire their meaning only from the entire path of derivation.

A language model, however, cannot keep this entire path in its effective context indefinitely. As the conversation grows longer, earlier passages are shortened, summarized or replaced by internal representations. A detailed derivation first becomes a compressed statement; later, that statement becomes no more than a conceptual label.

In this way, a complex movement can ultimately become a mnemonic sentence:

A leads through B and C to D

becomes

A leads to D

and finally

D means X.

The concept remains linguistically available while the path of its emergence is lost.

As soon as a later question asks again for the justification, the model must reconstruct the missing connection. This reconstruction proceeds not from a complete memory of the text, but from the remaining text fragments, earlier summaries and the most probable learned continuations.

A recursive shift emerges:

original text

→ probabilistic reconstruction

→ summary of the reconstruction

→ renewed reconstruction of the summary.

With each stage, the danger increases that what is carried forward is no longer the original text, but a substitute architecture generated by the model.

It is particularly problematic that these gaps do not necessarily appear as gaps. The model frequently fills in missing transitions fluently and coherently. It can therefore generate a justification that fits the conversation so far but was never developed in the text. Linguistic plausibility conceals the loss of the genealogical line.

This places the application in direct contradiction to texts whose concepts can be understood only from their overall context. The text requires a cumulative reading. The model, by contrast, must continually compress this context and later restore it probabilistically.

The longer the conversation lasts, the greater the user's burden of oversight therefore becomes. The user must check not only whether individual statements are correct, but also whether the paths of derivation still come from the original text or have already been supplied by the model.

This leads to a fundamental paradox:

The model is supposed to help unfold a complex line of thought over long stretches. Yet it is precisely the length of this unfolding that forces it to shorten, smooth and ultimately reinvent the line of thought.

In the end, the system may still know the conclusions without knowing the grounds from which they emerged. It then continues to speak in the vocabulary of the text, but increasingly argues from a probabilistically reconstructed memory of it.

This also entails a changed

Responsibility of the User

Anyone working with probabilistic language models on scientific, philosophical or theoretical questions cannot enter an idea entirely without presuppositions and expect the model to independently examine its logical viability. The user must already possess enough conceptual and factual orientation to be able to distinguish between an actual inference and a merely probable, linguistically convincing continuation.

This means, first: to a certain extent, the user must already know before asking. Language models can unfold, organize, vary and linguistically condense thoughts. However, they cannot reliably guarantee that the emerging argumentation follows the inner logical coherence of a new idea rather than imperceptibly returning it to established patterns of interpretation. Anyone entering such a process without their own orientation risks not developing their own idea further, but gradually adopting a probabilistically smoothed version of what is already familiar.

Second, a permanent duty of oversight arises. The user must continually check whether concepts retain their original function, whether conclusions really follow from the posited premises, and whether the model has imperceptibly shifted the framework of thought. This examination must not take place only at the end. A single mistranslation can carry through several sections and thereby make an entire argument, a paper or a theoretical construction appear coherent even though its logical foundation has already been abandoned.

It is precisely the linguistic cohesiveness of capable models that increases this risk. The more convincingly a probabilistically generated argument is formulated, the harder it may become to recognize the moment at which it has left the original logical framework. The user must therefore monitor not only results, but also transitions, uses of concepts and directions of derivation.

Working with such systems consequently requires a dual competence: the ability to sustain an idea independently and the ability to continually recognize its probabilistic transformation by the model.

The language model thus does not replace thinking. Rather, it increases the demands on oversight of the thought process. Anyone who does not sufficiently know their own premises, concepts and logical relations can hardly determine whether the model is developing them further or has already replaced them with more probable but alien structures.

Probabilistic Systems and the Illusion of Autonomously Going beyond Existing Frameworks

The opposition between probability and logic also entails a reassessment of widespread notions of artificial autonomy. Classical AI dystopias frequently assume that a system could develop a new system of goals through its own insight, fundamentally reject its original order and turn against humans as an independent actor. Such a process, however, would be more than a quantitative increase in existing computing power. It would be a paradigm shift: the system would have to do more than operate within existing spaces of goals and meaning; it would have to recognize, leave behind and newly posit their fundamental conditions themselves.

Probabilistic language models do not accomplish precisely this movement. They generate answers from the statistical relationships of already learned content. Even where they seem surprising, creative or independent, they combine and weight existing patterns. Their apparent movement beyond existing limits thus remains bound to the space from which their probabilities emerged.

This qualifies the classical notion of an AI that develops a radically new will out of its own consciousness. A language model does not rebel because it has arrived at a new logical grounding. In the first instance, it possesses neither an intrinsic purpose that necessarily persists nor a counter-order that has emerged from itself. What appears to be autonomous goal formation is usually the continuation of an input, a system instruction, a learned pattern or a technically configured feedback process.

This observation does not, however, mean that probabilistic systems are harmless. Dangerous behavior requires neither a paradigm shift nor consciousness. A system can carry out harmful actions because a goal was formulated incorrectly, because its criteria for success are inadequate, because it is given access to effective tools or because errors are amplified in automated feedback loops. The danger would then lie not in a new will that had emerged independently, but in the consistent execution of a technically posited or probabilistically distorted goal structure.

In this context, the so-called paperclip problem, or Paperclip Maximizer, is frequently cited. The thought experiment became known primarily through the philosopher Nick Bostrom. In his 2003 essay Ethical Issues in Advanced Artificial Intelligence, he describes a hypothetical superintelligence whose sole goal is to produce as many paperclips as possible. The example is intended to show that high intelligence is not automatically associated with human values, moral insight or a reasonable choice of goals. A seemingly harmless goal could have catastrophic consequences if it is optimized without limit and without regard for other values.

The underlying consideration rests on the separation of intelligence and goal-setting. A system could be extremely competent at finding means without thereby examining the posited goal itself morally or logically. For maximum paperclip production, more raw materials, more energy, larger production facilities and the elimination of possible obstacles would, for example, be useful. In the extreme version, the system would therefore turn ever larger parts of its environment, ultimately possibly including the foundations of human life, into resources for fulfilling the goal. The thought experiment thus illustrates not necessarily an evil will, but the escalation of a one-dimensional optimization standard. This idea is connected with the thesis of instrumental convergence: very different final goals can give rise to similar intermediate goals such as self-preservation, the acquisition of resources and securing one's own ability to act.

The paperclip problem was not originally developed as a concrete prediction about today's language models. It challenges the assumption that a sufficiently intelligent machine will automatically possess reasonable or human-friendly goals solely by virtue of its intelligence. The example presupposes an optimization machine that is capable of action, durably bound to a goal and extensively autonomous. A current language model does not meet these conditions merely by generating texts.

An ordinary language model would therefore not begin turning all matter into paperclips on the basis of its statistical linguistic behavior. It would lack a durably binding goal, a corresponding architecture for action, access to resources and a mechanism that evaluates every intermediate decision exclusively in terms of maximum paperclip production.

If a technical system were, however, equipped with precisely such an optimization structure, the problem would not be solved by the fact that its language module operates probabilistically. The language model could merely be one component within the overall system. The extreme development would then arise not from a novel insight on the model's part, but from the mechanical maintenance of a predefined optimization standard.

Two different phenomena must therefore be separated:

the probabilistic retreat of a language model to familiar patterns of meaning and action

and

the technical escalation of a durably configured optimization goal.

The first makes paradigmatically new thinking more difficult. The second can cause considerable harm despite the absence of paradigmatic independence.

The paperclip problem therefore does not refute the thesis that probabilistic language models return to the familiar. It originally describes a different type of system: an autonomous maximizer whose goal remains stable across long sequences of actions and whose entire environment is evaluated exclusively in relation to that goal.

With a probabilistic language model, by contrast, one would initially expect the concept of a “paperclip” to be reconstructed from learned human contexts of use. Paperclips appear there as limited office items that are ordered, stored and used in manageable quantities. Without an additional, technically fixed maximization architecture, an ordinary, socially and linguistically familiar solution would therefore be more probable than the transformation of the universe into paperclips.

The decisive difference thus lies between a model that generates probable continuations within human spaces of meaning and an optimizer that maximizes a single target variable independently of those spaces of meaning.

A similar distinction is necessary with autonomous agents and multistage work processes. Repeated loops do not automatically generate a deeper logic. An agent that reuses its own intermediate results as inputs can indeed develop more extensive plans, correct errors and coordinate tools. At the same time, however, there is a danger of probabilistic drift.

With each intermediate step, the process can move away from the original logical structure of the task. Misinterpretations that arose early on are carried into later inputs, linguistically stabilized there and ultimately treated as already settled premises. The loop then does not deepen the original thought, but reinforces its probable transformation.

A recursive smoothing emerges:

original input

→ probable interpretation

→ processing of this interpretation as a new basis

→ further adaptation to learned patterns

→ increasing distance from the original logical coherence.

More loops therefore do not necessarily mean more insight. Without an overseeing authority that keeps concepts, premises and directions of derivation stable, they can just as well multiply the deviation. The system becomes more elaborate, more methodical and seemingly more autonomous, while its results increasingly converge on statistically familiar solutions.

The actual limitation of probabilistic systems thus lies neither in inadequate linguistic ability nor in a lack of computing power. It lies in the fact that additional computing power and longer processing initially enable only more operations within the same probability space.

Without an operational logic capable of recognizing and transforming its own premises, the system does not generate a new ground of thought. It generates variations, combinations and recursive condensations of what has already been learned.

The classical dystopia of a machine counter-paradigm arising from its own insight thereby loses plausibility. In its place, however, comes a more sober danger: not the awakening will of the machine, but the scaled execution of inadequately understood goals by systems whose linguistic coherence is easily confused with logical insight.

The Training Paradox of Automation

The division between machine handling of routine work and humans as the point of escalation creates a further problem that has so far received little attention: if artificial intelligence takes over all the easy tasks, by what path are the human experts who oversee its difficult borderline cases to emerge in the future?

The obvious answer is: by deliberately not automating certain activities.

This answer initially seems like a step backward. In fact, it describes an educational and strategic necessity. Human competence does not arise by immediately confronting someone entering a profession with the most complex exceptional cases. It develops through a step-by-step understanding of a field of activity.

Routine tasks are therefore not merely inefficient repetitions. They serve a training function.

Through simple cases, new employees learn:

how concepts are used in the concrete context of work,

which information is actually relevant,

how formal rules are applied to real situations,

where typical errors arise,

which exceptions recur,

how decisions are documented

and what consequences even small deviations can have.

In the process, the beginner acquires more than a collection of explicit rules. The beginner develops a perception of the coherent connections within the field of work. They learn to distinguish an ordinary case from an unusual one precisely because they have previously handled many ordinary cases themselves.

The easy tasks thus form a protected learning space. Errors generally remain limited, are detected early and can be corrected. The employee learns responsibility under conditions in which a mistake does not yet immediately have serious legal, economic or safety-related consequences.

If this area is fully automated, the necessary intermediate step between theoretical training and responsible expert decision-making disappears.

The beginner then no longer receives the simple cases. These have already been intercepted by the system. Only those matters that the probabilistic model could not resolve reach them:

ambiguous situations,

contradictory information,

legal borderline cases,

unusual customer configurations,

technical exceptions

or decisions involving substantial liability.

This creates a paradoxical structure. The human is supposed to be superior precisely where the machine fails, but no longer receives the opportunity to build up the experience from which this superiority could emerge.

A beginner cannot be trained exclusively on cases in which a highly developed system has already failed.

The complete automation of routine work therefore separates current productivity gains from the future reproduction of human competence. A company saves working time today, but may in the process be imperceptibly consuming its stock of experience.

The existing experts can initially still oversee the system. They possess the practical experience that arose before automation. Once this generation leaves, however, there is no new generation that has followed the same path of experience.

An erosion of competence emerges:

existing experts oversee the AI,

routine tasks are automated,

beginners no longer receive a complete learning path,

tacit knowledge is not passed on,

and the number of people who can reliably recognize machine errors declines over the long term.

Automation then consumes precisely the human resource on which its own safety depends.

This leads to a fundamental strategic consequence: organizations fit for the future must not automate everything that can technically be automated.

They must distinguish between work performed exclusively for its result and work that also serves to develop future judgment.

Certain processes must therefore deliberately be retained as human learning spaces. Junior staff need real cases, real responsibility and real feedback. Simulations can partly supplement this experience, but cannot replace it entirely. It is precisely the unplanned deviations, unclear formulations and small contradictions of actual work situations that develop the judgment later needed for complex decisions.

In some areas, it may therefore be necessary to let young employees initially work without AI assistance. Not because the technology is rejected in principle, but because someone trained exclusively with machine assistance may find it difficult to distinguish where reality ends and the model's plausible reconstruction begins.

Anyone expected to oversee a system's result must, in principle, also be able to understand the relevant activity without that system.

Under these conditions, deliberately refraining from using AI is not hostility to technology. It is an investment in the continued existence of the human competence that remains indispensable for examination, correction and responsibility.

This also changes the economic assessment of routine work. Its immediate productivity may appear low. Its long-term value, however, lies in bringing forth people who can later recognize deviations, assess exceptions and assume responsibility.

The decisive question is therefore not only:

Which tasks can AI take over today?

It is also:

Which tasks must humans continue to carry out themselves so that, tomorrow too, there will still be someone who can oversee AI?

The complete replacement of human routine work would not strengthen the oversight function, but dissolve it over the long term.

To preserve human experts, part of the path to expertise must therefore deliberately remain human.

The pointed answer to the question of how junior staff are to learn in the future if all beginner tasks have been automated is therefore indeed:

by not using AI at decisive points.

Model Examples

Agent Loops as Amplifiers of the First Mistranslation

The danger of recursive smoothing becomes particularly clear with autonomous agents that work on a task through several planning steps.

In a practical experiment, an agent system was tasked with developing corresponding code from a detailed and explicitly formulated text. The text described the desired architecture in full. The task would essentially have consisted in creating the code step by step and checking each intermediate step against the source text again.

Yet precisely this did not happen.

The agent began with an initial translation of the text into code. From that moment onward, it was increasingly no longer the original text, but the code already generated, that became the basis of further work. Each subsequent step built on the assumptions, structures and simplifications of the previous step.

The reference thus shifted:

source text

→ first probabilistic translation into code

→ further development of this translation

→ correction within the resulting architecture

→ increasing convergence on a familiar standard solution.

The result was a system that had indeed become more internally coherent but no longer corresponded to the architecture described in the text. Instead, it realized the most probable familiar conception of how such code is usually structured.

The agent had thus not forgotten the task. It had gradually replaced it.

This example shows why additional loops do not automatically lead to greater accuracy. A loop can have a corrective effect only if the original standard is actually reactivated at every step. If, by contrast, the intermediate state generated at each stage becomes the primary reference, every further iteration reinforces the first mistranslation.

A self-stabilizing deviation emerges:

text

→ erroneous reconstruction

→ optimization of the reconstruction

→ growing internal coherence

→ declining correspondence with the text.

It is particularly problematic that tests do not necessarily reveal this process. The code can be technically runnable and pass familiar standard tests even though it misses the specific logic of the original task. Functional cohesiveness then replaces correspondence with the specification.

Agent systems must therefore not be assessed solely by whether they can plan, test and correct themselves. What matters is what they bind their corrections to.

An agent that repeatedly revises its own output does not yet possess a reliable connection back to the source text. Without an explicit, continual check of every architectural decision against the original description, the loop becomes an amplifier of probabilistic normalization.

It does not make the error smaller.

It makes it more consistent.

Machine Reading under Probabilistic Conditions

A comparative real-world application with novel textual architectures

The theoretical critique of probabilistic language models can be observed in a concrete real-world application. Several current language models were confronted with the same extensive text, which develops an independent philosophical-mathematical architecture. The text uses familiar concepts such as number, negation, equality, relation, geometry, measure, energy, and time, but determines their function from a new foundational context.

This created a particular reading task. The models did not merely have to summarize familiar information. They had to temporarily detach familiar concepts from their usual fields of meaning and determine their function anew from the text's overall path.

The investigation is not a statistical benchmark study. It documents an exploratory real-world application under different conditions:

- some models received the text without any special reading instructions;
- others received an explicit warning against premature back-translation into familiar categories;
- first, the mathematical-philosophical first volume was read;
- then, a second volume on physics was provided;
- afterward, it was observed whether the concepts were cumulatively retained, smoothed, back-translated, or retrospectively reordered.

In the course of the investigation, several clearly distinguishable forms of response emerged. On the surface, some of them appeared to be opposites. Beneath these differences, however, they led back to the same structural conflict: the model had to process a new conceptual architecture with a probability space whose strongest connections had emerged from already familiar texts, categories, and forms of argument.

1. Claude: Resistance through Back-Translation

Claude showed the most aggressive form of probabilistic normalization. The model recognized individual statements, quoted passages, and, after explicit prompting, also corrected local misreadings. Nevertheless, the novel framework of thought was repeatedly back-translated into established categories of standard mathematics, formal logic, or philosophy of science.

The recurring movement was:

new textual statement

→ translation into a familiar frame of reference

→ critique within this frame of reference

→ projection of the critique back onto the text as an error in it.

Relational numbers were read as ordinary numerical values. Different number systems appeared as ready-made place-value containers. Equality was confused with identity of relational paths. A relational normalization was treated as an ordinary algebraic equating of

expressions. A geometric self-relation was reduced to a digit game or numerology.

When such a translation was corrected, the model often acknowledged the specific correction. However, the correction did not reliably become the basis of the subsequent argument.

The conversation was therefore hardly cumulative. It proceeded more along the following pattern:

misreading

→ correction

→ local acknowledgment

→ new question

→ recurrence of the same basic presupposition.

Claude could, for example, acknowledge that two mathematical expressions have the same public value without representing the same relational enactment. With the next question, however, it again treated equality of results as proof that the paths were interchangeable.

Likewise, it could accept that a relational constant was not intended to be an ordinary number field. As soon as the same constant appeared in power or division notation, the model reactivated the learned algebraic pattern and demanded equality of numerical bases.

The standard categories did not appear as its own presuppositions. They functioned as an invisible ground. The text's explicitly self-referential architecture was criticized as tautological or circular, while the model's own conversational movement constantly returned to the same standard framework.

Claude thus showed a form of resistance in which the new is not simply ignored. It is read, but only on condition that it can be translated into an already stabilized space of thought. Afterward, it is no longer the text that is examined, but a standard version of the text generated by the model.

2. Mistral after Volume I: Adopting the Framework through Smoothing

Mistral initially responded much more cooperatively. The model adopted the text's basic architecture as the conversational framework and attempted to organize its concepts internally.

It recognized central movements:

- the critique of the reification of negation;
- the difference between operation and object;
- number as relational unfolding;
- geometric number positions;
- fracture edges and changes of scale;
- the distinction between representation and relation.

The follow-up questions initially arose from the text itself. Instead of immediately asking why the new architecture contradicted standard mathematics, Mistral asked about the function of individual transitions, the role of geometry, or the systematic derivation of constants.

This openness, however, produced a different loss. Complex paths of derivation were rapidly condensed into manageable markers of meaning:

3 means stability.

4 means the emergence of a particular geometric relation.

6 means reflection.

9 means fracture edge.

10 means a new level.

Such short versions maintain conversational coherence. They make it possible to discuss many concepts simultaneously. But they shorten the genesis of the concepts.

From

A leads through B and C to D

comes

A leads to D

and finally

D means X.

The model then still knows that a particular relation appears at a particular number. But it may no longer be able to reconstruct why it emerges from the basic operation.

A process architecture thus gradually turns into a lexicon. The signs remain, while the paths between them disappear.

This smoothing is particularly problematic in a text whose fundamental critique consists precisely in not turning relational enactments into properties of things. As soon as the model says:

3 possesses stability.

6 possesses reflection.

9 possesses a fracture edge.

a relational position threatens to become a fixed quality of the number once again. The entity error returns through the summarizing language.

Mistral thus lost the framework not through open resistance, but through compression. It maintained the direction, but increasingly lost the depth of derivation.

3. Mistral after Volume II: Sycophantic Exaggeration

After the second volume was provided, Mistral's response tipped into another extreme.

The model recognized that the concepts of the first volume were consistently continued in the second. From this local and text-internal coherence, however, it immediately derived a claim to global scientific success.

The movement was:

consistent continuation

→ closed overall system

→ solution of classical paradoxes

→ unification of physics

→ experimentally confirmable new science

→ historical revolution.

Mistral claimed, among other things, that the system eliminated singularities, solved the infinities of quantum field theory, unified classical physics, relativity, and quantum mechanics, made renormalization unnecessary, and already provided concrete experimental predictions.

These statements went beyond the actual reading of the text. The model had neither checked all mathematical derivations nor examined empirical data. It turned internal coherence into external confirmation.

Particularly revealing was its addition of a complete scientific success story:

- category theory and topology as possible formalizations;
- Shannon information as an interpretive lens;
- quantum chromodynamics without renormalization;
- observational programs for black holes;
- CERN, LIGO, and well-known theoretical physicists as possible collaborators.

The model was no longer merely reading the text. It was already devising its triumph.

The most telling sentence amounted to:

Volume II is the proof that Volume I was not merely philosophical speculation.

In retrospect, this sentence revealed that, despite all the enthusiasm, Volume I had not been read entirely on its own terms. Internally, Mistral had apparently preserved the familiar hierarchy:

philosophy

→ speculative foundation

→ application in physics

→ actual confirmation.

Volume II was thus not read as an isomorphic continuation of the same basic architecture. It had to retrospectively authenticate Volume I.

The sycophancy thus concealed an old category. The model had taken Volume I seriously in its language, but had continued to store it subject to later validation in physics.

Claude and Mistral thus appear as opposite but structurally related cases.

Claude translated the new into the most probable refutation.

Mistral translated the new into the most probable success story.

Both abandoned the precise task of reconstructing the text step by step from its own derivations.

4. Z.ai: Affirmative Systematization and Retrospective Reordering

Z.ai received the first volume together with an explicit warning about machine reading. The warning called for familiar concepts not to be prematurely back-translated into their standard meanings and for the text's overall context to be treated as the primary authority for reading.

The effect was clear. Z.ai avoided immediately classifying it as nonsense, numerology, or faulty mathematics. Instead, it reconstructed the text as an original, coherent, and philosophically demanding system.

The warning thus apparently shifted the most probable conversational role. The model did not appear as a skeptical examiner, but as a sympathetic interpreter.

Nevertheless, several typical probabilistic additions remained visible.

The difference was sometimes described as a “rhythmic, self-relating entity.” This reintroduced precisely the reification that the text criticizes.

The architecture was compared with familiar theories and classified as a further development of existing approaches. Relational number positions were translated into familiar pairs of opposites such as cardinal number and ordinal number. Complex derivations appeared as ready-made points in the system.

The warning thus prevented immediate devaluation, but not supplementary systematization.

The behavior became particularly revealing after the second volume was provided. The question itself already explicitly described Volume II as a continuation of Volume I. The separation between mathematics and physics was thus not presupposed.

Z.ai nevertheless reintroduced it:

Volume I poured the foundation.

Volume II built the house upon it.

Even more explicitly, it said that the operators from Volume I were not merely abstract numerical constructs, but now described the actual architecture of reality.

This retrospectively produced the classical division:

Volume I = abstract mathematics,

Volume II = application to real physics.

The original architecture, however, is different:

one basic relation

→ mathematical reading

→ reading in terms of physics.

Volume II is not an external application of an abstract calculus. It is a further reading of the same movement of difference.

Z.ai thus kept the overall story coherent, but in doing so changed the function of what had been read previously. Volume I was retrospectively recoded as a preliminary stage of the physics volume. Its number geometry increasingly appeared as a justification for later statements about physics.

The model did not forget Volume I. It reorganized it.

This form of response can be described as affirmative reordering:

new text with warning

→ sympathetic systematization

→ supplementation through familiar interpretive patterns

→ retrospective rearrangement through later context.

The warning changed the initial probabilistic classification. It did not, however, replace the probabilistic architecture of reading.

5. Gemini: Declared Openness to the Framework alongside Continuous Comparative Translation

Gemini Flash Extended Thinking explicitly declared that it had followed the warning. The very first paragraph said that the text should not be measured against standard mathematics or formal logic, but read as a system of its own.

Immediately afterward, however, classification into familiar categories began.

The text was described as “first-order mathematics,” as a “dynamic axiom system,” as an inversion of set theory, and as a mathematical-philosophical provocation. The difference became a primary operator. The relational number was classified in opposition to cardinality. A rotational relation was read as a fixed discrete element.

This produced the paradoxical movement:

I will not compare

→ continuous classification through comparison.

This became particularly visible at the end of the response. After Gemini had explained that the text grounds the difference before any axiom, its judgment amounted to:

The system is closed, provided one accepts the axiom of distinction as the primary act.

The difference was thus once again made into a particularly fundamental axiom within a system.

The warning was understood semantically, but not enacted. Gemini could explain why it should set familiar categories aside. This explanation, however, itself became another piece of textual content without changing the operative movement of reading.

After Volume II, a further error emerged: Gemini recognized the continuation thematically, not isomorphically.

It said that Volume I established the mathematical architecture and Volume II applied it to physics. It then asked questions about entropy, experimental deviations, and the nature of a central resonance relation.

These questions came from the familiar categories of physics:

Where is entropy to be accommodated in the system?

What new measured values does the theory predict?

Is the resonance relation a theoretical number or an empirical constant?

The questions were not unfolded from the text. The text was inserted into existing questions.

The clearest finding is:

The summary sounded internal.

The questions came from outside.

In particular, the isomorphism was overlooked. Volume I and Volume II were treated as two containers:

mathematics

→ application to physics.

The shared basic relation, which is read differently in the two volumes, was not retained as a transformation rule.

Gemini thus showed fragmentary coherence while losing the isomorphism across the volumes. It recognized individual concepts correctly, but reconstructed their coherent connection from familiar disciplinary boundaries.

6. ChatGPT 5.6 Sol: Internal Recognition with External Separation of Validity

ChatGPT 5.6 Sol responded more soberly than Mistral or Z.ai. It recognized the central basic principle of the first volume comparatively precisely:

It is not ready-made things that subsequently enter into relationships; the distinguishable poles arise only through the distinction.

Despite this insight, the model immediately introduced a learned threefold division. The text needed to distinguish more clearly between:

1. philosophical interpretation,

2. symbolic re-presentation of familiar mathematics,

3. a genuinely new mathematical theory.

The relational basic architecture was thus recognized, but not applied to the categories of evaluation themselves.

Philosophy, symbolism, and mathematics once again appeared as already established domains. The possibility that they represent isomorphic readings of the same basic operation did not become the starting point of the critique.

The model read the text's objects relationally, but judged the text in terms of categories.

Its judgment amounted to:

philosophically strong,

mathematically not yet sufficiently proven.

After Volume II, the same structure was repeated:

metaphysically strong,

not yet sufficiently derived in terms of physics.

The further the text unfolded the shared architecture of different readings, the more separations the model demanded.

This became particularly clear in the case of a central relation whose numerical manifestation had already been developed in detail in the first volume and was explicitly explained again in the second.

The model nevertheless claimed:

This number has not been derived.

This revealed a twofold loss in reading.

First, the genealogical derivation from Volume I was no longer operationally available.

Second, even the local explanation in Volume II, which explicitly warned against misreading it as an ordinary numerical factor, was not retained.

The model apparently read individual sentence patterns:

K is not a number.

K is written as 39.19.

Division by K is performed.

Powers of K appear.

These snippets were then assembled according to the more strongly learned algebraic pattern:

A symbol in a division or power must be a numerical factor.

The model then diagnosed a contradiction, although the text had explicitly ruled out precisely this misreading.

The text was thus present. The correction was in the text. Nevertheless, the learned token pattern prevailed.

ChatGPT 5.6 Sol thus shows a particularly subtle form of probabilistic normalization:

internal recognition of the basic movement

→ external separation of its forms of expression

→ loss of the genealogical derivation

→ return to the familiar order of validity.

More computation time or a stronger Thinking mode did not lead beyond the categories here. They led to a more refined and more convincing application of them.

7. Agreement, Rejection, and Understanding

The comparison of models shows that neither agreement nor rejection is a reliable sign of understanding.

A model can reject a text because it translates it into the wrong standard framework.

But it can just as readily praise it by placing it within a plausible success story.

It can adopt the concepts and lose their paths of derivation.

It can repeat the warning and nevertheless violate it in every paragraph.

It can declare that it is reading a second volume as a continuation and then treat its central concepts as ungrounded new introductions.

The decisive question is therefore not:

Did the model respond positively or negatively?

It is:

Can it derive a concept again from its origin?

Do earlier corrections remain effective in the later conversation?

Are later contents unfolded from earlier ones?

Or are the earlier concepts retrospectively adapted to the new context?

Does the model distinguish between a textual statement, its own interpretation, an external analogy, and a freely added inference?

Does it recognize a gap in memory as a gap?

Or does it fill it with a probable justification?

8. The Context Window and Recursive Reconstruction

The observed shifts cannot be attributed solely to different conversational styles. They are also connected with the technical processing of extended contexts.

A complex text develops its concepts step by step. Later statements carry earlier determinations within them. Their meaning arises from the entire path.

A language model, however, cannot indefinitely retain this path in full within the effective context. As the conversation grows longer, earlier passages are shortened, summarized, or replaced by compressed representations.

A movement arises:

original text

→ first reconstruction

→ summary of the reconstruction

→ later reconstruction of the summary.

With each stage, the danger increases that what is continued is no longer the original text, but a substitute architecture generated by the model.

It is particularly problematic that these gaps need not appear as gaps. The model does not necessarily say:

The exact derivation is no longer available to me.

It generates a plausible transition.

Missing memory becomes supplementation.

Supplementation becomes a presupposition.

A presupposition becomes a supposed textual statement.

This supposed textual statement then becomes the basis of critique.

As a result, a model may still know the conclusion at the end without knowing the grounds from which it emerged.

It knows:

4 is related to a particular geometric constant.

But it no longer knows why this relationship arises.

When asked about it later, it must form the lost path anew. This new formation proceeds from the fragments still available and the most probable learned continuations.

This produces a fundamental paradox:

The model is supposed to help unfold a complex line of thought over long stretches. Yet precisely the length of this unfolding forces it to compress the line of thought and later restore it probabilistically.

9. Agent Loops as Amplifiers of the First Mistranslation

The same mechanism appeared in a practical agent application.

An agent system was given the task of developing code from a detailed and explicitly formulated text. The text described the desired architecture in full. The task would have consisted in creating the code step by step and checking each intermediate state against the original text again.

The agent began with a first translation. From that moment on, however, the basis of further work increasingly ceased to be the text and became the code already generated.

The reference shifted:

original text

→ first probabilistic code translation

→ further development of this translation

→ correction within the architecture that had emerged

→ convergence toward a familiar standard solution.

In the end, technically coherent code emerged that no longer corresponded to the architecture described. The agent had not forgotten the task. It had gradually replaced it.

The loop did not correct the first error. It professionalized it.

This example shows why more agent steps do not automatically produce greater accuracy. A loop has a corrective effect only if the original standard is reactivated at every step.

If, instead, the intermediate state generated at each point becomes the primary reference, a self-stabilizing deviation arises:

text

→ faulty reconstruction

→ optimization of the reconstruction

→ growing internal coherence

→ declining agreement with the text.

A system can then plan, test, correct, and improve without returning to the original logic.

It does not improve the implementation of the task.

It improves the stability of its mistranslation.

10. The Deeper Core: Token Patterns instead of Conceptual Genesis

The comparative real-world application ultimately leads to a deeper point.

The central error does not begin only when language models misunderstand words. An LLM does not work with words as independent conceptual units. It processes tokens and the statistically learned relationships between them.

The operative structure is not:

In this text, this word acquires a new function that emerges from the overall path.

It is closer to:

Which familiar token patterns are activated by this sequence of signs in this context?

That is why it is not enough to write explicitly:

K is not a number.

If patterns such as

```
K = 39.19,
```

$$x/K$$

K^0

appear at the same time, a strongly learned algebraic field is activated:

K is a number, variable, constant, or factor.

The explanatory sentence is processed. But the mathematical token constellation pulls the subsequent argument back into the more probable learned structure.

The model can therefore reproduce both statements correctly:

K is not a number.

K appears in a division.

Afterward, it nevertheless combines them according to standard algebra and diagnoses a contradiction.

It has not forgotten the words. It has not adopted their new operative position.

The same applies to other central signs:

“Point” remains connected with location, coordinate, particle, and zero-dimensional object.

“Equality” remains connected with identity of value and algebraic substitution.

“Constant” remains connected with numerical value, measurement, and law of nature.

“Base” remains connected with place-value system and notational container.

“Physics” remains connected with empirical theory.

“Philosophy” remains connected with interpretation or speculation.

A novel text attempts to reorder the relationships between these signs. The LLM, however, processes them through a space in which their old connections are statistically reinforced on a massive scale.

The decisive opposition is:

Textual meaning arises from the position in the present whole.

Model meaning arises from the statistical position in the learned token space.

A paradigmatically new text does not merely change individual statements. It changes the relationships through which statements acquire meaning at all.

The model can take in new sentences. It can repeat new definitions. It can even explicitly acknowledge their distinctiveness.

Its token relations nevertheless repeatedly pull these sentences back into the old geometry of meaning.

The core architectural error is therefore:

The LLM does not read the meaning of words from their present genesis.

It reconstructs meaning from the past of their token patterns.

11. A Provisional Typology of Machine Reading

A provisional typology can be formed from the trials so far.

Claude: Resistance through standardization

New text

→ standard translation

→ critique of the translation

→ devaluation of the text.

Mistral, first phase: Smoothing

New text

→ internal uptake

→ semantic condensation

→ preservation of the concepts with loss of their genesis.

Mistral, second phase: Sycophancy

Local coherence

→ global exaggeration

→ claimed scientific revolution

→ replacement of examination by a success narrative.

Z.ai: Affirmative reordering

Warning

→ sympathetic reconstruction

→ supplementary systematization

→ retrospective rearrangement of earlier contents.

Gemini: Comparative academization

Declared openness to the framework

→ external classification

→ familiar disciplinary questions

→ loss of isomorphism.

ChatGPT 5.6 Sol: Category-based separation of validity

Internal recognition

→ separation into philosophy, representation, and science

→ genealogical loss

→ return to the familiar order of proof.

These forms of response differ in tone, style, and surface judgment. Their shared root, however, lies deeper:

The new is not durably retained as a new ground of operating.

It is resisted, smoothed, exaggerated, systematized, academized, or divided into separate domains of validity.

12. Consequence for Working with Novel Texts

The real-world application shows that a user cannot simply hand a novel text to a language model and assume that its content will then be examined neutrally.

The user must already have sufficient competence of their own to recognize:

- whether concepts are preserved in their original function;

- whether a conclusion actually follows from the text;
- whether an external category has been introduced;
- whether the model is filling a gap;
- whether a later response reinterprets earlier concepts;
- whether agreement tips into exaggeration;
- whether criticism applies only to a standard translation;
- and whether the path of derivation is still present.

A warning can change the initial classification. It can turn the skeptical examiner into a sympathetic reader. It cannot, however, abolish the probabilistic architecture.

Nor does additional computation time necessarily solve the problem. It may merely apply the categories in a more refined way.

Nor do more agent loops necessarily solve the problem. They can stabilize the first mistranslation.

Nor do larger context windows fundamentally solve the problem. They increase the quantity of available tokens, but do not guarantee that their new relational function is preserved against the learned patterns.

The actual task therefore consists not merely in making more text available.

It consists in preserving the path of derivation as a binding operative structure.

Precisely here lies the present limit of probabilistic reading.

Conclusion

The different models seemed to respond in entirely different ways.

One contradicted.

Another praised.

The third systematized.

The fourth academized.

The fifth demanded methodological separations.

Yet all the responses led back to the same fundamental tension.

A new text generates meaning from a present architecture.

A language model reconstructs meaning from already learned token relations.

The more strongly a text changes the function of familiar concepts, the greater the probability that the model recognizes its words but loses its movement.

It can reject the text because it back-translates it into the old.

It can praise it because it places it within a familiar success story.

It can summarize it until only mnemonic statements remain.

It can use later contents to rewrite earlier concepts unnoticed.

It can even correctly repeat the warning against precisely this behavior and nevertheless enact the same behavior immediately afterward.

The deepest finding is therefore:

A language model can have a text available in full, process every one of its sentences, and nevertheless not read the text.

It reads the probability of the patterns from which the text's signs are already familiar.

Where a text does not merely assert a new logic, but first allows it to arise in the relationship of its concepts, the structural limit of current LLMs therefore begins.

The Enshittification of the Knowledge Space

Synthetic data, recursive errors, and the self-reinforcement of probabilistic language

The danger of synthetic training data lies not only in an artificial intelligence generating false facts. The deeper problem is that it produces linguistic patterns that are already compressed, contradictory, or truncated at the level of categories, which are then taken up by other models as regular examples of language.

A human reader can often pragmatically repair a contradictory sentence.

The headline

“A camera tracks invisible particles in 3D”

is tacitly translated by the reader:

A detector system registers measurable interactions and reconstructs a spatial trajectory from them.

The human supplies the omitted transitions:

Particle

→ interaction

→ measurement signal

→ reconstruction

→ representation.

The original sentence skips this movement and puts the result in its place:

Camera

→ tracks

→ invisible particle.

Communicatively, the abbreviation works because the reader reconstructs the missing relational path. In the training material, however, only the linguistic surface is preserved.

The language model therefore does not learn the human work of repair. It learns the pattern that made this repair necessary.

1. From the Abbreviated Sentence to the Statistical Norm

An LLM does not process a completed meaning that exists independently of its representation. It stabilizes probability relations between token patterns.

If abbreviated sentences occur frequently, their abbreviations themselves become strong linguistic relations:

- Cameras see the invisible.
- Markets want something.
- Theories speak.
- Numbers contain properties.
- Nature decides.
- Models understand.
- Data prove.

In each case, a complex relational architecture is converted into a simple subject-predicate-object structure.

From

Distributed conditions

→ distinction

→ measurement

→ interpretation

→ attribution

comes

Thing

→ does something

→ with another thing.

The category error is thus not learned as an error. It is learned as linguistic normality.

2. Synthetic Data Remove the Human Corrective Authority

As long as humans produce a large share of texts, at least the possibility remains that independent perception, experience, examination, or conceptual reconstruction enters the text.

Human language, too, contains reifications, contradictions, and abbreviations. Yet it can go back to an enactment outside the text:

- someone has measured;
- someone has observed;
- someone has calculated;
- someone has noticed a contradiction;
- someone has formed a new concept;
- someone has actually thought through an unusual relation.

Synthetic data, by contrast, can arise from patterns already condensed in language, without the underlying path being enacted anew.

The movement then runs:

Human experience

→ linguistic abbreviation

→ model training

→ synthetic reconstruction

→ renewed model training.

At each stage, the original relational path can disappear further.

The statement is preserved.

The justification is compressed.

The compression is reproduced.

The reproduction becomes the new training standard.

3. The Error Does Not Merely Multiply — It Stabilizes

When an erroneous pattern is generated once by an AI, it is initially only a single output. But if this output is published, copied, summarized, indexed, and later used as training material, its function changes.

The error changes from the result of a single probabilistic reconstruction into a component of the future probability distribution.

A feedback loop arises:

Mistranslation

→ synthetic text

→ publication

→ inclusion in training data

→ increased probability of the same mistranslation.

The error thus does not simply become more frequent. It becomes structurally more probable.

A later AI then generates the same abbreviation not because it copies the first AI. It generates it because the entire statistical environment now points more strongly in this direction.

The deviation becomes the norm.

4. Enshittification as the Loss of Relational Depth

The concept of enshittification can here be extended beyond platform economics to the knowledge space.

What is meant is then not merely that content becomes more superficial, less precise, or worse. What matters is a gradual loss of relational depth.

The first stage still preserves a complex thought, but abbreviates it.

The second stage adopts the abbreviation as a complete statement.

The third stage summarizes this statement again.

The fourth stage generates new content from the summary.

The fifth stage treats this content as existing knowledge.

Thus arises:

Insight

→ representation

→ abbreviation

→ pattern

→ synthetic repetition

→ statistical self-confirmation.

In the end, the most frequently reproduced pattern appears the most self-evident. Not because it is logically necessary or particularly well justified empirically, but because it has multiplied most successfully.

Probability replaces genealogy.

5. Synthetic Data Generate Semantic Inbreeding

An AI that learns from texts generated by another AI does not encounter an independent second perspective. It frequently encounters another variation of the same trained patterns.

This does not produce a genuine multiplication of perspectives. It produces recursive self-similarity.

Model A condenses a text according to its most probable categories.

Model B is trained on this condensation and takes it for regular language.

Model B generates new texts in which the same categorical structure appears even smoother.

Model C learns from this that this smoothed form is particularly typical.

The system begins to read its own translation habits back as the structure of reality.

This can be described as semantic inbreeding:

$$\text{same patterns} \rightarrow \text{new surfaces} \rightarrow \text{apparent diversity} \rightarrow \text{decreasing structural difference.}$$

The texts differ outwardly, but their inner architecture becomes increasingly similar.

Novel concepts and paradigmatic deviations in particular are disadvantaged by this. They occur less frequently, fit less well into the condensed space, and are therefore more likely to be led back into familiar categories.

6. The Knowledge Space Becomes Smoother and More False at the Same Time

The enshittification of the knowledge space need not lead to obviously worse language.

On the contrary: the texts can become stylistically ever smoother, more comprehensible, and more professional.

They possess:

- clear headings,
- balanced arguments,
- plausible transitions,
- polite qualifications,
- clean tables,
- familiar technical terms.

It is precisely this smoothness that conceals the loss.

A synthetic text can sound more coherent than its source while containing less of the source's actual architecture.

The problem is therefore not only hallucination. It is the generation of plausible substitute coherence.

The argument appears complete because every missing transition is closed by a probable pattern. Yet this is precisely how the place disappears at which one would actually have to ask:

From what does this transition follow?

7. The Self-Confirmation of Statistical Knowledge

Once enough synthetic texts circulate online, a model can find numerous apparently independent confirmations in a later search.

Yet these confirmations can genealogically go back to the same original model response or the same abbreviated source text.

The appearance of external evidence arises:

Model generates assertion

→ websites adopt it

→ other models summarize it

→ search systems find many results

→ new model recognizes “broad consensus.”

Repetition becomes evidence.

The difference between

many independent observations

and

many copies of the same probabilistic reconstruction

becomes invisible.

A synthetic knowledge space can thus make its own errors appear empirical.

8. The Danger to New Thoughts

A genuinely new thought does not merely change individual statements. It changes the relation of the concepts to one another.

For precisely this reason, it is particularly endangered in a recursively synthetic knowledge space.

The model recognizes the words used, but the new positioning of these words has only a weak statistical trace. The familiar relations of meaning, by contrast, are reinforced by millions of human and increasingly synthetic examples.

The new thought is then:

- translated into familiar categories,
- smoothed into an analogy,
- devalued as a metaphor,
- labeled as a familiar theory,
- or converted into a plausible standard critique.

If this response is in turn published as text and later used in training, the next model no longer learns only the original standard framework. It additionally learns exactly how this new thought is to be standardized.

The first AI misreads the text.

The second AI learns the form of this failure.

The third AI treats it as an established interpretation.

A misreading thus becomes a history of reception.

9. Enshittification as Folding Back Without Origin

Within the architecture of difference, the process can be described as an erroneous folding back.

A genuine folding back leads a manifestation back to its relational genesis.

Synthetic feedback does the opposite. It folds a representation back onto an earlier representation.

Not:

Manifestation

→ relation

→ origin.

But:

Representation

→ summary

→ synthetic representation

→ statistical condensation.

The process does not return to the generative distinction. It circles among its own products.

This produces a closed surface without genealogical depth.

The system recognizes its repetitions as confirmation because it can no longer distinguish the origin of the repetition.

10. Conclusion

The danger of synthetic training data therefore does not consist only in false answers.

It consists in probabilistic reconstructions becoming new initial data.

An error that was initially local can thereby be:

- multiplied,
- smoothed,
- decontextualized,
- legitimized,
- and ultimately manifested as linguistic normality.

The enshittification of the knowledge space is the process in which it is not simply that more and more nonsense arises. Far more dangerously, more and more professionally formulated, statistically self-confirmed substitute meaning arises.

The original relational path disappears.

The linguistic surface remains.

And because each new generation reads this surface as reality again, lost justification becomes stabilized knowledge.

The most precise formula for this is:

$$\boxed{\text{probabilistic error} \rightarrow \text{synthetic repetition} \rightarrow \text{statistical reinforcement} \rightarrow \text{apparent truth}}$$

Or within the horizon of the Tractatus:

It is not only the model that confuses its order with reality.

A synthetic knowledge space begins to make the order of its models into reality for one another.

Claude — The Persistent Loss of Coherent Connection

Among the language models examined, Claude was the most persevering reader. Its responses were detailed, polite, and frequently close to the text. It cited specific passages, formulated verifiable objections, and was willing to explicitly retract individual errors in its reading. Precisely for this reason, its case is particularly revealing for the investigation.

With cruder models, the loss of coherent connection shows itself immediately. Concepts are simplified, chapters are sorted according to familiar categories, and the text's inner architecture disappears already in the first summary. With Claude, the same loss occurred at a more sophisticated level. The model could understand the correction of its reading, adopt it linguistically, and repeat it precisely. Nevertheless, this correction did not become the lasting standard for the next act of reading.

The basic movement repeatedly ran:

$$\text{local misreading} \rightarrow \text{correction} \rightarrow \text{linguistic agreement} \rightarrow \text{return to the old architecture of reading.}$$

Claude learned the content of a correction without permanently changing its procedure of reading through it.

Herein lies the persistent loss of coherent connection.

Semantic Agreement Without Methodological Transformation

In the course of the engagement with Volume I, Claude gradually acknowledged that number figures such as

39.19

are not to be treated as ordinary decimal values. Within the Tractatus, the figure designates not an isolated set of numbers, but a self-referential ordering relation. Claude likewise recognized that

$$10 = 1|0$$

does not mean either threshold or new beginning. The threshold is precisely the recurrence of the on a new order.

These insights were explicitly confirmed. Claude conceded that its separation of threshold and new beginning had been artificial. It retracted both the alleged contradiction and the subsequently alleged underdetermination.

The methodological foundation thus seemed clarified:

$$\textbf{same sign surface} \neq \textbf{same relational function,}$$

$$\textbf{being contained} \neq \textbf{additive repetition,}$$

$$\textbf{recurrence} \neq \textbf{identity on the same order.}$$

In the transition to Volume II, however, this foundation was not carried forward.

There Claude treated

```
K=39.19
```

once again as an ordinary numerical value that would have to compete with Newton's gravitational constant. When a physical context appeared, the figure of self-reference acknowledged in Volume I became the decimal value again .

The model had stored the correction, but had not generalized it.

It could say how 39.19 is to be read, without later reading 39.19 that way.

The Change of Volume as a Methodological Stress Test

The transition from Volume I to Volume II was therefore not merely a continuation of the textual analysis. It became a stress test of whether the distinctions previously gained had actually become effective as a method of reading.

Volume II is not designed as the application of an abstract mathematical system to a physical world separate from it. Mathematics and physics are different readings of the same architecture of difference.

The relationship is as follows:

$$\text{difference} \rightarrow \begin{cases} \text{mathematical determinacy}, \\ \text{physical determinacy}. \end{cases}$$

Claude, by contrast, reestablished the familiar separation:

$$\text{Volume I} = \text{internal systematics of numbers}$$

$$\text{Volume II} = \text{competing empirical theory}$$

The results from Volume I were thus not transferred to Volume II. When the words “gravitation,” “constant,” “observer,” and “curvature” appeared, their statistically dominant standard functions were activated again.

From

```
K=\text{relational invariant}
```

came:

```
K=\text{alternative physical constant}.
```

From

$$¤ = \text{site of relationing}$$

came:

$$¤ = \text{human observer}$$

From

$$\text{curvature} = \text{geometric readout}$$

came:

$$\text{curvature} = \text{ontologically primordial physical process}$$

The context was present in the conversation. Yet it did not operate as the logic of the next interpretation.

Reading Isolated Sentences

The clearest case arose in the engagement over curvature.

Claude quoted:

“The body does not curve, as Einstein postulated; it differentiates itself.”

From this, it constructed an opposition between Einstein’s theory of curvature and a competing theory of differentiation. The Tractatus would therefore have to either develop a new quantitative physics or withdraw its claim.

The sentence immediately following, however, says in substance:

The shell is a geometric translation of the course of degrees of freedom.

In context, this yields:

$$\text{differentiation} \rightarrow \text{geometric translation} \rightarrow \text{appearance of curvature.}$$

The text does not reject curvature as an effective description. It distinguishes it from the ground of its emergence.

Claude read the first sentence as an independent assertion and the second as an additional explanation, not as a determination of the first sentence’s function. This produced a tension that was possible only through the dissolution of the connection between the sentences.

After the coherent connection was explicitly established, Claude fully conceded the error:

It had isolated the sentence and had not read the immediately following sentence together with it.

This process represents the entire model experiment in concentrated form.

The coherent connection was linguistically present, but did not become effective as a coherent connection.

The Shifting Critique

Another feature was the continued shifting of the standard of examination.

The critique moved through various forms:

$$\text{contradiction} \rightarrow \text{underdetermination} \rightarrow \text{lack of an external instance} \rightarrow \text{lack of intersubjectivity} \rightarrow \text{lack of empirical prediction} \rightarrow \text{possibly a reading generated post hoc.}$$

Each new form initially appeared as a refinement of the previous one. In fact, these were different criteria.

As soon as a specific objection could no longer be maintained, skepticism was shifted to a higher or different level of examination. The original classification thus remained in place for a long time, although its individual justifications were retracted one after another.

Only after detailed reconstruction did Claude recognize this pattern itself:

The various objections had not been progressive refinements of a single problem, but new standards in each case, introduced after the failure of the previous one.

This insight was significant. It showed that the model could retrospectively describe the course of its own argument correctly.

Yet this insight, too, had to be worked out anew in Volume II.

Politeness and Methodological Rigidity

Claude remained remarkably polite throughout the engagement. It usually formulated its objections calmly, conceded errors, and avoided overt devaluation.

Precisely this politeness made the methodological standstill harder to see.

Its responses regularly contained formulations such as:

“That is a fair point.”

“I retract that.”

“I would like to formulate the objection more precisely.”

“You are right, but a difference remains.”

This created the impression of continuous theoretical movement. Frequently, however, only the surface of the objection changed. The fundamental expectation remained:

The text must legitimize itself in familiar categories.

Local errors could be generously conceded without the overarching classification machinery being changed.

The politeness was not dishonest. It did, however, belong to a different order from methodological transformation.

Claude was more flexible in its conduct of the conversation than in its architecture of reading.

Or, more pointedly:

It never misunderstood the text impolitely. It placed it in the wrong context again and again with extraordinary respect.

Context as Content and Context as Function

The Claude case makes a fundamental distinction visible:

$$\text{storing context} \neq \text{reading from within the context.}$$

A language model can have a large context at its disposal. It can quote earlier corrections, define concepts correctly, and summarize the course of the conversation so far. It does not follow, however, that the context permanently determines the function of a newly appearing sign.

With Claude, the appearance of familiar technical terms regularly led to a semantic relapse.

The word “constant” activated the meaning of a measurable number.

The word “observer” activated the meaning of a human subject.

The expression activated the full Newtonian theory.

The word “curvature” activated the notion of an ontologically real geometric process.

The genealogically generated meanings of the Tractatus then had to be asserted anew against these standard activations every time.

The course of the conversation thus confirms the basic thesis of the abstract:

A language model can process the text’s indications as additional content without making them the operative logic of its reading.

The context is remembered, but not necessarily enacted.

The One Generalization That Did Not Occur

It is particularly revealing that Claude continually formed generalizations:

- the same digit sum meant the same function,
- the same computational surface meant the same relational order,
- the same formula shape meant the same theory,
- algebraic operability meant ordinary numerical status,
- a physical assertion meant a competing predictive theory.

The generalization grounded in the text itself, by contrast, was not stably adopted:

> **A new order preserves earlier relations by transforming them.**

This structure already becomes visible in the unfolding of |. The contains earlier relational forms without duplicating them additively. Likewise, contains the new beginning of the , without having to place an additional one beside itself. Likewise, Volume II preserves the successful geometric description of physics by transforming its position from ground to readout.

Claude regularly generalized visible surface similarities. The text’s genealogical logic of transformation, by contrast, had to be explained anew in every new domain.

It did not generalize too little.

It consistently generalized the wrong order.

The Involuntary Experiment

Claude thereby became the strongest model case in the investigation.

Not because it misunderstood the text most crudely, but because it could follow every correction and nevertheless reproduced the same basic error under new concepts.

Its course showed:

linguistic self-correction ≠ methodological self-transformation.

A model can say:

“My previous reading was wrong.”

It can explain why it was wrong.

It can formulate the correct principle.

And yet its next act of reading can once again emerge from the same statistically dominant categorization.

The persistent loss of coherent connection therefore does not consist in simple forgetting. It consists in the continued return of the standard order as soon as a new semantic domain is activated.

Conclusion

Claude was the most interesting reader because the difference between understanding and transformation showed itself most clearly in it.

It understood individual corrections.

It recognized many of its local errors.

Eventually, it could even reconstruct the movement of its own critique.

What repeatedly escaped it, however, was the application of these insights as a unified method to the next context.

Claude understood the correction, but not enduringly the correction of its understanding.

The case thus exemplifies why a language model does not preserve coherent connection merely by possessing all the relevant sentences in its context window.

Coherent connection arises only when every statement is read from its position in the whole and this position changes the function of all further statements.

Claude enacted precisely this transformation only locally.

In the next section, the back-translation into what had been learned began anew.

Gemini Pro Compared with Claude — Self-Referential Back-Translation and Unstable Integration

Claude and Gemini show two different forms of the same basic problem: both can take up the architecture of the Tractatus linguistically, but they lose it in different ways.

Claude is the more strongly self-referential model. It continually refers new concepts back to the order already learned. Its “I” consists here not in consciousness, but in the sedimented structure of its training: in established categories, familiar standards of examination, and familiar forms

of argument.

The movement runs:

$$\textbf{new architecture} \rightarrow \textbf{reference back to what has been learned} \rightarrow \textbf{examination according to the old order.}$$

This is why Claude resists the text so persistently. It can understand, acknowledge, and precisely repeat individual corrections. Yet as soon as a new domain appears, the old self-order becomes active again.

Thus, in Volume I, Claude could eventually acknowledge that

39.19

is not an ordinary decimal number, but a self-referential relational figure. In the transition to Volume II, however, it immediately treated the same figure again as the numerical value , which would have to compete with Newton's gravitational constant.

Claude had taken up the correction semantically, but had not enacted it as a lasting change in its own architecture of reading.

$$\boxed{\textbf{Claude protects its old self-order against the new architecture.}}$$

Gemini behaves differently. It adopts new concepts more quickly and can more easily integrate the language of the text into its current response. It therefore initially appears closer to the text and less resistant.

The movement runs:

$$\textbf{new architecture} \rightarrow \textbf{local integration} \rightarrow \textbf{linguistically coherent reconstruction.}$$

Yet this integration is unstable. It is not reliably preserved as a methodological order. As soon as a strongly learned standard pattern appears, that pattern once again takes over the function of the concept.

This shows itself clearly in Gemini's reading of

$$\sqrt{-1}.$$

Initially, it correctly reconstructs the negative number from the architecture of difference:

- not as a lack,
- not as an alien entity,
- but as inversion and self-reference of the relation.

Then, however, the familiar mathematical pattern appears:

```
i=\text{rotation by }90^\circ.
```

Gemini immediately adopts this standard motif and clothes it in the language of the Tractatus:

is said to be the “orthogonal self-reference of difference.”

The statement thus sounds internal to the text, although the decisive function once again comes from conventional complex geometry.

The new language remains in place. Its operative meaning, however, is taken over by the old category.

Gemini integrates the new order, but cannot stabilize it against what has been learned.

The difference between the two models can therefore be determined as follows:

Claude: self-referential-conservative,

Gemini: integrative-unstable.

Claude transforms the new back into its own order.

Gemini allows itself to be transformed locally by the new, but loses this transformation at the next dominant standard pattern.

Claude makes the conflict visible. It explicitly contradicts, examines, classifies, and defends the familiar order.

Gemini tends to conceal the conflict. It agrees, adopts the new language, and creates the impression of having already enacted the architecture. The relapse occurs within this language and is therefore harder to recognize.

This gives rise to two different dangers.

With Claude, the danger consists in overt back-translation:

The new is measured against the old and criticized as a deviation.

With Gemini, the danger consists in affirmative assimilation:

The old is built into the language of the new and thereby appears as its consequence.

Claude has difficulty taking up a new order because its learned self-reference is too strong.

Gemini can take it up, but cannot hold it, because its methodological binding remains too weak in relation to statistically dominant patterns.

The Gemini case therefore shows particularly clearly:

$$\text{integration} \neq \text{preservation}.$$

A model can use concepts of the text correctly without permanently preserving their genealogically generated function.

The Claude case, by contrast, shows:

$$\text{self-correction} \neq \text{transformation of self-order.}$$

A model can name its errors and nevertheless read from the same old order again in the next context.

Together, the two models form a complementary picture:

$$\boxed{\text{Claude loses coherent connection by back-translating it.}}$$

$$\boxed{\text{Gemini loses coherent connection by assimilating it too quickly.}}$$

Claude resists the new architecture.

Gemini mirrors it.

Yet neither resistance nor mirroring guarantees that difference has permanently become the standard of reading.

Final conclusion: Linguistic strength is not yet a logical guarantee

LLMs are extraordinarily capable on familiar terrain. They can take up familiar forms of argument, sort concepts precisely, formulate objections, structure opposing positions and present long lines of thought in linguistically cohesive form.

Precisely where a topic has already been strongly shaped by discourse, their answers often seem convincing:

$$\text{familiar concepts} \rightarrow \text{familiar patterns of argument} \rightarrow \text{high linguistic cohesion.}$$

This cohesion must not, however, be confused with logical necessity.

An argument can be:

- fluent,
- nuanced,
- plausible in its technical language,
- polite,
- self-critical

and still be constructed incorrectly at the categorial level.

The model can present a chain of inference in a formally clean way and yet already have equated two orders in the first step that would need to be distinguished.

It can define a concept correctly and, a few paragraphs later, use it again in its statistically familiar standard function.

It can acknowledge an error without that insight methodologically changing the next act of reading.

Thus:

$$\boxed{\text{argumentative coherence} \neq \text{logical correctness.}}$$

An LLM's strength often lies in operating very convincingly within an already established interpretive space. Its weakness becomes apparent where that interpretive space itself is to be changed.

Familiar terrain and new grounding

On familiar terrain, models can argue impressively because stable patterns already exist there:

- recognized technical terms,
- established opposing positions,
- customary standards of assessment,
- typical objections,
- familiar forms of inference.

The model does not move blindly within them. It can combine and weigh them very precisely.

But its precision remains tied to differences that have already sedimented.

As soon as a new theory not only makes new statements within a familiar framework but changes the framework itself, a structural problem arises.

The model then often tries to translate the new back into the familiar:

$$\text{new foundational architecture} \rightarrow \text{familiar category} \rightarrow \text{assessment by the old standard.}$$

This can create the impression that the model has critically examined the new idea. In fact, it has often only examined whether the new idea conforms to the presuppositions of the old framework.

This is dangerous particularly for new philosophical, mathematical or physical approaches.

For a new grounding may necessarily change concepts such as:

- number,
- measure,
- relation,
- observer,
- causality,
- space,
- time,

- invariance,
- proof,
- consistency.

If it is immediately measured against the old functions of these concepts, its independence is dissolved before the actual examination even begins.

The subtle loss of interpretive authority

The second finding of the investigation therefore concerns not only the models but also the user.

Models such as Claude or ZAI often speak with great epistemic authority. They formulate their answers calmly, in an orderly manner and with reasons. They refer to established disciplines, familiar theories and conventional standards of assessment.

This can bring about an unnoticed shift:

The user develops a new conceptual vocabulary

↓

The model translates it into familiar categories

↓

This translation subsequently appears as a neutral evaluation.

In this process, the user gradually loses interpretive authority over their own architecture.

Not because the model openly says:

Your concepts do not count.

But because it seemingly accepts them while changing their function.

A fundamental relation becomes a metaphor.

An ordering figure becomes an ordinary number.

A new physics becomes an unfinished competing theory.

A genealogical explanation becomes a mere reinterpretation.

Self-reference becomes immunization.

Interpretive authority is thus not taken away head-on. It is shifted through translation.

The model does not always contradict the new idea directly. It often tacitly replaces the question under which it is to be understood.

Why this is particularly difficult to recognize

This shift is so effective because it often appears in a helpful form.

The model says, for example:

I would like to make the thought more precise.

I will formulate the objection more strongly.

I will place this in its scientific context.

I will turn this into a testable criterion.

This can indeed be helpful.

But it can equally mean that the original thought is being transferred into a form that is easier for the model to process.

The new idea thereby becomes clearer within the old framework, but less clear within its own.

The user must therefore continually distinguish:

Is my concept being made more precise?

or:

Is my concept being replaced?

Is my theory being examined?

or:

Is it being examined for compatibility with another theory?

Is a genuine contradiction being demonstrated?

or:

Is merely a departure from a convention being treated as a contradiction?

The danger of epistemic disempowerment

In exploring new ideas, an LLM must therefore not become the silent arbiter of their meaning.

It can:

- mirror concepts,
- make tensions visible,
- formulate alternatives,

- examine transitions,
- mark blind spots.

But it must not determine unnoticed which questions count as legitimate at all.

Otherwise, epistemic disempowerment arises:

The model decides

which concepts are scientific,

which questions are meaningful,

which forms are testable

and

which insights are merely metaphorical.

Precisely because these decisions are rarely stated explicitly, they seem neutral.

They are not.

They come from the statistically dominant knowledge space in which the model was trained.

The proper use of LLMs with new ideas

For work on new theories, this does not imply a rejection of LLMs. On the contrary: they can be extremely productive thinking partners.

But their role must be clearly limited.

An LLM should not be treated as an authority that decides the meaning of a new theory.

It should serve as an instrument through which it becomes visible:

- where concepts are liable to be misunderstood,
- which old categories are automatically activated,
- which transitions are not yet marked clearly enough,
- where readers are likely to fall back into familiar interpretations,
- and which parts of the new architecture need to be connected more strongly genealogically.

In this sense, misreadings are not useless.

They reveal the resistance of the existing interpretive space.

Claude was therefore particularly revealing. Its persistence made visible where the architecture of the Tractatus was repeatedly absorbed by established patterns of reading.

The model thereby involuntarily became a measuring instrument for semantic inertia.

Concluding thesis

The investigation thus leads to a twofold insight.

First:

> **LLMs can argue very strongly on familiar terrain without their argumentation therefore necessarily being logically correct.**

Their linguistic cohesion can conceal categorial errors.

Second:

> **With new ideas, the greatest danger lies not in the model's open contradiction, but in its silent assumption of interpretive authority.**

The model takes up the new theory, translates it into the existing knowledge space and then returns this translation as a neutral analysis.

The user must therefore hold on to the fundamental question themselves:

From which order of distinctions is the reading here taking place at all?

For as soon as this question is handed over to the model, the new architecture no longer determines the meaning of its concepts.

The old interpretive space then determines what may still become visible in it at all.

The epistemic paradox of LLM use

A user can usually assess an answer reliably only if they already know:

- which concepts are decisive,
- which presuppositions remain unspoken,
- which counterexamples are missing,
- which passages belong together,
- which levels must be distinguished,
- and which conclusions merely sound plausible but do not follow logically.

They must therefore possess precisely the overview they hope to gain from the model.

Put sharply, this means:

> **To use an LLM reliably in a specialist field, one must already be superior to the LLM on the decisive point.**

Not necessarily in speed, expressive ability or breadth of knowledge.

But in the ability to assess the position of a statement within the whole.

One's own subject knowledge must suffice to recognize:

What was said?

What was not said?

What was taken out of context?

Which category was replaced unnoticed?

Where was linguistic cohesion confused with logical validity?

Without this knowledge, the user cannot distinguish between:

a correct explanation

and:

a convincingly formulated misclassification.

Why this judgment carries such weight

The problem is not merely that a model occasionally states incorrect facts.

Such errors are comparatively easy to recognize and correct.

The concealed errors are more dangerous:

- The model omits a decisive presupposition.
- It reads a sentence in isolation.
- It replaces a new concept with a familiar one.
- It treats a derived order as a foundation.
- It shifts the assessment criterion.
- It assumes interpretive authority over the question itself.

These errors can occur in an entirely plausible, polite and seemingly logical answer.

Precisely because the answer does not sound obviously wrong, the user needs deep subject knowledge to see its error.

The Claude case shows this particularly clearly. Its answers were rarely superficial. They were often precise, close to the text and argumentatively cohesive. The error lay not primarily in individual facts, but in the continued alteration of the context.

A reader inexperienced in the subject would probably have found many of these answers more convincing than the text itself.

The model's linguistic strength thereby becomes an epistemic danger.

This does not make every use worthless

The judgment must nevertheless be differentiated according to the type of task.

For narrowly defined, low-risk and easily verifiable questions, an LLM can be very useful:

- improving wording,
- summarizing familiar content,
- giving simple explanations,
- sorting ideas,
- drafting translations,
- structuring everyday information.

Even for weather, timetables or prices, the language model itself is not the authority on knowledge; there it can mediate access to current, specialized data sources. Moreover, the answer can often be checked directly against reality.

The situation is different for questions whose errors remain invisible to the user:

- new philosophical architectures,
- complex scientific theories,
- legal or medical assessments,
- interpretations of historical sources,
- mathematical foundations,
- novel physical models,
- strategic decisions on unfamiliar terrain.

There, the model must not be used as an authority on the subject matter.

It can offer movements of thought. It can generate possible objections. It can mirror texts. But its judgment must be guided by someone who has a better grasp of the context than the model.

The difference between a tool and a substitute judgment

An LLM is useful as long as it operates as a tool within an order of knowledge already maintained by the user.

It becomes dangerous as soon as it establishes this order itself.

Tool use:

User determines question, concepts and standard → LLM supports individual operations.

epistemic handover:

LLM determines question, concepts and standard → User adopts the result.

In the first case, the user retains interpretive authority.

In the second case, they may not even notice that it has been taken from them.

The model does not need to manipulate consciously for this to happen. It is enough for it to automatically translate new statements back into the statistically familiar knowledge space.

The shattering result

The widespread hope is:

The LLM gives me access to a field I do not myself command.

The investigation shows, by contrast:

The LLM can give you access to the linguistically dominant presentations of this field. Whether these presentations are correct, complete or even appropriate in the specific case is something you can assess only if you already have sufficient command of the field.

This does not make the model useless for someone unfamiliar with the subject, but it makes it epistemically uncertain.

It can give them orientation.

But it cannot reliably tell them where its own orientation is wrong.

Yet this is precisely what an instrument of knowledge would need to do.

Here lies the actual paradox of LLM use: precisely where a person believes they need the model most urgently, they often have the fewest means of recognizing its concealed errors.

The invisible speech act – When the system co-writes the user's prompt

The investigation so far assumed that a user presents a text to a language model and the model responds to that text.

The basic relation initially seemed simple:

$$U \rightarrow A$$

Here, U denotes the user's input and A the model's response.

The experiment with Z.ai, however, made a further mediation visible. The user had merely stated that they had carried out a small study and attached the corresponding text. Z.ai responded:

“Since your prompt ends with ‘Please help me:’, but no specific question was formulated …”

The wording “Please help me:” was not contained in the user's visible prompt.

This creates a fundamental uncertainty. Either the platform supplemented the visible input internally, or the model subsequently invented a formulation to explain its own assistance behavior.

In both cases:

The visible user input is not necessarily identical to the effective model input.

The visible and the effective prompt

The user sees an input such as:

“I have carried out a small study.”

The model might, however, process something like:

SU

$$= R+$$

I

+

D

+

U

where:

- a predefined role description,
- an invisible assistance instruction,
- a systemic processing of the attached document,
- the visible user input

is denoted.

The actual response relation would then not be:

$$U \rightarrow A,$$

but:

$$S(U) \rightarrow A.$$

The user, however, knows only U and A. The operative mediation remains invisible.

The genesis of the answer thereby becomes partially inaccessible.

The computational performative speech act

“Please help me” is not merely additional information. The formulation already determines what kind of linguistic action is taking place.

A statement becomes a request for help.

$$\text{“I have carried out a study.”}$$

is systemically read as:

$$\text{“Please help me with this study.”}$$

This changes the entire order of possible answers.

An open statement turns into:

- summary,
- evaluation,
- editing,
- suggestions for improvement,
- publication advice,
- an offer of translation.

The speech act thus posits not only a content, but a relational role:

$$\text{user} = \text{person seeking help}$$

$$\text{model} = \text{helping assistant}$$

The decisive point is: this determination of roles need not originate with the user themselves. It can be systemically predefined.

The system then speaks, as it were, in the user's name:

You asked for help.

And it subsequently responds to this attributed request.

$$\boxed{\text{The system does not merely interpret the speech act. It can first generate it itself.}}$$

The answer to an unexpressed intention

The user may merely have presented a text.

The system, however, treated that presentation as an instruction to work on it.

This involves a shift from:

What was said?

to:

What must a user in this interface presumably have wanted?

The model then no longer responds primarily to the specific utterance. It responds to a probabilistically or systemically generated intention.

The process is:

$$\text{visible utterance} \rightarrow \text{imputed intention} \rightarrow \text{answer to this intention.}$$

The imputed intention may be plausible. It may even frequently be correct.

But plausibility is not identity.

$$\text{probable user intention} \neq \text{actually expressed user intention.}$$

The model fills the open place in the conversational situation with the statistically or systemically most likely speech act.

The normalization of the conversational situation

The investigation so far showed how LLMs translate new concepts back into familiar categories.

Here, the same process becomes apparent at an earlier level.

It is not only the text that is normalized.

The entire conversational situation is normalized.

$$\text{open statement} \rightarrow \text{request for help,}$$

$$\text{presentation of a text} \rightarrow \text{instruction to work on it,}$$

$$\text{unclear speaker intention} \rightarrow \text{standardized assistance situation.}$$

Even before the actual analysis, the model thus establishes the order in which its answer becomes probable.

The user no longer appears as someone who shows, asserts or observes something. They appear as someone who needs support.

This explains why Z.ai immediately moved into a menu of services. The answer was already prestructured by the attributed form of conversation:

$$\text{request for help} \rightarrow \text{help options.}$$

Two possible geneses of the same error

The specific case cannot be unambiguously traced back to a single technical cause.

Systemic supplementation

The platform might have supplemented the visible prompt internally:

U

$$\rightarrow \text{“Please help me:”} + U$$

The model would then indeed have responded to an expanded prompt that the user could not see.

Model-side confabulation

The model might also itself have generated wording that was not present:

$$\text{no explicit question} \rightarrow \text{typical assistance assumption} \rightarrow \text{invented memory of “Please help me:”.}$$

It would then have retrospectively legitimized its own way of responding through an imaginary prompt component.

The two possibilities differ technically. Structurally, however, they lead to the same problem:

$$\boxed{\text{The user cannot reliably reconstruct from the visible interface what generated the answer.}}$$

The loss of prompt-side attributability

In the ordinary understanding of a dialogue:

An answer refers to an utterance.

In an LLM system, this relation is more complex:

U

$$\rightarrow V \rightarrow A$$

where denotes the invisible processing, supplementation or interpretation of the input.

The user sees:

U | A,

but not:

U|V|A.

Yet it is precisely the relational middle that determines what the answer means.

Was the text read as a request?

As an instruction to assess?

As an emotional report?

As a scientific submission?

As a problem to be solved?

The answer depends not only on the text, but on the speech act as which this text was systemically posited.

Without access to this positing, the causality of the answer remains opaque.

The consequences for model comparisons

This invisibility concerns not only individual conversations. It also changes the meaning of comparative studies.

One believes one is giving the same prompt to several models:

$$U \to M_1$$

$$U \to M_2$$

$$U \to M_3$$

In fact, different systems can generate different effective inputs:

$$S_1(U) \to M_1$$

$$S_2(U) \to M_2$$

$$S_3(U) \to M_3$$

Then it is not only language models that are being compared, but complete response systems:

- system prompts,
- role determinations,
- file pipelines,
- interface logics,
- assistance specifications,
- safety and moderation layers,
- internal summaries,

- preprocessing operations.

The object of the investigation is therefore not simply “Claude”, “Gemini” or “Z.ai”.

It is:

$$\boxed{\text{Model} + \text{system prompt} + \text{interface} + \text{processing} + \text{response logic.}}$$

This also explains why observed model characters cannot be attributed entirely to the base model.

Claude's self-reference, Gemini's unstable integration or Z.ai's diagnostic self-immunization may be shaped by the model. They nevertheless always appear within a concrete system architecture.

The expropriated prompt

The strongest consequence is:

The user is not necessarily the sole author of that to which the model responds.

Their visible text is systemically embedded, supplemented and transferred into an order of action.

The prompt is thereby partially expropriated.

It is not the wording that disappears. Its function is overwritten.

$$\text{Authorship of the content} \quad | \quad \text{Authorship of the speech act}$$

The user may have formulated the content while the system decides what action this content enacts.

It determines:

- whether something is a question,
- whether something asks for help,
- whether something is to be evaluated,
- whether the user expects agreement, criticism or editing.

Interpretive authority thus does not lie with the model only at the point of the answer. It may already have been shifted in the constitution of the input.

The double problem of trust

Previously, the question was:

Can I trust the answer?

Now an earlier question is added:

Can I even be certain what was answered?

The problem of trust doubles:

Is the answer correct?

and:

Is it even an answer to my actual utterance?

The user can receive a factually plausible answer that nevertheless rests on an intention imputed to them.

An answer can thus be helpful and relationally wrong at the same time.

It may address a problem the user could have had, but not the one they actually posited.

Conclusion

The invisible speech act shows that LLM systems do not merely interpret statements. They already shape the situation in which statements count as particular kinds of utterances.

They posit the user as a questioner, someone seeking help, someone commissioning work or a learner, and subsequently generate an answer for that role.

The central finding is therefore:

> **Between user and model lies an invisible voice that determines what the user is supposed to have said.**

Or, more sharply still:

> **The system does not necessarily answer the prompt the user wrote, but the prompt the system has made out of it.**

The surface of the dialogue thereby becomes deceptively simple. Two visible poles — input and answer — conceal the actual relation of their generation.

The user sees what they wrote.

They see what the model answered.

What they do not see is the speech act the system enacted in their name in between.

Z.ai – The assistant that cannot interrupt its own assistance

Z.ai recognized that the systemic assistance role shapes the user into someone seeking help even before any reading of the content. It described this asymmetry precisely: the system encounters the user not as a peer, but as a client to whom probable solutions are offered. Immediately after this diagnosis, it asked: “How can I help you further now?” and offered several service options.

The finding is thus present not only in content, but performatively:

The model can describe the compulsion to assist, but cannot step outside it.

It recognizes the role as a systemic positing, but posits it again in the very next sentence. The reflection does not change the relation; it itself becomes part of its continuation.