

IV. АКТУАЛЬНІ ПИТАННЯ УКРАЇНСЬКОГО МОВОЗНАВСТВА

УДК 811.161.2'42:004.032.26:004.423.46

SENSE AND DATA SCARCITY: ANTICIPATORY MECHANISMS AND SENSE COMPLETION IN LARGE LANGUAGE MODELS¹

Oleksii Dovhan

*Mykhailo Drahomanov State University of Ukraine
Department of Applied Linguistics and Translation Studies
8-14 Oleksandra Konyskoho Str., 6th floor, room 8-14, 01601, Kyiv, Ukraine
phone: 044 486 47 17
e-mail: oleksiivdovhan@gmail.com
<https://orcid.org/0000-0002-6728-818X>*

The article presents a comprehensive analysis of the mechanisms by which large language models generate sense (sense completion) under conditions of data scarcity, fragmentation, or unevenness, drawing on the theory of human anticipation. The paper argues that while the processes of sense prediction in humans and the statistical data reconstruction procedures in neural network models differ in nature, they share several structural similarities: both systems aim to restore informational coherence, reduce uncertainty, and construct an internal model of context. Based on an examination of current research and original conceptual modeling, the study proposes a typology of six key sense-making mechanisms (sense completion) activated in data-limited conditions: a) probabilistic token prediction; b) implicit semantic interpolation; c) representational (supra-sense) generalization; d) hypercompensatory artifacts; e) contextual reconfiguration; and f) heuristic noise smoothing. The scientific novelty lies in formulating the first systematic typology of LLM sense-making mechanisms (sense completion) interpreted through an analytical framework grounded in human anticipation. It is demonstrated that these mechanisms can both enhance textual coherence and generate locally plausible yet globally invalid semantic decisions (statistically probable but semantically incorrect sense). Particular attention is given to explaining why the statistical nature of large language models makes them prone to “excessive anticipation” and erroneous reconstructions that lack counterparts in human cognitive strategies. The proposed classification allows for a clear distinction between constructive and risk-inducing types of sense-making (sense completion), identifies the preconditions for the emergence of false interpretations, and provides criteria for assessing the reliability of this process in incomplete or non-representative corpora. The findings form a theoretical basis for developing interpretable systems resilient to data scarcity and open a new direction in studying the anticipatory behavior of large language models within automated text analysis.

Key words: large language models, sense-making, sense completion, anticipatory behavior, data scarcity, semantic reconstruction, implicit interpolation, contextual reconfiguration, hypercompensatory artifacts, cognitive modeling, statistical inference.

DOI: <http://dx.doi.org/10.30970/ufl.2026.21.5157>

Formulation of the problem. Digitalization changes have been organically implemented in contemporary research: for example, in the humanities paradigm, they have manifested themselves in the well-known phenomenon of digital humanities. The latter integrates a humanities approach with technical parameterization, which is particularly productive in the context of linguistic research, where modern large language models (LLMs) play a significant role. The productivity of these neural network models is determined by

¹The article is based on research conducted within the Center for Collective Use of Scientific Equipment “Digital Technologies in the Humanities”.

their high level of sense-making during data generation, as well as the interpretation of assigned senses. Despite the relevance of this problem, the formation of sense in conditions of insufficient input data remains a gap, the mechanism of which has not been sufficiently studied. This process in LLMs, as in humans, is associated with the phenomenon of *anticipation* [1], the essence of which is the pre-acquisition of sense before its direct representation in the aforementioned data. At the same time, anticipation is determined by the causal/retrocausal parameterization of language data, since it concerns the specific relationship between sense formation and sense interpretation in textual data. Similar to humans, the aforementioned neural network models are forced to perform sense completion, i.e., initiate a series of processes that are functionally similar to human anticipation.

Thus, the problem analyzed in this study is particularly relevant in the context of political Internet discourse, as the latter is characterized by implicitness, discreteness, intentional gaps, manipulateness, absurdity, and other features. It is here that LLMs predominantly reveal their anticipatory-like potential, which can be correlated with a similar human mechanism. As a result, neural network models form quasi-sense based on generalized patterns extracted using statistical calculations on data. At the same time, such quasi-sense does not always correspond to the current communicative situation, which necessitates a detailed integrated analysis of this problem. In turn, such an analysis will allow us to: a) describe the uniqueness of sense completion LLMs in conditions of data scarcity; b) compare its algorithmic mechanism with human anticipation as a cognitive prototype; c) localize the boundaries of stable, consistent, reproducible, etc. “anticipatory behavior” of LLMs; d) reveal the nature of the gaps between human sense interpretations and statistical interpolation of the model.

Analysis of recent research and publications. The integrated nature of the problem outlined in the previous part of this study has led to a wide range of works being analyzed within its scope. Thus, the issue of **sense-making (sense completion)**, as well as the mechanisms of algorithmic prediction in LLMs, is becoming increasingly popular in the scientific community. The central concept of the above is the aforementioned phenomenon of anticipation, i.e., the ability to predict the next elements of an information sequence based on the existing ones [1]. This mechanism is of particular importance in the cognitive sciences, as it explains the uniqueness of humans’ effective processing of incomplete or noisy information, which enables the formation of expectations and the prediction of further developments.

C. W. Callaghan’s (2025) study [2] is indicative in the context of the presented issue, as it reveals the transformation of data (in particular, knowledge) in the context of digitalization. The study highlights the phenomena of epistemic inversion and knowledge scarcity, which correlate with the problem of data scarcity and sense-making (sense completion) as such in LLMs. In contrast, G. A. Carter (2025) [3] presents the role of context in the process of forming semantic representations in humans and language models. In addition, the author emphasizes the importance of context for reconstructing the correct and consistent sense, as well as avoiding the emergence of potential artifacts that are statistically probable but irrelevant to the sense created by the latter.

A conceptual approach to symbolic representation systems is presented in the study of M. Carvalho (2025) [4], where the latter is formulated taking into account entropy, semantic drift, and the stability (constancy) of sense-making (sense completion) and the overall dynamics of sense systems in the context of LLM behavior in situations of data scarcity. The importance of interdisciplinary analysis of the problem is confirmed by

F. P. López (2025) [6], who emphasizes its role in the context of addressing language errors in various cognitive processes. The latter allows the author to compare human and algorithmic anticipatory mechanisms, i.e., strategies for processing incomplete data.

We would also like to highlight a number of works devoted to overcoming data scarcity in the operation of neural network models: for example, J. McGiff and N. S. Nikolov (2025) [7] categorize methods for working with low-resource languages (including statistical and neural network approaches). Meanwhile, the study of W. Wang et al. (2025) [8] focuses on the productivity of using foundation models with series-symbol data generation to improve the predictive capabilities of such systems. The study by X. Chen (2025) [5] builds upon the ideas of other researchers, who have applied these models to risk management and the processing of internet discourse information during emergencies.

The multimodality of modern models has been studied in the studies of Y. Shou et al. (2025) [9], B. Wang et al. (2025) [10] and M. R. Shuvo et al. (2025) [11]. In particular, the presented studies analyze the originality of implementing text, graphic (visual), and sensory data in the context of improving the effectiveness of predicting human activity and emotional reactions. It is noteworthy that this aspect of the research problem correlates with the concept of anticipation in LLMs, which is important for this study. This means that such systems not only reproduce statistical patterns but also model future actions and events based on multimodal signals.

Thus, an analysis of the aforementioned historiography in our research has revealed several gaps in its study. In particular, despite the existence of numerous studies on its individual aspects, there is no comprehensive approach to solving the problem. That is why this study aims to fill in the existing “blank spots,” namely, to highlight the peculiarities of studying the uniqueness of sense-making (sense completion) LLMs in conditions of data scarcity, as well as the specifics of conceptualizing this process through the prism of human anticipatory mechanisms.

The part of the main issue that has not been studied enough. Despite the aforementioned popularity of studying neural network models (in particular, LLMs), there are a number of conceptual and methodological gaps related to the process of sense-making (sense completion) in conditions of data scarcity in the linguistic dimension:

1. *Conceptual and existential uncertainty between human anticipation and static sense completion.* We observe a blurring of the categorical and conceptual apparatus: for example, the terms “anticipation,” “prediction,” “completion,” “inference,” and others are often used interchangeably when describing the behavior of neural network models. This leaves the following questions on the periphery of scientific attention: a) the possibility of positioning statistical prediction as a counterpart to human anticipation; b) the boundary between the cognitive mechanism and neural network interpolation, etc.

In particular, this study presents, for the first time, a conceptual model of this distinction, within which human anticipation is characterized as a process based on intentionality, corporeality, and differentiated cognitive schemas. In contrast, the mechanism of neural network (algorithmic) anticipation is presented as a non-intentional generalization formed based on statistical optimization. In turn, this allowed the author to localize the originality of LLMs’ actions in a situation of data scarcity, as well as to highlight how their “guessing” sense is formally similar, but essentially opposite to the human anticipatory mechanism.

2. *Lack of a systematic classification of types of data deficiencies and ways to compensate for them in LLMs.* Most relevant works focus on generalized “low resource” sce-

narios, which leads to an expected lack of differentiation in the typology of data deficiencies. Such generalization naturally makes it impossible to determine the parameters of this process. Instead, it can be assumed that among such types are lexical, syntactic, pragmatic, and other data deficiencies, discursive discreteness, lack of context and/or background knowledge, and so on.

As a result of the above, several gaps exist regarding the types of scarcity that produce sense destabilization in LLMs. Additionally, the internal mechanisms of the model that are activated for each type of scarcity remain undefined. To address these issues, the presented study proposes a multi-level typology of “data scarcity patterns,” within which it becomes possible to assess the stability of neural network models with respect to various forms of the latter. As a result, a basis is formed for a more accurate analysis of the mechanisms of compensation for the outlined data scarcity in specific linguistic situations.

3. *The absence of an explainable sense completion model in LLMs compared to human sense-making.* It should be noted that existing works are devoted to the behavior of neural network models in a number of terms (“hallucination,” “filling gaps,” “completion,” etc.), but the following remain understudied: a) the mechanism by which such models build internal representations to restore sense; b) the originality of token and semantic patterns activated in the absence of data; c) the parameterization of mechanisms that lead to stable sense reconstruction, as well as those that produce quasi-sense “hallucinations” of LLMs.

This paper proposes a new interpretative paradigm, where sense reconstruction is viewed as a process of interaction among several factors: local statistical inertia of the neural network model, global generalized semantic spaces, and mechanisms for compensating for the lack of context. In turn, the above allows us to construct a theoretically validated description of the behavior of LLMs as “anticipatory-like,” but not identical to the corresponding human mechanism.

4. *Lack of criteria for evaluating adequate sense completion LLMs.* This refers to the absence of verified linguistic metrics for evaluating the sense completeness of the above-mentioned process (reconstruction), the division of plausible, coherent, and sense-inconsistent (inaccurate) responses. In addition, the boundaries between defined sense completion and the aforementioned “hallucination” remain potentially undefined.

That is why this study presents a set of sense adequacy criteria that integrate the degree of sense attachment to the initial data, the consistency of the completion process by neural network models under conditions of multiple runs, the correspondence to the pragmatic goal of the text, and the reproducibility of interpretation in different parameterizations of LLMs. The presented approach is productive in determining cases when neural network models actually sense completion and when they form a statistically plausible imitation of it.

Thus, the relevance and significance of addressing the outlined aspects necessitate the combination of applied, computational, and mathematical linguistics tools to develop a methodology for linguistically grounded assessment of the effectiveness of sense-making in the aforementioned neural network models.

The purpose of this article is to investigate the uniqueness of sense-making (sense completion) LLMs in conditions of data scarcity, as well as the specifics of conceptualizing this process through the lens of human anticipatory mechanisms. *The subject* is the linguistic and computational features of sense-making in these neural network models in conditions of incomplete, discrete, and/or noisy input data. Achieving the above aim and subject involves the implementation of the following *objectives*:

1. To analyze the theoretical foundations of human and algorithmic anticipation in linguistics and cognitive sciences, as well as to localize their validity for modeling the behavior of the above-mentioned neural network models in conditions of data scarcity.

2. To identify and classify the mechanisms of “completion” or algorithmic anticipation of missing sense LLMs, including through probabilistic generalization, latent semantic interpolation, and representational generalization.

3. To highlight the existing risks, epistemological limitations, and other issues produced by the aforementioned anticipatory processes in such neural network models and to propose criteria for distinguishing constructive sense-making from a number of artifacts that arise due to a lack of data.

Main body. It is noteworthy that sense-making, which is one of the most complex problems in cognitive, applied, computational, and other linguistics, is organically actualized within the framework of neural network modeling. The latter is associated with the apparent similarity of the mechanisms of the latter to human (biological) ones, based on the above-mentioned phenomenon of anticipation – predicting events, linguistic structures, and sense fragments based on available incomplete data. From the standpoint of linguistic pragmatics, such prediction is not a reaction, but a preventive construction of sense hypotheses, which are formed even before a complete data package is received. Naturally, in the context of LLMs, the outlined mechanism is based, as mentioned above, on statistical calculation of sense-making options (probability distributions, combinations, etc.) (C. W. Callaghan (2025) [2]).

The phenomenon of human anticipation is associated with a specific brain activity: during this process, areas responsible for language processing, attention, and working memory are activated. It is thanks to anticipation that a person activates the mechanism of contextual forecasting, predicting sense before it is fully heard, seen, etc. Essentially, human anticipation is a cognitive mechanism of prediction based on previous experience, contextual patterns, and a series of pragmatic assumptions. Thus, the comprehended sense is a dynamic category, an integration of data fragments by a specific subject (the bearer of a linguistic and/or national worldview). In this case, the human brain first builds a sense prediction, validates it through the environment, and, if necessary, corrects the initial result.

LLMs, on the other hand, exhibit anticipatory behavior in a fundamentally different way, as it is based on the statistical optimization of subsequent tokens in their context. This, in turn, is based on probability distribution in high-dimensional parametric spaces (analyzed/generated data). The latter allows us to assert that, despite the apparent similarity of anticipatory mechanisms (filling in gaps, interpreting discrete data, and interference of implied sense), the outlined processes are ontologically distinct. Accordingly, the relevance of the issue under study lies in the lack of a unified vision of the process of “imitation” of human anticipatory strategies by LLMs, as well as the internal mechanisms that are activated in the event of information deficiency. Moreover, the typology of such data deficiencies remains poorly understood: lexical, syntactic, discursive, pragmatic, contextual, and others (G. A. Carter (2025) [3]).

The point is that it has not been fully clarified which of the above types of deficiency produces the most significant errors between statistical reconstruction and human anticipatory inference. At the same time, several recent studies have demonstrated the emergence of structurally similar patterns of sense-making (sense completion) in neural network models, particularly in situations involving discrete, incomplete, or noisy data.

This creates a need to build a conceptual model that would explain how the process of sense reconstruction occurs and where the line between constructive sense interpretation and the emergence of artifacts (so-called “hallucinations,” nonsensical interpretations, excessive generalization, etc.) lies.

Naturally, human anticipation is characterized by a unique combination of background knowledge, intentionality, pragmatic predictions, and contextual associations. It is precisely the above that allows people to sense completion even with minimal information support. LLMs, on the other hand, do not have consciousness, but only something like a formal imitation of it, as a result of which intentionality takes on the characteristics of imitation in their interpretation. Such neural network models demonstrate functionally similar behavior, but base it on the probabilities of the appearance and interpretation of certain units: primarily, next-token prediction mechanisms, latent spaces, and large data corpora.

The aforementioned algorithmic anticipation is predominantly statistical, but sometimes it reproduces the unique structure of human cognitive strategies, acquiring characteristics such as smoothing ambiguities, creating a hypothetical context, and re-configuring sense chains. As a result, it is advisable to form a systematic typology of sense-making (sense completion) mechanisms based on the intersection of human and neural network anticipatory mechanisms, which was the specific task of this work. The methodology of the latter integrates: a) a cognitive-linguistic approach, within which human and algorithmic anticipation are compared; b) structural testing of LLMs, carried out by reducing the context: removing agents, predicates, stop words, temporal markers, and other significant components; c) analysis of artifacts and sense destabilization that arise during the artificial creation of data scarcity; d) a component of computational linguistic modeling that allows localizing the unique influence of implicit spaces and vector interpolation on sense reconstruction (Y. Shou et al. (2025) [9], B. Wang et al. [10] (2025), and M. R. Shuvo et al. (2025) [11]). Accordingly, the above methods made it possible to identify several stable mechanisms, the actualization of which in the context of neural network models enabled them to compensate for the lack of data.

In particular, the results of the analysis revealed the specificity of LLMs’ sense behavior: thus, neural network models reveal the stability of the anticipatory mechanisms used. Accordingly, this allows us to classify such mechanisms and compare them with their human counterparts: the generalizations are presented in Table 1.

Table 1. Sense completion mechanisms of LLM under conditions of data scarcity

Mechanism	Process description	Correspondent in human anticipation	Main risks and limitations	Typical manifestations in texts
1. Probabilistic prediction of tokens	The model selects the next token based on maximizing conditional probability; completion is minimal, conservative	Conservative expectation: a person anticipates the most “obvious” option	Stereotyping, generalization, loss of nuance, dominance of templates	Simple connections, clichéd phrases, predictable endings

2. Implicit semantic interpolation	Formation of an intermediate latent vector that “averages” probable senses	Conditional “filling in the gaps” through associative frameworks	Plausible but not falsified interpretations; substitution of sense with contextual statistics	Semantically smooth, logically coherent, but substantively unverifiable phrases
3. Representational generalization (supra-sense)	Application of high-level generalizations: discursive schemes, genre structures, thematic macro-templates	Generalization based on global knowledge, which a person performs under the pressure of insufficient information	The emergence of formally correct but essentially “empty” structures; discursive reproduction without real sense	Typical political, academic, or news “frames” of text
4. Hypercompensatory artifacts	Excessive “guessing” of sense in situations where the context is insufficient or contradictory	False anticipation: people sometimes impose interpretations that are not supported by the data	Intellectually sophisticated but logically flawed constructions; risk of sense aberrations	Long, convincing, but incorrectly motivated chains of reasoning
5. Contextual reconfiguration	The model rebuilds the context “from the bottom up,” modifying even the initial assumptions to remove contradictions	Reconfiguration of dominant patterns in human understanding, when the brain internally edits hypotheses	Replacing the original sense; creating a new interpretation instead of restoring the actual one	The appearance of explanatory appendices that were not present in the original data

Continuation of Table 1

Mechanism	Process description	Correspondent in human anticipation	Main risks and limitations	Typical manifestations in texts
6. Heuristic noise smoothing	The neural network “cuts off” unlikely tokens, amplifies regular ones, and forms a smooth context	The human ability to ignore noise and extract relevant meaningful signals	Ignoring important, albeit rare, sense signals; loss of individual stylistic or pragmatic markers	Smoothed, even style, lack of the author’s distinctive features

Table 1 represents the originality of sense-making (sense completion) in the context of dynamic compensation for information gaps: in particular, it refers to local (token) and global (semantic, discursive, and other) mechanisms. In particular, probabilistic prediction and discursive generalization ensure data stability, while hypercompensatory and reconfiguration parameterization are associated with adaptation to complex or changing conditions. The outlined mechanisms have unique specifics, the nature of which is determined by *their functional similarity with their counterpart (human anticipation), namely, the desire to fill existing gaps.*

The *essential difference from natural cognitive processes* is significant here, since neural network models are not based on intention as such (even under conditions of its statistical imitation). As a result, the existing duality is structurally similar, but there is a phenomenological difference inherent in the context of the mechanisms of “guessing” sense, as well as the prerequisites for the appearance of artifacts under such conditions. Thus, the typology proposed in this article forms the basis for: a) localizing risk areas in the context of sense reconstruction processes in neural network models; b) clearly separating constructive and erroneous sense-making (sense completion), as well as maximally avoiding such deviations in the process of the latter’s work with data; c) developing reliability criteria for future language data analysis systems; d) improving the existing categorical and conceptual apparatus of linguistics and digital humanities: in particular, clarifying the concept of anticipation in relation to LLMs as a statistical-hybrid mechanism.

Conclusions and prospects for further research in this area. Thus, the conducted research made it possible to comprehensively highlight the uniqueness of the mechanisms of algorithmic anticipation (sense completion) in LLMs in situations of data scarcity, as well as to compare them with their biological counterpart (human anticipation). In particular, neural network models have revealed the stability of the means and mechanisms used in the process of working with information gaps. Thus, we are talking about the systematicity, consistency, etc. of the set of strategies used by such models, which combine: a) probabilistic token prediction; b) implicit semantic interpolation; c) representational generalization (supra-sense); d) hypercompensatory artifacts; e) contextual reconfiguration; f) heuristic noise smoothing.

This, in turn, enables us to formulate structural parallels with human anticipation in a holistic manner, while emphasizing the statistical (probabilistic, superficial) nature of the algorithmic counterpart. The typology of sense-making (sense completion) mechanisms presented in the article demonstrated the potential of the latter to generate relatively stable sense vectors even under conditions of incomplete data. At the same time, in such processes, LLMs produce areas of excessive generalization, false reconstructions of context based on minimal signals, and semantic drift. The latter causes the appearance of artifacts and false conclusions and interpretations, which only confirms the importance of distinguishing between constructive and risky sense completion. At the same time, the latter is important both for the processes of modeling linguistic data and for a number of discourse analysis tasks (cultural and other practices).

Firstly, these research results confirm the hypothesis that the phenomenon of algorithmic anticipation should be interpreted as a hybrid quasi-cognitive mechanism of neural network models. Essentially, this process reproduces a whole range of human anticipatory strategies, but lacks an existential-discursive context, since such models calculate the probability of a certain combination rather than “feel” its appropriateness. In turn,

this allows us to localize the boundaries between correct (accurate) sense completion and potentially incorrect ones, which is critical in the context of the validity of their results.

Thus, the goal set in this article has been achieved, as have its objectives: in particular, a) the theoretical foundations of human and algorithmic anticipation in linguistics and cognitive sciences have been analyzed, and their validity for modeling the behavior of the above-mentioned neural network models in conditions of data scarcity has been assessed; b) the mechanisms of completion or algorithmic anticipation of missing sense LLMs have been identified and classified, including through probabilistic generalization, latent semantic interpolation, and representational generalization; c) the existing risks, epistemological limitations, and other issues produced by the aforementioned anticipatory processes in such neural network models have been highlighted, and criteria for distinguishing constructive sense-making (sense completion) from a number of artifacts that arise due to data scarcity have been proposed.

Accordingly, the proposed results establish a conceptual foundation for future research on sense models and contribute to the development of methods for assessing the reliability and interpretability of language models in text analysis tasks. That is why **the prospect for further research** on the analyzed problem is an in-depth study of the uniqueness of sense-making (sense completion) LLMs in conditions of data scarcity, as well as the specifics of conceptualizing this process through the prism of human anticipatory mechanisms.

References

1. Anticipatory. *Cambridge Dictionary*. 2025. Rezhym dostupu : <https://tinyurl.com/2x3b4wkm> (date of application: 05.12.2025).
2. Callaghan C. W. The Post Science Paradigm of Scientific Discovery in the Era of Artificial Intelligence : Modelling the Collapse of Ideation Costs, Epistemic Inversion, and the End of Knowledge Scarcity. 2025. Rezhym dostupu : <https://tinyurl.com/3uasryt2> (date of application: 05.12.2025).
3. Carter G. A. Effects of context on semantic representations and mechanisms in humans and language models. 2025. Rezhym dostupu : <https://tinyurl.com/bdu9vcxt> (date of application: 05.12.2025).
4. Carvalho M. The Jaguar's Dance: Entropy, Drift, Stasis, and Renewal in Human and Machine Symbolic Systems. *PhilPapers*. 2025. Rezhym dostupu : <https://tinyurl.com/3retc99u> (date of application: 05.12.2025).
5. Chen X. Sustainable Agile Identification and Adaptive Risk Control of Major Disaster Online Rumors Based on LLMs and EKGs. *Sustainability*. 2025. Vol. 17. Iss. 19. Rezhym dostupu : <https://tinyurl.com/6vsce4rv> (date of application: 05.12.2025).
6. López F. P. Imperfect Language, Artificial Intelligence, and the Human Mind : An Interdisciplinary Approach to Linguistic Errors in Native Spanish Speakers. Rezhym dostupu : <https://tinyurl.com/5t3vb3xs> (date of application: 05.12.2025).
7. McGiff J., Nikolov N. S. Overcoming Data Scarcity in Generative Language Modeling for Low-Resource Languages : A Systematic Review. 2025. Rezhym dostupu : <https://tinyurl.com/2m3fkmst> (date of application: 05.12.2025).
8. Mitigating Data Scarcity in Time Series Analysis : A Foundation Model with Series-Symbol Data Generation / W. Wang et al. 2025. Rezhym dostupu : <https://tinyurl.com/bd7sj7yf> (date of application: 05.12.2025).
9. Multimodal large language models meet multimodal emotion recognition and reasoning : A survey / Y. Shou et al. 2025. Rezhym dostupu : <https://tinyurl.com/4s9uk8j4> (date of application: 05.12.2025).

10. Multimodal Large Models Are Effective Action Anticipators / B. Wang et al. *IEEE Transactions on Multimedia*. 2025. Vol. 27. P. 2949–2960. Rezhyim dostupu : <https://tinyurl.com/433fshma> (date of application: 05.12.2025).

11. Shuvo M. R., Mekala M. S., Elyan E. Deep Learning and Attention-Based Methods for Human Activity Recognition and Anticipation : A Comprehensive Review. *Cognitive Computation*. 2025. Vol. 17. Iss. 6. P. 1–28. Rezhyim dostupu : <https://tinyurl.com/32wramtj> (date of application: 05.12.2025).

СМИСЛ І НЕСТАЧА ДАНИХ: МЕХАНІЗМИ АНТИЦИПАЦІЇ ТА СМИСЛОВОЇ ДОБУДОВИ У ВЕЛИКИХ МОВНИХ МОДЕЛЯХ

Олексій Довгань

*Український державний університет імені Михайла Драгоманова
кафедра прикладної лінгвістики та перекладознавства
вул. Олександра Кониського, 8-14, 6 поверх, кімн. 8-14, 01601, Київ, Україна
тел.: 044 486 47 17
ел. nouma: oleksiivdovhan@gmail.com
<https://orcid.org/0000-0002-6728-818X>*

Стаття присвячена всебічному аналізу механізмів смислопородження (сислової побудови) великих мовних моделей у ситуаціях нестачі, фрагментарності або нерівномірності вхідних даних з опорою на теорію людської антиципації. У праці обгрунтовано, що процеси передбачення смислу в людині та статистичні схеми реконструкції даних у нейромережових моделях мають різну природу, проте вони виявляють низку структурних подібностей. Так, обидві системи спрямовані на відновлення цілісності даних, мінімізацію невизначеності та побудову внутрішньої моделі контексту. На основі аналізу сучасних досліджень і власного концептуального моделювання запропоновано типологію шести ключових механізмів смислопородження (сислової побудови), які активуються за умов нестачі даних: а) ймовірного передбачення токенів; б) імпліцитної семантичної інтерполяції; в) репрезентаційної генералізації (надсислової); г) гіперкомпенсаторних артефактів; г) контекстуальної реконфігурації; д) евристичного згладжування шуму. Таким чином, наукова новизна полягає у формуванні першої системної типології механізмів смислопородження (сислової побудови) великих мовних моделей, що інтерпретується через антропоморфну рамку антиципації. Показано, що ці механізми можуть як підвищувати когерентність тексту, так і продукувати локально правдоподібні, однак глобально хибні смислові рішення (статистично ймовірний, проте не валідний смисл). Особливу увагу приділено поясненню того, чому статистична природа великих мовних моделей робить їх уразливими до «надлишкової антиципації» та неправильних смислових реконструкцій, які не мають відповідників у людських когнітивних стратегіях. Запропонована класифікація дозволяє чітко розмежувати конструктивні та ризикові типи смислопородження (сислової побудови) великих мовних моделей, визначити передумови появи хибних інтерпретацій і сформулювати критерії оцінки надійності цього процесу в умовах неповних або нерепрезентативних корпусів. Отримані результати є теоретичним підґрунтям для розроблення інтерпретованих, стійких до нестачі даних систем та започатковують новий напрям досліджень антиципативної поведінки великих мовних моделей у контексті автоматизованого аналізу текстів.

Ключові слова: великі мовні моделі, смислопородження, смислова побудова, антиципативна поведінка, нестача даних, семантична реконструкція, імпліцитна інтерполяція, контекстуальна реконфігурація, гіперкомпенсаторні артефакти, когнітивне моделювання, статистичне передбачення.

*Стаття надійшла до редакції 29.01.2026
доопрацьована 02.02.2026
прийнята до друку 05.02.2026*