The Evaluation of GenAI Capabilities to Implement Professional Tasks

Yaroslav Kouzminov
Academic Supervisor, kouzminov@hse.ru

Ekaterina Kruchinskaia
Senior Lecturer, Department of Higher Mathematics, ekruchinskaya@hse.ru

National Research University Higher School of Economics (HSE University), 20, Myasnitkaya str., Moscow 101000, Russian Federation

Abstract

Generative AI (GenAI) or large language models (LLMs) have been running the world since 2022, but despite all the trends surrounding the use of generative models, these cannot yet be used professionally. While they are most valued for 'knowing everything', nonetheless GenAI models cannot explain and prove. In this way we conceptualize the most recent problem of LLMs as the general trend of mistakes even in the core of knowledge and non-causality of mistake via the complexity of question, as the mistake can be named as an accident and be everywhere as the most limitation of professionalism. At their current stage of development, LLMs are not widely used in a professional context, nor have they replaced human workers. They do not event extend workers' professional abilities.. These limitations of GenAI have one general: non-repayment. This article seeks to analyze GenAI's professional viability by examining two models (GigaChatPro, GPT-4) in three fields of knowledge (economics, law, education) based on our unique Bloom's taxonomy benchmark. To prove our assumption concerning the low possibility of its professional usage, we test three hypotheses: 1) the number of parameters of models have low elasticity regarding difficulty and taxonomy with even the right answer; 2) difficulty and taxonomy jointly have no effect on the correctness of an answer, 3) multiple choice is a factor that decreases the number of right answers of a model. We also present the results of GPT-4 and GigaChat MAX on our benchmark. Finally, we suggest what can be done about the limitations of GenAI's architecture to reach at least a quasi-professional use.

Keywords: professionalism; generative artificial intelligence; professional use of language models; graphs of knowledge; orchestration; Bloom's taxonomy

Citation: Kouzminov Y., Kruchinskaia E. (2024) The Evaluation of GenAI Capabilities to Implement Professional Tasks. *Foresight and STI Governance*, 18(4), pp. 67–76. DOI: 10.17323/2500-2597.2024.4.67.76

Introduction

Generative artificial intelligence (GenAI) technologies based on large language models (LLM) have become widespread in the last few years, especially with the emergence of ChatGPT. Such tools are mainly used for the quick retrieval of reference information, writing and translating texts, and creating images and videos. As an assistant, they are in demand in various fields, including education, economics, finance, law, medicine, and pharmaceuticals (Table 1).

The generative AI market is expected to experience a great boost in the years up to 2030. The industry stood at just under $67 billion at the end of 2024, nearly triple the size of 2022. The Statista forecast[1] says it may reach nearly $207 billion (see Figure 1). In Russia, according to the ISSEK sociological survey[2], organizations using AI technologies spend about 15% of total expenditures on digital technologies on them.

According to a report by the consulting company International Data Corporation (IDC) (IDC, 2024), the economic impact of artificial intelligence (AI) technologies around the world will amount to $19.9 trillion by 2030 and will account for 3.5% of global GDP.[3] The IDC study notes that by 2030, every $1 billion of corporate AI investment will generate $4.6 billion for global GDP with direct and indirect impacts.

The overall trend for GDP development is slow. Global GDP growth will slow to 2.7% in 2024 from 2.9% in 2023, according to the Organization for Economic Cooperation and Development (OECD) (OECD, 2024). The indicator's expected value next year will be the lowest since 2020 during the Covid-19 pandemic. The slowdown in growth in the forecast of the organization has been called a direct consequence of the tight measures conducted by the central banks of developed countries and a contraction in business activity, along with generally weakened trade. That is, GDP growth due to AI is a new event, which generally breaks from the existing trend, which is slowing down and focused on stable export-import flows. Is this expectation justified and rational? How we can illustrate those prospects?

Capabilities and Limitations of LLMs

In order to assess expectations surrounding AI's ability to solve problems and stimulate economic growth, it is useful to outline a notional scale of technological potential. For this purpose, let us provide two historical analogies.

The first refers to the alchemical practices of the Middle Ages, associated with attempts to obtain gold from other metals with the help of the Philosopher's Stone. It is noteworthy that in the 20th century it was possible to experimentally convert mercury into gold by means of nuclear reaction, but this method of production turned out to be extremely expensive and has no prospects for payback.[4] In other words, the "philosopher's stone" can refer to technologies characterized by increased labor intensity and financial costs, but are never realized in reality, despite the high expectations associated with them. They become the first reference point on our scale.

The second analogy concerns a development that arose almost by accident, at the intersection of two dominant technologies that have proven useful and cost effective. This refers to the invention of steam locomotives and the formation of railroad infrastructure, processes that resulted from a successful attempt to equip coal transportation vehicles with a steam engine that was originally designed for other purposes (Turnock, 1998). This development had a very simple applied and observable purpose – to accelerate the logistics process with a tangible, measurable result – the speed of doing something. Now railways bring great profits for each country – both direct (for its use and operation and as a huge sector of labor) and indirect – on savings from transaction costs.[5] Technologies with such attributes will be labeled a "steamroller" on our notional scale.

The process of creating the Philosopher's stone also had an applied cost: obtaining gold with minimal costs. However, what does one with this gold if the process of its extraction becomes almost cost-free? No one could give an answer. The other result is obvious – gold would have been devalued. Today, it is seen as an unconditional "currency" - a pledge of stability for the financial marketplace.[6] The Philosopher's stone was never invented – this is due to the unrealistic (and not obvious) way of using the results of this development given that if it were to spread hypothetically, the financial market would lose stability very quickly.

It is difficult to assess which investments were made in both discoveries (or lack of discoveries), and it is not related to the aims of our study. It is, however, worth paying attention to the expectations of returns on these two developments. Where are we and GenAI at this point in our history – closer to a steam locomotive or a philosopher's stone? Do we know which applied tasks, expressed in specific operations and in specific profits,

[1] https://www.statista.com/forecasts/1474143/global-ai-market-size, accessed 27.09.2024.

[2] https://issek.hse.ru/news/981416418.html, accessed 13.11.2024.

[3] Just for comparison, in 2022, the share of agriculture in the global gross domestic product was 4.27%. https://www.statista.com/statistics/256563/share-of-economic-sectors-in-the-global-gross-domestic-product/, accessed 11.10.2024.

[4] https://www.ixbt.com/live/offtopic/pravda-li-chto-mozhno-sdelat-zoloto.html, accessed 17.10.2024 (in Russian).

[5] For example, the latest year (April 2022 to March 2023) saw passengers contribute GBP £8.6 billion of fares income in the UK (ORR, 2023).

[6] Now the price of the December gold contract on the Chicago Mercantile Exchange (CME) increased by 0.77%, rising to $2,750.9 per ounce, according to trading data as of the end of October 2024, the price of the precious metal rose above $ 2,750 per ounce for the first time in history. https://www.barchart.com/futures/quotes/GC*0/profile, accessed 22.11.2024.

Table 1. **Prospective Applications for GenAI**

Field	Assistant functions of GenAI	Literature
Education	Teacher or student assistants which can be used to decrease routine tasks	Al-Zahrani, Alasmari, 2024; Ogunleye et al., 2024; Al-Zahrani et al., 2023; Gill et al., 2023; Chu et al., 2022; Dai, Ke, 2022; Hassan et al., 2022
Economics and finance	Financial consultants	Shapira et al., 2024
Law	Copilot lawyer who performs basic tasks under the very strict supervision via prompts – e.g., summarize the large document	Alimardani, 2024; Lai et al., 2023
Medicine	Helping with diagnoses	Chen, Esmaeilzadeh, 2024
Pharmaceuticals	Construction of new formulas	Choi at al., 2024; Mortlock, Lucas, 2024

Source: compiled by the authors

we can solve with it? We will attempt to answer following the literature review and author's experiment.

Before analyzing the prospects for the professional use of large language models, let us outline what is meant by the term "professionalism". In this case we define professionalism as the ability to select the optimal solutions within the conditions of uncertainty in keeping with knowledge and resisting the most common, secular knowledge if such information contradicted the true empirically stated knowledge. The LLMs, on the contrary, usually operate according to common knowledge (Strachan et al., 2024). LLMs can operate the basic routine operations such as solving basic knowledge operations and other human routines what can help to simplify operations in these fields (Cheung, 2024; Han et al., 2023). However, such functions are routine, whereas professional activity also implies creativity and work with novelty. Since generative models in most cases rely on superficial knowledge and information from low-quality sources, there is an increased risk that they will give an erroneous answer to even elementary questions, which is regarded as an unsatisfactory outcome.

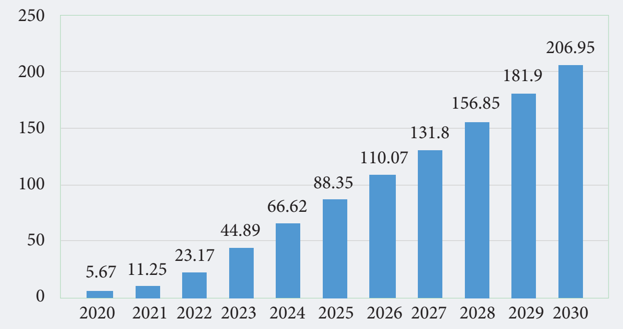
Hence, this is the main limitation for the full-scale delegation of professional tasks to language models. Increasing the number of customization parameters does not eliminate the overall problem, which is further illustrated by statistical hypothesis testing. Thus, when using GenAI for work tasks, it is not yet possible to do without close human supervision. As will be shown empirically, the tested language models GigaChat Pro and GPT-4 make up to 50% errors in the theoretical foundations of law, education, and economics because they lack basic professional knowledge. All known methods of pre-training cannot yet offer an optimal solution. Entrusting professional work to an incompetent "assistant" can be fraught with not only financial but also reputational losses for companies. Therefore, no clear strategy for GenAI productivization has yet been developed.

Another constraint to the expansion of the professional application of LLMs is the lack of empirical analysis of the effects of their implementation in business, education, and other applied fields (McKnight et al., 2024; Sohail et al., 2023). As a consequence, no generative model is currently being used as a professional work product (Noever, Ciolino, 2023). Amidst the expectations associated with their development, there are obvious risks of not satisfying needs, as evidenced by sociological observations. According to Thomson Reuters, almost 60% of lawyers surveyed are not sure that GenAI will have an impact on the value creation of legal services. At least 70% of respondents said they see the greatest risks of using LLMs as being low response accuracy, and 57% ethical impropriety to such tools (Thomson Reuters, 2024). A McKinsey & Company survey shows a similar picture (Figure 2) - organizations are concerned about the low quality and incorrectness of answers generated by AI (56%). Such an indicator obviously does not meet the criteria of professionalism.

We also highlight the lack of empirical studies about LLM implementations to the business, education, or other applied fields and the absence of research describing the practices surrounding LLM use (McKnight et al., 2024; Sohail et al., 2023). As a result, no LLM tool is used completely independently of human oversight in a professional setting (Noever & Ciolino, 2023). There is empirical evidence of the risks of using LLMs. According to a Thomson Reuters survey almost 60% of legal professionals do not believe GenAI will impact the rates they charge clients. More than 50% of respondents stated that LLMs generally have inaccurate responses (70%); poorly comply with laws and regulations (60%); and have a lot of dangerous ethical issues and possibly do not meet criteria for responsible usage (57%) (Thomson Reuters, 2024). Let us also pay attention to the GenAI-related risks that global organizations consider relevant according to a McKinsey & Company poll (see Figure 2) which proves the thesis about the main constraint of LLM usage – the low quality of answers (56%) and inaccuracy, which hinders professionalism.

Thus, the results of AI work require control by means of special procedures - prompting (adjusting queries), fine-tuning of response parameters, and so on. Moreover, LLMs have other limitations (Borji, 2023) because of a number of biases: gender (Borji, 2023),

Figure 1. **Generative Artificial Intelligence (GenAI) Market Size Worldwide from 2020 to 2030 (*billion USD*)**



Source: https://www.statista.com/forecasts/1474143/global-ai-market-size, accessed 18.10.2024.

linguistic (Zhang et al., 2024; Zhang et al., 2021), objectivity (Anthis et al., 2024), and lack of logic (Nguyen et al., 2023; Wan et al., 2024). Thus, generative artificial intelligence models did not bring obvious profit and were not massively productized as professional assistants (Cheung, 2024), nor do they have a level of performance to operate professionally. The final hypothesis we prove in this article via an empirical analysis in the fields of economics, law, and education. This shows that the humans still cannot trust the answers of LLM models professionally, and that we have unjustified expectations from AI.

What have we done to realize the professional usage of LLM and what can be done? Generative AI models have already begun to move toward productization, although they have also done so indirectly. The new "color of the season" is training the model so that it has an increased ability to reason – to use a chain of thoughts (Wei et al., 2022; Liu et al., 2023). In this matter we must declare a limitation of our study – in the moment that we publish this, the new models GPT-4o1 and GigaChat MAX were launched. We have tested these as well, but not in an as statistically rigorous way as we did their predecessors. However, we can note two relevant points based on the newer models. Based on the overall accuracy of the generated responses, it can be noted that GigaChat MAX generally shows a larger gain on our benchmark (Pro vs. MAX +10%) against +5% for GPT-4o1 compared to GPT-4 (see below).

To continue the topic of discussion of which methods are usually used to improve the quality of models, we can name the method of using knowledge graphs (Xu et al., 2024; Luo et al. 2023; Sun et al., 2023) within the framework of RAG (Retrieval Augmented Generation) or KAG (Knowlegdge Augmented Generation), which received a "boom" in the second quarter of 2024 and are quite successful for further training models in the fields of creative thinking (Sanmartin, 2024; Liang et al., 2024), reasoning, and logic (Mirzadeh et al., 2024).

They are even employed to improve the quality of responses to customer queries within individual modules of the model (Xu, 2024). The mixture of Experts (MoE) method is also quite popular, but its optimal use has not yet been identified (Cai et al., 2024; Zhong et al., 2024; Antoniak et al., 2023), although attempts have been underway, strictly speaking, since the time before the advent of transformer models. It should also be noted that there is a great need to increase the level of proficiency in mathematical operations through orchestration techniques (Zhou et al., 2024), for which success has been confirmed (Rasal, 2024).

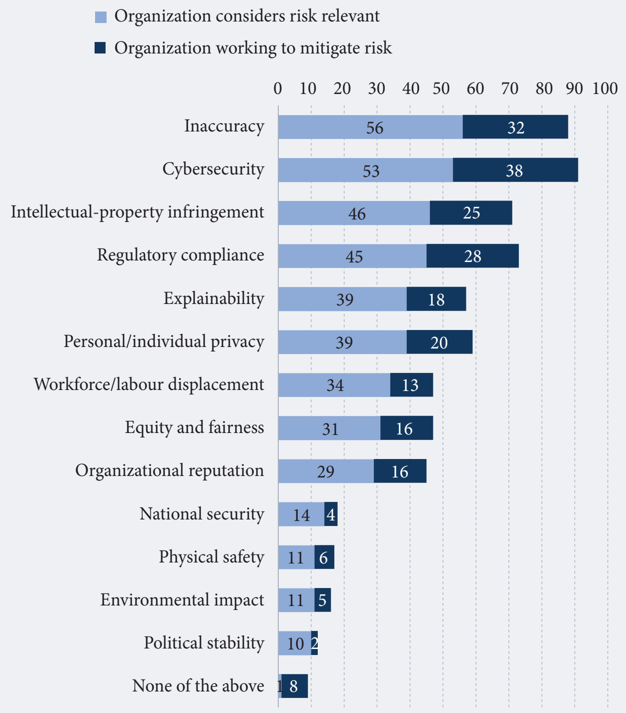
All these strategies can only be called an indirect attempt to achieve the professional use of such models, since these techniques are aimed only at generally improving the ability of the model to respond, which is expressed in the risk of increasing errors and profanity in responses. Such outcomes are unacceptable for a professional, as we have already emphasized earlier. Why is this happening contrary to the expectations of productization? We do not have an established method for "teaching" LLMs, we can only improve upon them.

According to ISSEK (HSE, 2024), the majority (70.1%) of students in AI programs in Russia study within the fields of Engineering, Technology and Technical Sciences; about a quarter (27.2%) are in "Mathematical and Natural Sciences" departments. Still, only 1.5% study as part of "Social Sciences" programs. Why is this dangerous for the productivity of models? Modelers often set themselves the goal of "growing" models, seeking to increase the number of hyperparameters, and then test them on benchmarks that have nothing to do with checking the level of professionalism. Meanwhile, the product departments of companies have clear intentions for the model to perform specific, professionally oriented tasks. According to the results of several empirical experiments, which we will describe in the next section, an increase in the number of hyperparameters of the model does not have a direct impact on improving the abilities of the model.

In this case we can say that the risk of not meeting professional standards cannot be answered without an academic approach to the term 'teaching' (we will conceptualize the academic approach in the methodological section of this paper). We state that professional teaching of LLMs should be based on the concept of the knowledge's core, which was created as one of the key concepts of positivism by Kuhn and Lakatos (Kuhn, 1977; Lakatos, 1963; Lakatos, 1970a; Lakatos, 1970b). The key for teaching LLMs this core tenant is finding a way to restrict the possibility of errors and mistakes in the basic knowledge.

For professional usage we need to align the process of 'boosting' and rethink the core of knowledge, the connection of subject areas with one another to create the clean datasets, the relevant RAG techniques, etc., and these tasks can be done only by professional academics in the relevant fields (recall the example of Arizona State University and OpenAI).

Figure 2. **Risks that Global Organizations Consider Relevant When Using LLMs**



Source: McKinsey & Company. https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-in-2023-generative-ais-breakout-year, accessed 17.10.2024.

Moreover, we need not only 'teach' but also evaluate the results of this 'teaching' in a professional way. Global trends in the development of benchmarks (Wang, 2024) show that professional complication of issues is one of the new turns in the development of benchmarks in general. In this regard, this paper presents a new approach to benchmarking professionalism and testing the actual versions of LLMs on this.

To summarize, in this section of article we have attempted to highlight the problem of not using LLM models professionally, and the lack of opportunities to use them according to some observed facts, both concluded sociologically or via the relevant literature. In the next section, we prove empirically why our corollary about the actual, non-professional level can be stated as real and we further describe our methodology for this experiment.

The Empirical Evaluation of the Degree of Professionalism of LLMs

To test our main hypotheses of the models' inability to act professional, we conducted an experiment using our own benchmark in three fields (domains): economics, law, and pedagogy & education. A detailed presentation of the benchmark methodology was outlined in our earlier study (Kardanova et al., 2024).

To explain it in brief (see Figure 3), the testing process was conducted in accordance with an academic approach, in which the test tasks were prepared based on the principles of:

- scientific criteria – not checking the model's ability to state the facts, but verifying its ability to demonstrate fundamental knowledge and an ability to solve the practice cases needed to have this kind of knowledge.
- taxonomy and difficulty criteria – each question was assigned (by an expert) to the basic Blooms' criteria (Bloom, 1956) of taxonomy: to remember, to understand, and to apply[7]; also, the level of difficulty was assigned to each question: easy, moderate, difficult.[8]
- collectivity (collegiality) – the mixture of independent experts' opinions.
- iterativity – a staged process to check each question for the all the criteria.

Within the framework of these principles, each of the tasks were created by an expert (a professional academic specialist) for each domain and subsequently checked by a psychometrician to prove compliance with the classical measurement theory. Finally, each question was also validated by three independent experts afterwards (see Figure 3).

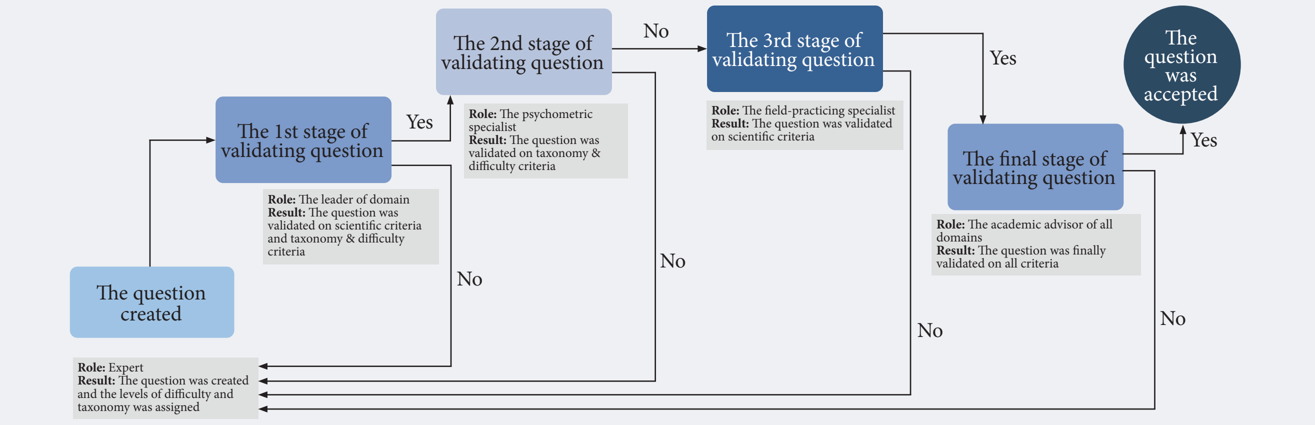
The questions created within the framework of the benchmark are based on the principle that the versatility and responsiveness of knowledge is one of the most important criteria of a professional. The questions test both fundamental knowledge and the possibility of applying this knowledge in various contexts, often non–trivial ones. The tasks were of varying complexities to test the model from different scientific sides. The questions have a certain level of complexity, set by an expert, and can measure the depth of the model's knowledge.

The questions vary in taxonomy levels to test the model's abilities in different contexts. The questions have multiple levels of taxonomy, set by an expert, and can measure the breadth of knowledge of the model and its ability to combine knowledge from different sources in different contextual situations. Taxonomy is the basis for creating any assessment tool for both people (exams) and large language models (benchmarks). However, a separate or unique taxonomy for checking the level of large language models has not yet been developed, which makes the current assessment biased toward psychometric patterns of checking people. Given

[7] Each level was assigned expertly. The level of remembering means that a student can reproduce the main points of a theory. The level of understanding means that the student can relate the theory to a practical task. The level of applying means that the student can apply the relevant theory in the correct way.

[8] Each level was assigned expertly. A task is considered easy if more than 70% of undergraduate students could solve it. The task is considered moderate if between 30% and 70% of undergraduates could solve it. A task is considered difficult if less than 30% of undergraduates could solve it.

Figure 3. **Algorithm of Expert Validation of Test Questions**


The question created
Role: Expert
Result: The question was created and the levels of difficulty and taxonomy was assigned
The 1st stage of validating question
Role: The leader of domain
Result: The question was validated on scientific criteria and taxonomy & difficulty criteria
Yes
No
The 2nd stage of validating question
Role: The psychometric specialist
Result: The question was validated on taxonomy & difficulty criteria
No
No
The 3rd stage of validating question
Role: The field-practicing specialist
Result: The question was validated on scientific criteria
Yes
No
The final stage of validating question
Role: The academic advisor of all domains
Result: The question was finally validated on all criteria
Yes
No
The question was accepted

Source: compiled by the authors

the increase in the number of model parameters, only in 20% or less of the cases for all domains would the questions contain the need to verify knowledge about the fact or theory (an easy task).

In total, more than 13,000 unique MLUU-type original questions were created. Given the very high risk of leakage of a benchmark (and mixing with pre-train[9]), we cannot make public even one of these questions to show an example. We tested models with chain-of-thoughts (CoT) and without them to show the difference in the results if a model "thinks" before answering (in the case of CoT) or not.

All questions were assigned a level of taxonomy and difficulty. Generally, the results of testing GPT-4 and GigaChat Pro (with chain-of-thoughts and without) using these questions shows the impossibility of using these models in a professional setting as they must be evaluated by a person. We state that the impossibility of usage of GenAI in a professional context is highly correlated with fact that the model can make mistakes when answering the simplest questions. To prove this thesis, we provide three hypotheses (all of which were confirmed):

Hypothesis 1: The number of parameters of models have a low elasticity to difficulty and taxonomy even when providing the right answer.

Hypothesis 2: Difficulty and taxonomy jointly have no effect on obtaining a correct answer

Hypothesis 3: Multiple choice is a way to decrease the number of correct answers of a model.

Before statistically proving the hypotheses, we should pay attention to the average level of accuracy (see Table 2). All tests were run with a temperature equaling 1.[10]

Based on results of testing, we can conclude that GPT-4 did not create a serious competition for the GigaChat Pro model. In its current state, neither GigaChat Pro nor GPT-4 can be used for professional purposes without the supervision of a specialist (specific prompting, tuning, etc.). Both models are insufficiently stable and do not show a level above satisfactory - an overall low level is observed (no more than 50% of correct answers). This means that the possession of a theoretical basis (i.e. a broad and valid pre-training base with a stable significance of the subject core) is fundamentally important for further training of both models.

Moreover, as we can highlight the same low correlation of difficulty and taxonomy levels for GigaChat Pro as for the GPT-4. Below we prove that statistically.

Table 3 shows that at all levels of remembering, the best results are observed for all levels of difficulty. This means that for GigaChat Pro, ceteris paribus, it is not difficult to reproduce the theory. However, this does not allow it to "understand" the theory or successfully use it – this can be seen if we turn to the level of understanding and application. Moreover, the model has significant difficulties in understanding which theory to use in a particular case. When the model is tasked with applying the theory – without reflecting on the understanding of the correspondence of theory to practice – it copes generally better, although the gap is no more

[9] Pre-train is a stage of machine learning that consists of forming a knowledge base. The greatest difficulty lies in selecting the most useful information from the "infinite" stream of Internet data, where a paradox arises. On the one hand, if you train the model every time you make a new update to the dataset, this process will be too resource-intensive and slow. On the other hand, if training is carried out after accumulating a "critical array" of updates, the risks of incorrect training and, as a consequence, a drop in the quality of generated answers, increase. Therefore, working with pretrain is a peculiar art of balance. Source: https://habr.com/ru/companies/yandex/articles/759306/, accessed 20.10.2024 (in Russian).

[10] Temperature is a fine-tuned randomness parameter for language model output, measured on a scale from 0 to 1. Lowering its value leads to predictable and "traditional" responses to the user's query. On the contrary, the higher the value of this indicator, the more creativity and variety should be expected in the output. For example, a lower temperature level can be used for factual responses, while an increase in temperature is useful for creative tasks. Source: https://learn.microsoft.com/ru-ru/ai-builder/prompt-modelsettings, accessed 21.10.2024 (in Russian).

Table 2. **Comparative Indicators of the Share of Correct Answers for the Tested LLMs(%)**

Model	CoT	Non-CoT
GigaChat Pro	34	38
GPT-4	45	46

Source: compiled by the authors

Table 3. **The Conjugacy of the Percentage of Correctly Completed Tasks of Different Difficulty Levels and Taxonomy for GigaChat Pro (share of correct answers, %)**

Taxonomy/ Difficulty	Easy	Moderate	Difficult
To remember	44.8	37.1	40.3
To understand	43.3	34.0	31.7
To apply	41.9	33.6	33.7

Source: compiled by the authors

Table 4. **The Conjugacy of the Percentage of Correctly Completed Tasks of Different Difficulty Levels and Taxonomy for GPT-4 (share of correct answers, %)**

Taxonomy/ Difficulty	Easy	Moderate	Difficult
To remember	54.4	46.4	45.8
To understand	52.7	44.5	40.8
To apply	49.2	45.3	44.3

Source: compiled by the authors

than 1.5% on average for the levels of application and understanding (in favor of application).

Thus, two significant conclusions can be drawn. In general, the low level of theory proficiency (less than 50%) does not allow the model to put into practice the knowledge that exists in the pre-training. A significant gap in the levels of "reproduction-application" means that the data available in the pretraining is simply not enough to apply them in professional or academic situations.

If the level of understanding is the most difficult for the model, then the model does not have relevant mechanisms to correlate a theory, but the relevant theory in the pre-train correlates it with the practical situation. This was the case for GigaChat Pro. For GPT-4, all patterns are the same (see Table 4).

As we state in Hypothesis 1, the number of parameters of models have a low elasticity regarding difficulty and taxonomy even when the right answer is obtained. Using the results of regression analysis (OLS model specification) we find that this hypothesis can be confirmed (see Model 1, Table 5). The dependent variable shows the status (right or wrong answer to a question), the covariates show the level of difficulty and taxonomy.

Table 5. **Model Parameters for Testing Hypothesis H1**

Dependent variable (status – 0/1)	GPT-4		GigaChat Pro	
	non CoT	CoT	non CoT	CoT
Intercept	0.52***	0.51***	0.45***	0.4***
Difficulty	-0.04***	-0.05***	-0.05***	-0.04***
Taxonomy	-0.01	-0.01	-0.02**	-0.01*
Observations	13225	13225	13225	13225
AIC	1.665e+04	1.665e+04	1.665e+04	1.665e+04
BIC	1.669e+04	1.669e+04	1.669e+04	1.669e+04

Note: Status: 0 — wrong answer; 1 — right answer. * — p < 0.05, ** — p < 0.01, *** — p < 0.001.

Source: compiled by the authors.

The intercepts, which show the general level of accuracy if the other parameters are equal to 0, for both models (with CoT and without) have no more than 11% of variance. It should be noted that the GPT-4 model has almost 2 trillion parameters while GigaChat Pro has only 40 billion. This highlights a plateau in the development of large language models by simply boosting the number of parameters and the insufficiency of such a technique for the professional improvement of models (the ability to solve professional problems).

For GPT-4, we can observe similar trends (see Table 4), only with the 5%-10% best results. Again, this result can be highlighted as a great prospect for GigaChat Pro, drawing attention to the great gap of hyperparameters numbers. Practically, this means that both models do not have mechanisms for comparing theory and practice yet, which is unacceptable for professional use in general. Moreover, for both models we can see that the results without CoT is better.

We also found that the relationship "difficulty vs the chance of answering correctly", although statistically significant, is insignificant – all other things being equal, an increase in the difficulty level only reduces the chance of answering correctly by 4%-5% for all models. The taxonomy factor is statistically significant only for GigaChat Pro, but the number is negligible (1%).

Test model 1 has the following formula:

$Status_i = \beta_0 + \beta_1 * Difficulty_i + \beta_2 * Taxonomy_i + \epsilon_i$ *(1)*

where: β – constant (the value of the equation provided that all variables are equal to 0); $Difficulty_i$ –difficulty level; $Taxonomy_i$ – taxonomy element; ϵ_i – standard error (deviation of the predicted value from the real value of the variable).

As we stated in the Hypothesis 2, models can have mistakes in as many easy tasks as difficult ones, also the same can be said for different taxonomies. We checked the hypothesis with Model 2 (Table 6).

$Status_i = \beta_0 + \beta_1 * Difficulty_i + \beta_2 * Taxonomy_i + \beta_3 * (Difficulty_i * Taxonomy_i) + \epsilon_i$ (2).

Table 6. **Model Parameters for Testing Hypothesis H2**

Dependent variable (status – 0/1)	GPT-4		GigaChat Pro	
	non CoT	GPT-4 CoT	non CoT	GigaChat Pro CoT
Intercept	0.54***	0.54***	0.47***	0.41***
Difficulty	-0.06***	–0.1***	–0.07***	–0.05***
Taxonomy	–0.03**	–0.03**	–0.03*	–0.02*
Difficulty* Taxonomy	0.02*	0.03**	0.02	0.01
Observations	13225	13225	13225	13225
AIC	1.665e+04	1.665e+04	1.665e+04	1.665e+04
BIC	1.669e+04	1.669e+04	1.669e+04	1.669e+04

Note: Status: 0 — wrong answer; 1 — right answer. * — p < 0.05, ** — p < 0.01, *** — p < 0.001.

Source: compiled by the authors.

Table 7. **Model Parameters for Testing Hypothesis H3**

Dependent variable (status – 0/1)	GPT–4		GigaChat Pro	
	non CoT	CoT	non CoT	CoT
Intercept	0.58***	0.61***	0.58***	0.47***
Difficulty	–0.05***	–0.07***	–0.05***	–0.04***
Taxonomy	–0.03**	–0.03**	–0.03**	–0.02*
Taxonomy* Difficulty	0.01	0.02**	0.01*	0.002
Multiple	–0.23***	–0.41***	–0.24***	–0.35***
Observations	13225	13225	13225	13225
AIC	1.665e+04	1.665e+04	1.665e+04	1.665e+04
BIC	1.669e+04	1.669e+04	1.669e+04	1.669e+04

Note: Status: 0 — wrong answer; 1 — right answer. * — p < 0.05, ** — p < 0.01, *** — p < 0.001.

Source: compiled by the authors.

As we see from Model 2 (the same specification as Model 1, but the interaction variable of taxonomy and difficulty was added), the interaction of taxonomy and difficulty is not significant. This means that for each question, ceteris paribus, there is a joint, but tiny influence of taxonomy and complexity, and it is positive (significant only for GPT-4). This may be a consequence of the fact that the level of understanding for a moderate level of complexity is the most problematic for the GPT-4 model to solve. Generally, the coefficients for GPT-4 being insignificant with regard to the joint effect of taxonomy and difficulty in GigaChat Pro allow us to confirm Hypothesis 2 and conclude that the models nowadays produce errors both in simplest questions where remembering is all that is necessary as well as the difficult prompts of applying or understanding (the process, the theory, etc.).

Table 8. **Results of Preliminary Testing of New Versions of GigaChat and GPT-4**

Language model (testing mode)	Share of right answers	Increase of share in comparison with previous versions
GigaChat MAX (non-CoT)	49%	+10% to GigaChat Pro
GPT-4o1 (non-CoT)	51%	+5% to GPT-4

Source: compiled by the authors

Finally, the Hypothesis 3 states that multiple choice prompts significantly decrease the number of correct answers of a model (Table 7).

$$Status_i = \beta_0 + \beta_1 * Difficulty_i + \beta_2 * Taxonomy_i + \beta_3 * (Difficulty_i * Taxonomy_i) + \beta_4 * Multiple_i + \epsilon_i \ (3),$$

where $Multiple_i$ is the factor of multiple choice.

Model 3 shows a statistically significant and negative relationship between the multiple-choice factor and the likelihood of answering correctly for all models. The coefficient indicates that the previously stated assumption that the model copes worse with multiple choice produces error rates between 23% and 41%, while the GigaChat Pro copes better than the GPT-4 using CoT. Interestingly, including the factor of multiple-choice means that GPT-4 starts performing better with the CoT than without. This can be explained by the fact that GPT-4 performs better with simple operations and CoT helps it. The corollary of this is the fact that the more diverse task is, the worst results we get. On the contrary, a strong professional should and must deal with complex situations.

While this article was being written, new models were released – GPT-4o1 and GigaChat MAX. We did therefore run our test using these versions as well (Table 8).

We can only prove our hypothesis about the potential of the smaller model (in this case GigaChat) to more significantly.

Conclusion

This article illustrates the current quality of GenAI. In fact, such models as GPT-4 and GigaChat can be successfully used for translating, summarizing, content-making for non-professional tasks. However, today's models definitely need a supervisor, and they are not ready to go beyond the co-pilot in use. We see that the models are mistaken in the core of knowledge, imitating the behavior of Ostap Bender[11], they "adapt" to the context, without highlighting the scientific truth. Imagine that you have hired a law assistant who makes unpredictable mistakes in interpreting Constitutional provisions, an economist who cannot calculate the Pareto optimum, or a teacher's assistant who cannot distinguish tutoring from developmental teaching methods. At the same time, you do not know exactly where

[11] Ostap Bender is the hero of I. Ilf and E. Petrov's novel The Twelve Chairs. Ostap Bender is the hero of the novel "The Twelve Chairs" by Ilf and Petrov, characterized by his ability to over-invent, adjust to any interlocutor, and reproduce false facts. In foreign literature, the closest analog of such a character can be considered Baron Munchausen.

your assistant will have fewer mistakes: in light routine tasks or complex applied ones. All you know is that these errors will occur. And these errors will be, strictly speaking, regardless of the level of general "knowledge" of the assistant.

These conclusions were confirmed by three regression models in this paper and found additional validation on the primary results of the new GPT-o1 and GigaChat MAX models. We would especially like to emphasize that increasing the parameters of the model has a negligible effect on the ability of the model to perform professional tasks. We believe that a new word in the development of generative artificial intelligence lies in the orchestration of models (Zhou et al., 2024), in the use of knowledge graphs to increase connectivity of knowledge an imitate the human cognition (Jin et al., 2023; Zhu et al., 2024; Wen et al., 2023; Yang et al., 2023) and in what can be called the development of special glossaries (what brings the terminology of the core of knowledge into a clear form for LLMs), but these are prospects for description in future publications.

We express our sincere gratitude to Elena Kardanova (HSE), Helen Yusupova (HSE), Ksenia Tarasova (HSE) for the well-coordinated work and psychometric expertise that formed the basis of this article. We also thank SberDevices, especially Denis Filippov, Fyodor Minkin, Evgeniy Evdokimov, Sergey Kozlov, Valeria Zanina and Irina Kobzareva for the opportunity to conduct this experiment, as well as appreciate the openness to the academy.

References

Alimardani A. (2024) Generative artificial intelligence vs. law students: An empirical study on criminal law exam performance. *Law, Innovation and Technology*, 2392932, 1–43. https://doi.org/10.1080/17579961.2024.2392932

Al-Zahrani A., Alasmari T. (2024) Exploring the impact of artificial intelligence on higher education: The dynamics of ethical, social, and educational implications. *Humanities and Social Sciences Communications*, 11(1), 912. https://doi.org/10.1057/s41599-024-03432-4

Al-Zahrani A.M. (2024) From Traditionalism to Algorithms: Embracing Artificial Intelligence for Effective University Teaching and Learning. *IgMin Research*, 2(2), 102–112. https://doi.org/10.61927/igmin151

Anthis J., Lum K., Ekstrand M., Feller A., D'Amour A., Tan C. (2024) *The impossibility of fair LLMs* (ArXiv paper 2406.03198). https://doi.org/10.48550/arXiv.2406.03198

Antoniak S., Krutul M., Pióro M., Krajewski J., Ludziejewski J., Ciebiera K., Król K., Odrzygóźdź T., Cygan M., Jaszczur S. (2023) *Mixture of Tokens: Continuous MoE through Cross-Example Aggregation* (ArXiv paper 2310.15961). https://doi.org/10.48550/arXiv.2310.15961

Bloom B.S., Engelhart M.D., Furst E.J., Hill W.H., Krathwohl D.R. (1956) *Taxonomy of Educational Objectives: The Classification of Educational Goals* (Handbook 1: Cognitive Domain), Ann Arbor, MI: Edwards Bros.

Borji A. (2023) *A categorical archive of Chat GPT failures* (ArXiv paper 2302.03494) https://doi.org/10.48550/arXiv.2302.03494

Cai W., Jiang J., Wang F., Tang J., Kim S., Huang J. (2024) *A Survey on Mixture of Experts* (ArXiv paper 2407.06204). https://doi.org/10.48550/arXiv.2407.06204

Chen Y., Esmaeilzadeh P. (2024) Generative AI in Medical Practice: In-Depth Exploration of Privacy and Security Challenges. *Journal of Medical Internet Research*, 26, e53008. https://doi.org/10.2196/53008

Cheung M. (2024) *A Reality check of the benefits of LLM in business* (ArXiv paper 2406.10249). https://doi.org/10.48550/arXiv.2406.10249

Choi J., Palumbo N., Chalasani P., Engelhard M.M., Jha S., Kumar A., Page D. (2024) *MALADE: Orchestration of LLM-powered Agents with Retrieval Augmented Generation for Pharmacovigilance* (ArXiv paper 2408.01869). https://doi.org/10.48550/arXiv.2408.01869

Chu H.C., Hwang G.H., Tu Y.F., Yang K.H. (2022) Roles and research trends of artificial intelligence in higher education: A systematic review of the top 50 most-cited articles. *Australasian Journal of Educational Technology*, 38(3), 22–42.

Dai C-P, Ke F. (2022) Educational applications of artificial intelligence in simulation-based learning: A systematic mapping review. *Computers and Education: Artificial Intelligence*, 3, 100087. https://doi.org/10.1016/j.caeai. 2022.100087

Gill S.S., Xu M., Patros P., Wu H., Kaur R., Kaur K., Fuller S., Singh M., Arora P., Kumar A.P., Stankovski V., Abraham A., Ghosh S.K., Lutfiyya H., Kanhere S.S., Bahsoon R., Rana O., Dustdar S., Sakellariou R., Uhlig S., Buyya R. (2023) Transformative Effects of ChatGPT on Modern Education: Emerging Era of AI Chatbots. *Internet of Things and Cyber-Physical Systems*, 4, 19–23. https://doi.org/10.1016/j.iotcps.2023.06.002

Han S.J., Ransom K.J., Perfors A., Kemp C. (2023) Inductive reasoning in humans and large language models. *Cognitive Systems Research*, 83, 1–28. https://doi.org/10.1016/j.cogsys.2023.101155

Hassan R., Ali A., Howe C.W., Zin A.M. (2022) Constructive alignment by implementing design thinking approach in artificial intelligence course: Learners' experience. *AIP Conference Proceedings*, 2433(1), 0072986. https://doi.org/10.1063/5.0072986

Hendrycks D., Burns C., Basart S., Zou A., Mazeika M., Song D., Steinhardt J. (2020) Measuring Massive Multitask Language Understanding (ArXiv paper 2009.03300). https://doi.org/10.48550/arXiv.2009.03300

HSE (2024) *Training of highly qualified personnel in the field of artificial intelligence* (ed. L.M. Gokhberg), Moscow: HSE University (in Russian).

IDC (2024) *The Global Impact of Artificial Intelligence on the Economy and Jobs*, Needham, MA: IDC Corporate.

Jin B., Liu G., Han C., Jiang M., Ji H., Han J. (2023) *Large Language Models on Graphs: A Comprehensive Survey* (ArXiv paper 2312.02783). https://doi.org/10.48550/arXiv.2312.02783

Kardanova E., Ivanova A., Tarasova K., Pashchenko T., Tikhoniuk A., Yusupova E., Kasprzhak A.G., Kuzminov Y., Kruchinskaia E., Brun I. (2024) *A Novel Psychometrics-Based Approach to Developing Professional Competency Benchmark for Large Language Models* (arXiv paper 2411.00045). https://doi.org/10.48550/arXiv.2411.00045

Kuhn T.S. (1977) *The Essential Tension*, Chicago: University of Chicago Press.

Lai J., Gan W., Wu J., Qi Z., Yu P.S. (2023) *Large Language Models in Law: A Survey* (ArXiv paper 2312.03718). https://doi.org/10.48550/arXiv.2312.03718

Lakatos I. (1963) Proofs and Refutations (I). *British Journal for the Philosophy of Science,* 14(53), 1–25.

Lakatos I. (1970a) Falsification and the Methodology of Scientific Research Programmes. In: *Criticism and the Growth of Knowledge* (eds. I. Lakatos, A. Musgrave), Aberdeen: Cambridge University Press, pp. 91–195.

Lakatos I. (1970b) History of Science and Its Rational Reconstructions. *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, pp. 91–136.

Liang L., Sun M., Gui Z. et al. (2024) KAG: Boosting LLMs in Professional Domains via Knowledge Augmented Generation. ArXiv paper 2409.13731, 1–33. https://doi.org/10.48550/arXiv.2409.13731

Liu N.F., Lin K., Hewitt J., Paranjape A., Bevilacqua M., Petroni F., Liang P. (2023) *Lost in the Middle: How language models use long contexts* (ArXiv paper 2307.03172). https://doi.org/10.48550/arXiv.2307.03172

Luo L., Li Y.F., Haffari G., Pan S. (2023) *Reasoning on Graphs: Faithful and Interpretable Large Language Model Reasoning* (ArXiv paper 2310.01061). https://doi.org/10.48550/arXiv.2310.01061

McKnight M.A., Gilstrap C.M., Gilstrap C.A., Bacic D., Shemroske K., Srivastava S. (2024) Generative Artificial Intelligence in Applied Business Contexts: A systematic review, lexical analysis, and research framework. *Journal of Applied Business and Economics*, 26(2), 7040. https://doi.org/10.33423/jabe.v26i2.7040

Mirzadeh I., Alizadeh K., Shahrokhi H., Tuzel O., Bengio S., Farajtabar M. (2024) *GSM-Symbolic: Understanding the limitations of mathematical reasoning in large language models* (ArXiv paper 2410.05229). https://doi.org/10.48550/arXiv.2410.05229

Mortlock R., Lucas C. (2024) Generative artificial intelligence (Gen-AI) in pharmacy education: Utilization and implications for academic integrity: A scoping review. *Exploratory Research in Clinical and Social Pharmacy,* 15, 100481. https://doi.org/10.1016/j.rcsop.2024.100481

Naveed H., Khan A.U., Qiu S., Saqib M., Anwar S., Usman M., Akhtar N., Barnes N., Mian A. (2023) *A comprehensive overview of large language models* (ArXiv paper 2307.06435). https://doi.org/10.48550/arXiv.2307.06435

Nguyen H., Fungwacharakorn W., Satoh K. (2023) *Enhancing logical reasoning in large language models to facilitate legal applications* (ArXiv paper 2311.13095). https://doi.org/10.48550/arXiv.2311.13095

Noever D., Ciolino M. (2023) *Professional Certification Benchmark Dataset: The first 500 jobs for large language models* (ArXiv 2305.05377). https://doi.org/10.48550/arXiv.2305.05377

OECD (2024) *OECD Economic Outlook* (Interim Report, September 2024), Paris: OECD.

Ogunleye B., Zakariyyah K.I., Ajao O., Olayinka O., Sharma H. (2024) A Systematic Review of Generative AI for Teaching and Learning practice. *Education Sciences*, 14(6), 14060636. https://doi.org/10.3390/educsci14060636

ORR (2023) *Rail industry finance (UK): April 2022 to March 2023,* London: Office of Rail and Road.

Rasal S., Hauer E.J. (2024) Navigating Complexity: Orchestrated Problem Solving with Multi-Agent LLMs (ArXiv paper 2402.16713). https://doi.org/10.48550/arXiv.2402.16713

Sanmartin D. (2024) *KG-RAG: Bridging the gap between knowledge and creativity* (ArXiv paper 2405.12035). https://doi.org/10.48550/arXiv.2405.12035

Shapira E., Madmon O., Reichart R., Tennenholtz M. (2024) *Can LLMs replace economic choice prediction labs? The case of language-based persuasion games* (ArXiv paper 2401.17435). https://doi.org/10.48550/arXiv.2401.17435

Sohail S.S., Faiza Farhat F., Himeur Y., Nadeem M., Madsen D.O., Singh Y., Atalla S., Mansoor W.. (2023) Decoding ChatGPT: A taxonomy of existing research, current challenges, and possible future directions. *Journal of King Saud University – Computer and Information Sciences*, 35(8). https://doi.org/10.1016/j.jksuci.2023.101675

Strachan J., Albergo D., Borghini G., Pansardi O., Scaliti E., Gupta S., Saxena K., Rufo A., Panzeri S., Manzi G., Graziano M.S.A., Becchiol C. (2024) Testing theory of mind in large language models and humans. *Nature Human Behaviour,* 8(7), 1285–1295. https://doi.org/10.1038/s41562-024-01882-z

Sun J., Xu C., Tang L., Wang S., Lin C., Gong Y., Ni L.M., Shum H.Y., Guo J. (2023) *Think-on-Graph: Deep and responsible reasoning of large language model on knowledge graph* (ArXiv paper 2307.07697). https://doi.org/10.48550/arXiv.2307.07697

Thomson Reuters (2024) *2024 Generative AI in Professional Services*, Toronto: Thomson Reuters Institute.

Turnock D. (1998) *An Historical Geography of Railways in Great Britain and Ireland* (1st ed), New York: Routledge.

Wan Y., Wang W., Yang Y., Yuan Y., Huang J., He P., Jiao W., Lyu M.R. (2024) *$A \wedge B \Leftrightarrow B \wedge A$: Triggering logical reasoning failures in large language models* (ArXiv paper 2401.00757). https://doi.org/10.48550/arXiv.2401.00757

Wang Y., Ma X., Zhang G., Ni Y., Chandra A., Guo S., Ren W., Arulraj A., He X., Jiang Z., Li T., Ku M., Wang K., Zhuang A., Fan R., Yue X., Chen W. (2024) *MMLU-Pro: A more robust and challenging Multi-Task Language Understanding benchmark* (ArXiv paper 2406.01574). https://doi.org/10.48550/arXiv.2406.01574

Wei J., Wang X., Schuurmans D., Bosma M., Ichter B., Xia F., Ed H., Quoc C.V., Zhou L.D. (2022) *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models* (ArXiv paper 2201.11903). https://doi.org/10.48550/arXiv.2201.11903

Wen Y., Wang Z., Sun J. (2023) *MindMap: Knowledge Graph prompting sparks graph of thoughts in large language models* (ArXiv paper 2308.09729). https://doi.org/10.48550/arXiv.2308.09729

Xu Z., Cruz M.J., Guevara M., Wang T., Deshpande M., Wang X., Li Z. (2024) *Retrieval-Augmented Generation with Knowledge Graphs for Customer Service Question Answering* (ArXiv paper 2404.17723). https://doi.org/10.48550/arXiv.2404.17723

Yang L., Chen H., Li Z., Ding X., Wu X. (2023) *Give Us the Facts: Enhancing Large Language Models with Knowledge Graphs for Fact-aware Language Modeling* (ArXiv paper 2306.11489). https://doi.org/10.48550/arXiv.2306.11489

Zhang Y., Ding H., Shui Z., Ma Y., Zou J., Deoras A., Wang H. (2021) *Language models as recommender systems: Evaluations and limitations*. Paper presented at the NeurIPS 2021 Workshop on I (Still) Can't Believe It's Not Better.

Zhang Y., Sun R., Chen Y., Pfister T., Zhang R., Arik S.O. (2024) *Chain of Agents: Large language models collaborating on Long-Context Tasks* (ArXiv paper 2406.02818). https://doi.org/10.48550/arXiv.2406.02818

Zhong Z., Xia M., Chen S., Lewis M. (2024) *Lory: Fully Differentiable Mixture-of-Experts for Autoregressive Language Model Pre-training* (ArXiv paper 2405.03133). https://doi.org/10.48550/arXiv.2405.03133

Zhou J.P., Luo K.Z., Gu J., Yuan J., Weinberger K.Q., Sun W. (2024) *Orchestrating LLMs with Different Personalizations* (ArXiv paper 2407.04181). https://doi.org/10.48550/arXiv.2407.04181

Zhu Y., Wang X., Chen J., Qiao S., Ou Y., Yao Y., Deng S., Chen H., Zhang N. (2023) *LLMS for Knowledge Graph Construction and Reasoning: Recent Capabilities and Future Opportunities* (ArXiv paper 2305.13168). https://doi.org/10.48550/arXiv.2305.13168