Training Large Language Models for Autonomous "Out-of-the-Box" Thinking

1. Introduction: The Challenge of Autonomous Creative Reasoning in Large Language Models

Large language models (LLMs) have achieved remarkable progress in recent years, demonstrating exceptional proficiency in various natural language processing and generation tasks [1]. These models have become integral to applications ranging from machine translation and text summarization to content creation and question answering. While their capabilities are impressive, the notion of "out-of-the-box" thinking, characterized by the generation of novel, unexpected, and insightful ideas that transcend conventional human patterns, remains a significant challenge. This form of thinking involves creating genuinely new concepts and establishing connections that might not be immediately obvious. Current LLMs often rely on extensive training data and, in many cases, specific prompts to guide their output towards desired results [4]. Their performance is intrinsically linked to the patterns learned from the vast datasets they are trained on and the instructions they receive to perform particular tasks. This reliance raises a fundamental question: Is it possible to train or fine-tune LLMs to autonomously exhibit genuine "out-of-the-box" thinking and creativity, without the necessity of explicit system prompts to steer their reasoning? This inquiry forms the central focus of this research proposal.

The potential impact of truly creative and autonomous LLMs is substantial across numerous domains. Imagine artificial intelligence capable of independently formulating groundbreaking scientific theories, conceiving entirely new forms of artistic expression, or devising innovative solutions to intricate global issues. This stands in contrast to the present limitations of LLMs, which, despite their ability to mimic creativity, often fall short of generating truly original ideas [4]. Understanding and potentially overcoming these limitations carries profound theoretical and practical implications. From a theoretical standpoint, it could significantly enhance our understanding of intelligence and the very nature of creativity. In practical terms, it could pave the way for transformative applications that were previously considered beyond the reach of artificial intelligence. This research proposal will delve into the complexities of achieving autonomous creative reasoning in LLMs, exploring potential training methodologies, suitable datasets, and novel evaluation metrics.

2. Deconstructing "Out-of-the-Box" Thinking: Emergence, Creativity, and Novelty in AI

The phenomenon of emergent abilities in large language models refers to capabilities that are not evident in smaller-scale models but manifest as the model size and complexity increase [1]. These abilities often appear unexpectedly and were not explicitly programmed into the models. A key characteristic of these emergent abilities is their seemingly abrupt appearance and unpredictable nature, often transitioning from near-random performance to a high level of competence at a certain scale [7]. However, some research suggests that these emergent abilities, particularly those beyond basic linguistic tasks, might be a consequence of in-context learning, where the model leverages examples provided in the prompt to perform tasks it was not explicitly trained for [10]. This perspective challenges the interpretation of these abilities as genuine reasoning or creative thinking that arises intrinsically from the model's architecture and training. Furthermore, there is an ongoing debate about whether these emergent abilities are fundamental properties of scaling AI models or merely artifacts of the metrics used to measure their performance [7]. Certain metrics might create the illusion of a sudden leap in capability, whereas a more continuous measure might reveal a gradual improvement with scale. The very definition of "emergence" in the context of LLMs is subject to interpretation, encompassing properties not explicitly trained for, advantages gained from pre-training, or simply capabilities that appear only in larger models [8]. This ambiguity underscores the need for a clear and precise understanding of what constitutes emergence when discussing the potential for "out-of-the-box" thinking.

Creativity in the realm of artificial intelligence prompts a re-evaluation of traditional definitions, which often center on human attributes like intention, emotion, and originality [12]. Human creativity typically stems from individual experiences, emotions, thoughts, and a unique combination of knowledge and skills, often driven by intentionality and imbued with emotional connection [15]. In contrast, AI creativity, particularly in LLMs, is based on algorithms and the vast datasets they are trained on, learning patterns and structures to generate new content. While AI can produce outputs that might be perceived as creative, its underlying mechanisms differ significantly from human creative processes, often lacking intrinsic intention and emotional depth [15]. Nevertheless, research indicates that AI systems can indeed produce outputs considered creative, such as generating art and literature, and in some instances, their outputs have been judged as comparable to or even surpassing human creations in specific creative tasks [16]. This raises questions about whether creativity should be considered an exclusively human trait. The role of AI in creativity can also be viewed as a tool to augment human capabilities, providing new perspectives and automating tasks, rather than achieving independent creative capacity [15]. The definition of creativity in the context of AI often revolves around the

capability of producing innovative, novel, and effective ideas [14].

Novelty, in the context of this research, needs to be defined beyond mere statistical unexpectedness. While a language model might generate statistically improbable sequences of words, these might not necessarily constitute insightful or impactful new ideas. "Out-of-the-box" thinking, therefore, is more closely aligned with the generation of high-level novelty that transcends simple recombination of existing information. It involves the formation of connections and the conceptualization of ideas that are genuinely new and potentially significant, not just statistically rare. For LLMs to achieve this, they likely need to exhibit a form of internal reasoning that enables them to generate concepts that are not merely derived from patterns present in their training data but represent true conceptual innovation.

3. The Limitations of Current LLMs: Barriers to True Novelty

A fundamental limitation of current large language models in achieving true novelty lies in their reliance on training data and pattern matching [2]. These models learn by identifying statistical regularities and patterns within the massive datasets they are trained on. Their primary function is to predict the next word in a sequence based on the preceding context, effectively making them sophisticated pattern recognition systems. This inherent mechanism constrains their ability to generate genuinely novel ideas that are not, in some form, present within their training data [4]. The "creativity" exhibited by LLMs is often a reflection of their ability to mimic patterns and styles observed in their training corpus, rather than a manifestation of true understanding or the capacity to create entirely new concepts from first principles [4]. Consequently, there is a significant potential for LLMs to regurgitate or closely mimic segments of their training data, which directly hinders their capacity for originality [8]. This tendency to reproduce learned patterns, especially when prompted on topics with limited coverage in their training data, can even lead to copyright violations [24].

Another significant barrier to true novelty in LLMs is their lack of real-world understanding and embodiment [15]. These models operate solely on textual data and lack the grounding in physical experiences, emotions, and real-world interactions that characterize human cognition. Human creativity is often deeply inspired by personal experiences, emotional responses, and the complex interplay between the physical world and our internal states. The absence of this embodied understanding in LLMs might fundamentally limit their ability to generate truly insightful and relevant novel ideas that resonate with the complexities of the human experience [15]. The distinction between human creativity, rooted in personal experiences and emotions, and AI

creativity, based on algorithms and data, underscores this limitation.

Current LLMs also face challenges in handling novel or rare situations [24]. Their training on vast amounts of existing text means they may struggle to reason effectively about or generate coherent content related to events or concepts that are infrequent or entirely absent from their training distribution. In such scenarios, the probabilistic nature of these models can lead to the generation of outputs that, while seemingly plausible, might be factually incorrect or even nonsensical [4]. For instance, they might struggle with rare events, high recall setups, or high coverage scenarios [25]. The inability to distinguish between reality and fiction when confronted with novel information further highlights this limitation [24].

Finally, the "black box" nature and lack of explainability in the internal reasoning processes of LLMs present a considerable obstacle to intentionally fostering and evaluating "out-of-the-box" thinking [3]. Understanding why an LLM generates a particular output remains a significant challenge. This lack of transparency makes it difficult to discern the mechanisms behind any observed instances of novelty and complicates efforts to guide the models towards more creative and original outputs. The inability to fully comprehend the decision-making processes within LLMs hinders our ability to strategically intervene and enhance their capacity for autonomous creative reasoning.

4. Exploring Potential Training Methodologies for Autonomous Novelty

One promising avenue for training large language models to exhibit autonomous novelty lies in harnessing the concept of intrinsic motivation [26]. Inspired by the human drive for curiosity and exploration, intrinsic motivation in AI refers to mechanisms that encourage an AI system to learn and explore its environment autonomously, without relying solely on external rewards. Research could focus on designing intrinsic reward functions that incentivize LLMs to seek out novel states, concepts, or relationships within their data space [28]. These reward functions could be based on measures of unexpectedness, surprise, or the potential for generating new information. Various approaches to implementing intrinsic motivation could be explored, including rewarding the model for encountering novel data patterns, maximizing surprise in its predictions, or minimizing prediction errors, all of which could drive the model towards less familiar and potentially more innovative areas of its learned representation [28]. Furthermore, there is potential in utilizing LLMs themselves to provide feedback on the novelty of generated content, serving as an intrinsic reward signal within a reinforcement learning framework [32]. For example, one LLM could be trained to generate content, while another could be trained to evaluate its novelty relative to a

corpus of existing knowledge. Methods like ELLM (Exploring with LLMs) demonstrate the use of pre-trained language models to suggest useful goals for exploration and reward agents for achieving them [33]. Similarly, Motif leverages LLM preferences to construct intrinsic rewards for training agents, showing promising results in challenging environments [34]. LanGoal combines LLM guidance with intrinsic exploration rewards to enable agents to propose meaningful goals, further highlighting the potential of this direction [36].

Beyond standard supervised learning, advanced reinforcement learning (RL) techniques offer another potential pathway to train LLMs for novelty. RL allows models to learn through interaction with an environment, receiving rewards for desired outcomes. In this context, reward functions could be designed to explicitly incentivize the generation of unexpected or insightful outputs. This could involve human evaluators providing feedback on the novelty of the model's responses, or the development of automated metrics that correlate with novelty and insight. Multi-agent reinforcement learning, where multiple LLM agents interact and potentially compete, could also be explored as a means to foster emergent creative behaviors. The dynamic interaction between agents could lead to novel strategies and solutions that might not arise from training a single agent. While it might seem counter to the "without system prompts" requirement, reinforcement learning from human feedback (RLHF) could be employed during the training phase to guide the model towards generating novel and insightful responses, even if the final inference stage is prompt-free. Research into using RL to train reasoning models to dynamically allocate computational resources based on the complexity of a task could also contribute to more efficient and potentially more insightful reasoning processes [37]. For instance, RL has been used to optimize the novelty of recommendations [32] and to improve reasoning capabilities based on the correctness of final answers [38].

Adversarial training, a technique commonly used in generative adversarial networks (GANs), could also be employed to encourage creative generation in LLMs. In this approach, one LLM acts as a generator, attempting to produce novel and creative content, while another LLM acts as a discriminator, trying to distinguish this content from outputs generated through standard methods or present in the training data. This adversarial process could push the generator to explore less conventional and more original outputs in order to "fool" the discriminator. A key challenge in this approach lies in defining the criteria for the discriminator to effectively identify non-novel content. The discriminator would need to have a robust understanding of what constitutes originality in the given domain.

Finally, training LLMs to perform reasoning and generate novel ideas within a

continuous latent space, rather than being constrained by the discrete nature of language tokens, presents an intriguing possibility [39]. This approach could allow for more flexible and abstract forms of reasoning, potentially leading to more "out-of-the-box" thinking. By operating in a continuous space, the model might be able to explore conceptual relationships and generate ideas that are not directly tied to specific words or phrases in its vocabulary. The Coconut paradigm, which explores training LLMs to reason in a continuous latent space, demonstrates the potential for this approach to yield novel reasoning patterns and outperform traditional chain-of-thought reasoning in certain logical tasks [39].

5. Datasets and Benchmarks for Training and Evaluating Novelty

Training large language models for autonomous "out-of-the-box" thinking necessitates the use of appropriate datasets and benchmarks. Existing datasets used for training and evaluating creativity in AI, such as those for creative writing, art generation, or problem-solving, could provide a starting point. However, their suitability for fostering autonomous novelty without specific prompts needs careful assessment. Many current datasets are designed for specific tasks or stylistic imitation. Datasets containing examples of human creative breakthroughs or solutions to challenging problems from scientific literature, patent databases, or historical records of innovation might be more relevant.

The creation of novel datasets specifically designed to encourage and evaluate autonomous novelty in LLMs might be essential. Such datasets could include open-ended creative prompts with no single correct answer, examples of highly novel and impactful ideas across various domains, abstract concepts and relationships that require non-conventional thinking, and problems that demand creative solutions by integrating disparate pieces of information. Cross-disciplinary datasets could be particularly valuable in encouraging the model to make novel connections between different fields of knowledge.

Scholarly publications and their citation networks offer a unique resource for creating benchmarks for novelty, as explored in the SchNovel benchmark [14]. In academic research, novelty is a key criterion for publication. The relative time of publication and the patterns of citation can serve as proxies for novelty, with papers published later and those that are highly cited often considered more novel and impactful within their respective fields. The SchNovel benchmark, for instance, uses pairs of scholarly papers with different publication dates to evaluate an LLM's ability to assess novelty [14]. This approach provides a structured way to evaluate a model's understanding of

novelty in a specific domain.

6. Novel Evaluation Metrics for Assessing Autonomous Creativity

Standard NLP evaluation metrics like BLEU, ROUGE, and perplexity are generally insufficient for assessing the creativity and novelty of LLM outputs [42]. These metrics primarily focus on the similarity of generated text to reference texts or the statistical likelihood of the generated output, rather than its originality or insightfulness. A truly creative and novel output might deviate significantly from existing references.

Qualitative evaluation methods, such as the Consensual Assessment Technique (CAT), which relies on expert human judgments to assess various dimensions of creativity including fluency, coherence, and originality, are crucial for evaluating subjective qualities like creativity [42]. While human evaluation can be subjective and time-consuming, it remains a vital component in assessing the nuanced aspects of creative work [43].

Research into novel quantitative metrics is also necessary. This could involve developing metrics that measure the semantic distance of generated content from existing data in a learned embedding space, quantifying the unexpectedness or surprisal of the output based on the model's training data, or attempting to assess the potential impact or usefulness of a novel idea [42]. A framework for assessing scientific creativity based on dimensions like originality, feasibility, fluency, and flexibility, as proposed in some studies [45], could be adapted for evaluating more general forms of "out-of-the-box" thinking. For example, originality could be assessed by judge LLMs, while feasibility and fluency could be evaluated based on predefined criteria. The Figural Interpretation Quest (FIQ) has also been used to compare AI and human creativity, focusing on flexibility and subjective creativity [18].

A comprehensive evaluation strategy will likely require a combination of qualitative human expert judgment and novel quantitative metrics to provide a balanced assessment of autonomous creativity in LLMs [42].

Metric Category	Specific Metric Examples	Advantages	Disadvantages	Suitability for Autonomous Novelty

Standard NLP	BLEU, ROUGE, Perplexity	Easy to compute, widely used for text generation	Focus on similarity/likelihood, not novelty or insight	Low
Qualitative	Consensual Assessment Technique (CAT)	Direct assessment of creativity by experts	Subjective, time-consuming, requires expert panels	High (for subjective aspects of originality and insightfulness)
Quantitative Novelty	Semantic distance from training data, Surprisal metrics	Potentially captures statistical unexpectedness	May not correlate with meaningful or impactful novelty	Medium (needs refinement to focus on semantic and conceptual novelty)
Impact/Usefulness	(To be developed) - e.g., proxy measures like user engagement or expert ratings of potential impact	Measures the potential real-world value of the novel idea	Difficult to define and quantify, often requires time to assess	High (if a robust and timely metric can be developed)
Creativity Dimensions	Originality (e.g., assessed by judge LLMs), Feasibility, Fluency, Flexibility	Provides a structured assessment of different aspects of creativity	May be domain-specific and require adaptation for general "out-of-the-box" thinking	Medium to High (can capture different facets of novelty and creativity)

7. Feasibility Analysis and Potential Challenges

Training large language models to achieve autonomous "out-of-the-box" thinking presents several significant challenges. The computational resources required for training and fine-tuning such large models are substantial [3]. Employing advanced reinforcement learning or adversarial training methodologies could further amplify these computational demands, requiring significant infrastructure and time.

The availability of appropriate data is another critical factor. While LLMs are trained on

massive datasets, curating datasets specifically designed to foster and evaluate autonomous novelty poses a unique challenge [22]. Such datasets might need to include examples of human creativity, abstract concepts, and problems demanding innovative solutions, which can be difficult to collect and structure effectively.

Defining and objectively measuring creativity remains an inherently complex task, even in humans [12]. Developing robust evaluation metrics for autonomous AI creativity that go beyond standard NLP measures and capture the essence of "out-of-the-box" thinking is a major hurdle.

Ensuring the safety and ethical behavior of highly autonomous and creative AI systems is paramount. The potential for generating harmful or misleading content necessitates the incorporation of safety mechanisms and ethical guidelines throughout the research and development process [2].

Finally, there is a concern that training AI models for creativity might inadvertently lead to a homogenization of ideas if the models primarily rely on patterns learned from existing data [14]. Addressing this potential for a lack of true diversity in AI-generated creative content will be an important consideration.

8. Conclusion: Towards Truly Creative and Autonomous Language Models

This research proposal outlines a multifaceted approach to investigate the feasibility of training large language models for autonomous "out-of-the-box" thinking. It acknowledges the remarkable progress of LLMs while highlighting their current limitations in generating truly novel ideas without explicit prompting. The proposal delves into the concepts of emergent abilities, creativity in AI, and the nuances of novelty, emphasizing the need for clear definitions and appropriate evaluation metrics. Several potential training methodologies are explored, including harnessing intrinsic motivation, leveraging advanced reinforcement learning techniques, employing adversarial training, and investigating reasoning in latent spaces. The importance of developing specialized datasets and novel evaluation metrics that go beyond standard NLP measures is also underscored.

Future research directions could explore the impact of different model architectures on creative potential, investigate the benefits of incorporating multimodal information into the training process, and develop more sophisticated intrinsic motivation mechanisms that better mirror human-like curiosity and the drive for discovery. Ultimately, achieving truly creative and autonomous language models remains a long-term endeavor with the potential to revolutionize various fields. This research

aims to contribute to this exciting frontier by exploring novel training paradigms and evaluation strategies that could pave the way for AI systems capable of genuine "out-of-the-box" thinking.

Πηγές αναφοράς

1. Emergent Abilities of Large Language Models - OpenReview, πρόσβαση Μαρτίου 23, 2025, https://openreview.net/pdf?id=yzkSU5zdwD
2. Exploring the Emergent Abilities of Large Language Models - Deepchecks, πρόσβαση Μαρτίου 23, 2025, https://www.deepchecks.com/exploring-the-emergent-abilities-of-large-language-models/
3. Unveiling the power and limitations of large language models - 6Clicks, πρόσβαση Μαρτίου 23, 2025, https://www.6clicks.com/resources/blog/unveiling-the-power-of-large-language-models
4. How to Overcome the Limitations of Large Language Models - Deepchecks, πρόσβαση Μαρτίου 23, 2025, https://www.deepchecks.com/how-to-overcome-the-limitations-of-large-language-models/
5. Improve the Diversity and Novelty of Contents Generated by Large Language Models via inference-time Multi-Views Brainstorming - arXiv, πρόσβαση Μαρτίου 23, 2025, https://arxiv.org/abs/2502.12700
6. Emergent abilities of large language models - Google Research, πρόσβαση Μαρτίου 23, 2025, https://research.google/pubs/emergent-abilities-of-large-language-models/
7. Are Emergent Abilities of Large Language Models a Mirage Or Not? | by Novita AI - Medium, πρόσβαση Μαρτίου 23, 2025, https://medium.com/@marketing_novita.ai/are-emergent-abilities-of-large-language-models-a-mirage-or-not-c53cd56d8686
8. A Sanity Check on 'Emergent Properties' in Large Language Models - Hacking semantics, πρόσβαση Μαρτίου 23, 2025, https://hackingsemantics.xyz/2024/emergence/
9. Emergent Abilities in Large Language Models: A Survey - arXiv, πρόσβαση Μαρτίου 23, 2025, https://arxiv.org/html/2503.05788v2
10. [R] Are Emergent Abilities in Large Language Models just In-Context Learning? - Reddit, πρόσβαση Μαρτίου 23, 2025, https://www.reddit.com/r/MachineLearning/comments/19bkcqz/r_are_emergent_abilities_in_large_language_models/
11. How Quickly Do Large Language Models Learn Unexpected Skills? - Quanta Magazine, πρόσβαση Μαρτίου 23, 2025, https://www.quantamagazine.org/how-quickly-do-large-language-models-learn-unexpected-skills-20240213/
12. The Replacement of What? Artificial Intelligence, Creativity and (More-than-)Humanness - Adalberto Fernandes, 2025 - Sage Journals,

πρόσβαση Μαρτίου 23, 2025, https://journals.sagepub.com/doi/10.1177/09732586241275955?icid=int.sj-full-text.citing-articles.3
13. Redefining Creativity in the Era of AI? Perspectives of Computer Scientists and New Media Artists - Taylor and Francis, πρόσβαση Μαρτίου 23, 2025, https://www.tandfonline.com/doi/full/10.1080/10400419.2022.2107850
14. Evaluating and Enhancing Large Language Models for Novelty Assessment in Scholarly Publications - arXiv, πρόσβαση Μαρτίου 23, 2025, https://arxiv.org/html/2409.16605v1
15. The Difference Between Human Creativity and Generative AI Creativity, πρόσβαση Μαρτίου 23, 2025, https://futuristspeaker.com/artificial-intelligence/the-difference-between-human-creativity-and-generative-ai-creativity/
16. Generative artificial intelligence, human creativity, and art | PNAS Nexus | Oxford Academic, πρόσβαση Μαρτίου 23, 2025, https://academic.oup.com/pnasnexus/article/3/3/pgae052/7618478
17. Artificial Intelligence Capability and Organizational Creativity: The Role of Knowledge Sharing and Organizational Cohesion - Frontiers, πρόσβαση Μαρτίου 23, 2025, https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2022.845277/full
18. Full article: Artificial Creativity? Evaluating AI Against Human Performance in Creative Interpretation of Visual Stimuli - Taylor & Francis Online, πρόσβαση Μαρτίου 23, 2025, https://www.tandfonline.com/doi/full/10.1080/10447318.2024.2345430
19. Artificial creativity - Akademische Gesellschaft, πρόσβαση Μαρτίου 23, 2025, https://www.akademische-gesellschaft.com/comminsights/human-versus-artifical-creativity/
20. Generative AI — Never Truly Creative? | by Axel Schwanke | Medium, πρόσβαση Μαρτίου 23, 2025, https://medium.com/@axel.schwanke/generative-ai-never-truly-creative-68a0189d98e8
21. AI vs. Human Creativity: Can Machines Truly Be Original? - Connective Web Design, πρόσβαση Μαρτίου 23, 2025, https://connectivewebdesign.com/blog/ai-vs-human-creativity
22. Novelty of LLMs | Theory - DataCamp, πρόσβαση Μαρτίου 23, 2025, https://campus.datacamp.com/courses/large-language-models-llms-concepts/building-blocks-of-llms?ex=1
23. From Large Language Models to Reasoning Language Models - Three Eras in The Age of Computation. - YouTube, πρόσβαση Μαρτίου 23, 2025, https://www.youtube.com/watch?v=NFwZi94S8qc
24. LLMs: The Dark Side of Large Language Models Part 2 - HiddenLayer, πρόσβαση Μαρτίου 23, 2025, https://hiddenlayer.com/innovation-hub/the-dark-side-of-large-language-models-part-2/

25. Concerns and limitations of Large Language models : r/datascience - Reddit, πρόσβαση Μαρτίου 23, 2025, https://www.reddit.com/r/datascience/comments/102z0kl/concerns_and_limitations_of_large_language_models/
26. saturncloud.io, πρόσβαση Μαρτίου 23, 2025, https://saturncloud.io/glossary/intrinsic-motivation-in-ai/#:~:text=Definition,perceived%20value%20of%20its%20actions.
27. Intrinsic Motivation in AI - Saturn Cloud, πρόσβαση Μαρτίου 23, 2025, https://saturncloud.io/glossary/intrinsic-motivation-in-ai/
28. Intrinsic motivation (artificial intelligence) - Wikipedia, πρόσβαση Μαρτίου 23, 2025, https://en.wikipedia.org/wiki/Intrinsic_motivation_(artificial_intelligence)
29. Intrinsic Motivation - Lark, πρόσβαση Μαρτίου 23, 2025, https://www.larksuite.com/en_us/topics/ai-glossary/intrinsic-motivation
30. intrinsic motivation (artificial intelligence) - Autoblocks AI, πρόσβαση Μαρτίου 23, 2025, https://www.autoblocks.ai/glossary/intrinsic-motivation
31. The Emerging Neuroscience of Intrinsic Motivation: A New Frontier in Self-Determination Research - PMC - PubMed Central, πρόσβαση Μαρτίου 23, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC5364176/
32. Optimizing Novelty of Top-k Recommendations using Large Language Models and Reinforcement Learning | OpenReview, πρόσβαση Μαρτίου 23, 2025, https://openreview.net/forum?id=D1DE6b0Csh&referrer=%5Bthe%20profile%20of%20Amit%20Sharma%5D(%2Fprofile%3Fid%3D~Amit_Sharma3)
33. Guiding Pretraining in Reinforcement Learning with Large Language Models, πρόσβαση Μαρτίου 23, 2025, https://proceedings.mlr.press/v202/du23f/du23f.pdf
34. Motif: Intrinsic Motivation from Artificial Intelligence Feedback - OpenReview, πρόσβαση Μαρτίου 23, 2025, https://openreview.net/forum?id=tmBKIecDE9
35. [2310.00166] Motif: Intrinsic Motivation from Artificial Intelligence Feedback - arXiv, πρόσβαση Μαρτίου 23, 2025, https://arxiv.org/abs/2310.00166
36. Generate explorative goals with large language model guidance - OpenReview, πρόσβαση Μαρτίου 23, 2025, https://openreview.net/forum?id=hCfhfwSfCg
37. Training Language Models to Reason Efficiently - arXiv, πρόσβαση Μαρτίου 23, 2025, https://arxiv.org/html/2502.04463v2
38. Teaching Large Language Models to Reason with Reinforcement Learning with Alex Havrilla - 680 - YouTube, πρόσβαση Μαρτίου 23, 2025, https://www.youtube.com/watch?v=SokWUnHaoQI
39. [2412.06769] Training Large Language Models to Reason in a Continuous Latent Space, πρόσβαση Μαρτίου 23, 2025, https://arxiv.org/abs/2412.06769
40. [2409.16605] Evaluating and Enhancing Large Language Models for Novelty Assessment in Scholarly Publications - arXiv, πρόσβαση Μαρτίου 23, 2025, https://arxiv.org/abs/2409.16605
41. Zhiyuan Peng - Google Scholar, πρόσβαση Μαρτίου 23, 2025, https://scholar.google.co.kr/citations?user=aXRHp5UAAAAJ&hl=fil
42. AI Creativity Assessment Methods | Restackio, πρόσβαση Μαρτίου 23, 2025, https://www.restack.io/p/ai-creativity-answer-assessment-methods-cat-ai
43. Top Methods for Effective AI Evaluation in Generative AI, πρόσβαση Μαρτίου 23,

2025, https://www.galileo.ai/blog/top-methods-for-effective-ai-evaluation-in-generative-ai
44. Beyond Accuracy: The Hidden Power of Metrics & Tools in AI Development - Omnisearch, πρόσβαση Μαρτίου 23, 2025, https://omnisearch.ai/blog/metrics-tools-in-ai-development
45. LiveIdeaBench: Evaluating LLMs' Scientific Creativity and Idea Generation with Minimal Context - arXiv, πρόσβαση Μαρτίου 23, 2025, https://arxiv.org/html/2412.17596v2
46. Large Language Model Displays Emergent Ability to Interpret Novel Literary Metaphors, πρόσβαση Μαρτίου 23, 2025, https://www.tandfonline.com/doi/full/10.1080/10926488.2024.2380348
47. Large Language Model Displays Emergent Ability to Interpret Novel Literary Metaphors, πρόσβαση Μαρτίου 23, 2025, https://doi.org/10.1080/10926488.2024.2380348