AssisterrAI: The DeAI Gig Economy for Mixtures of Small Language Model (SLM) Agents

Version 1.0: December 2024

Abstract

The current AI industry is heavily centralized and often fails to provide fair compensation to data contributors. It predominantly relies on Large Language Models (LLMs), which are often designed for general-purpose reasoning across diverse contexts. While perhaps effective for broad, summative responses, LLMs fall short in specialized business applications where deeper analysis and domain-specific problem solving are essential. Furthermore, the requirements for running LLMs have become increasingly more complicated and costly, while returns on investment have been unsatisfactory.

Verticalized AI is increasingly envisioned as the future of the industry, where problem-specific, task-focused systems may deliver precise and functional solutions. This paradigm shift is poised to challenge LLM market dominance by offering more cost-efficient and percipient language model alternatives. Small language models (SLMs) show great promise in this area, built to be domain-specific, customizable, high throughput, and revisable. Modern SLMs integrate a Mixture of Experts (MoE) and a Mixture of Agents (MoA) architectures that combine verticalization advantages with the contextual breadth and interactivity of LLMs, all without sacrificing adaptability.

AssisterrAI is at the forefront of the vertical AI paradigm, providing a distributed participant ecosystem for MoA SLMs. We provide no-code infrastructure for the development of SLMs catered to domain-specific use cases. Assisterr's ecosystem is an internal free market based around the peer review, model creation, and data validation of collaborative contributors, producing a self-sustaining decentralized AI (DeAI) economy. This SLM Factory utilizes MoA architectures built by reciprocatively compensated AI development gig workers incentivized by the Assisterr's native token.

Assisterr's DeAI economy combines the multidisciplinary expertise of contributors with our curated development tools, fostering a constantly evolving and rapidly go-to-market (G2M) deployment strategy suited to real-world problems. Distributed participants integrate into Assisterr's multimodal and multi-agent development environments and are given partial ownership of the SLMs. Each model is managed with its own decentralized autonomous organization (DAO) to handle treasury & governance. These models are listed on the marketplace, and crowdsourcing is enabled for datasets, computation power, and other resources. AssisterrAI's on-chain data provenance mechanism ensures the transparency and traceability of all data contributions, validations, and compensation.

TABLE OF CONTENTS

1. Introduction: Technicalities of Language Models

The business world seems on the verge of being transformed by large language models. The recent jump in the capabilities of these models has opened the imagination of developers and investors to a new range of applications. Since the resounding success of OpenAI's GPT-3, capital expenditure in generative AI, in particular LLMs, has ballooned. Though now projected to approach a market size of $1Tn by 2030,[1] persistent technological deficiencies threaten the solution efficiency and returns on investment for LLMs.

1.1 Current Landscape: Challenges Facing LLMs

1.1.1 Technical Problems

Large language models underlie the recent advances in the field and have absorbed the brunt of investment. However, investments in AI have started to outpace returns.[2] This has expanded the gap between speculative and real-world value. At present, models often show poor performance for domain-specific use cases which are contingent on unique details and specialized reasoning.[3] BigTech has driven rapid industry growth by focusing on general-purpose models that fail to meet the specialized needs of business. This problem is a consequence of the following limitations:

1. **Data Exhaustion**: LLMs are running out of available high-quality data. Big Tech companies have harvested much of the textual data available on the internet, limiting future improvements.[4] Estimates suggest that the stock of available data may deplete by 2028, bottlenecking advances.[5] New approaches are needed to meet increasing demands for well-matched training data.

2. **Data Quality & Provenance**: Tracking metadata across transformation steps is essential for ensuring training quality,[6] which is lackluster and costly in current LLM practice. Additionally, models are often tuned to low-quality data, for which curation should be better. This affects LLM reasoning performance and makes high-quality models burdensome to develop.

3. **Distributed Knowledge**: There is no single platform or "source of truth" for knowledge as the world constantly evolves. Rules, regulations, and best practices change, and businesses continually adapt to new challenges. LLMs, trained on static datasets, often must keep up with these dynamics.

4. **Memory & Hallucinations**: Despite their massive parameter sizes, LLMs face memory bottlenecks due to computational demands in real-time applications, especially in agentic models.[7,8] This results from storage and dynamic data sourcing shortcomings. Scaling LLMs with these

[1] The Goldman Sachs Group, Inc. 'Gen AI: Too much spend, too little benefit?' Global Macro Research, Issue 129, 25 June 2024, www.gs.com/research/hedge.html.

[2] Skye Jacobs, 'Big Tech needs to generate $600 billion in annual revenue to justify AI hardware expenditure', TechSpot, July 2024, https://www.techspot.com/news/103699-big-tech-needs-generate-600-billion-annual-revenue.html

[3] Plaat, Aske, et al. "Reasoning with large language models, a survey." arXiv preprint arXiv:2407.11511 (2024).

[4] Metz, Cade, et al. "How Tech Giants Cut Corners to Harvest Data for AI." ARTIFICIAL INTELLIGENCE-AI, 15 Apr 2024, The New York Times

[5] Villalobos, Pablo, et al. "Will we run out of data? an analysis of the limits of scaling datasets in machine learning." *arXiv preprint arXiv:2211.04325* (2022).

[6] Wei, Jason, et al. "Finetuned language models are zero-shot learners." *arXiv preprint arXiv:2109.01652* (2021).

[7] Verma, Ajay. "Memory Management Challenges in Large Language Models." Artificial Intelligence in Plain English, 2021, https://ai.plainenglish.io/memory-management-challenges-in-large-language-models-a54439df39cd.

[8] Zhang, Zeyu, et al. "A survey on the memory mechanism of large language model based agents." *arXiv preprint arXiv:2404.13501* (2024).

handicaps often reduces performance to factual inaccuracies or logical inconsistencies and can show misalignment between user commands and the reasoning aptness of models.[9]

5. **Model Collapse & Retraining**: Retraining LLMs and on recursive datasets has often introduced defects that impair memory and analysis logic, undermining or collapsing task execution.[10, 11] The lack of coherence in generative LLMs can make them incapable of accommodating progressively dynamic scenarios when handling a task.

6. **Privacy & Security**: Centralized LLMs may face issues with data leaking when user data may be extracted for training on future functions. This is a security problem commonly associated with the data protection layer of LLMs. Furthermore, embedded biases are a continuing concern with institutionally backed models, which can influence outputs, perpetuate disinformation, engage in exploitative user data farming, and selective output logic. SLMs can be a solution where only limited data requirements can constrain the gap from which data privacy can be exploited.

7. **Complex Business Use Cases**: Businesses often need AI solutions that go beyond basic model outputs. They require a full solution involving task decomposition, context enrichment with real business data, and automated decision-making. LLMs are not designed to handle the complexity of abstract tasks that require a combination of different processes and automated decisions.[12,13]

8. **Ineffective Enhancements (MoE & CoT)**: While Big Tech has introduced enhancements such as Mixture of Experts (MoE)[14] and Chain of Thought (CoT)[15] to stabilize LLM outputs and improve general reasoning, they are still less effective than SLMs at domain-specific solutions.

Data Exhaustion in Focus:

The trend with LLM evolution has been expanding the recruitment of data for training larger and larger models. The size of constantly scaled-up LLMs and the large amounts of data required make the development pipeline more rigid and stifle quick iterations. This is compounded by the exponentially rising costs associated with progressively larger models. Due to these limitations, businesses are still struggling to build effective in-house AI solutions, which slows overall AI adoption and commercialization. We conclude that to achieve further innovation, a new paradigm is needed. *Figure 1* shows the bottleneck of available data for training scaled-up LLMs expected to be reached.

[9] Plaat, Aske, et al. "Reasoning with large language models, a survey." arXiv preprint arXiv:2407.11511 (2024).

[10] Alemohammad, Sina, et al. "Self-consuming generative models go mad." *arXiv preprint arXiv:2307.01850* (2023).

[11] Briesch, Martin, Dominik Sobania, and Franz Rothlauf. "Large language models suffer from their own output: An analysis of the self-consuming training loop." *arXiv preprint arXiv:2311.16822* (2023).

[12] Hadi, Muhammad Usman, et al. "A survey on large language models: Applications, challenges, limitations, and practical usage." *Authorea Preprints* (2023).

[13] Valmeekam, Karthik, et al. "Large language models still can't plan (a benchmark for LLMs on planning and reasoning about change)." NeurIPS 2022 Foundation Models for Decision Making Workshop. 2022.

[14] Cai, Weilin, et al. "A survey on mixture of experts." *arXiv preprint arXiv:2407.06204* (2024).

[15] Wei, Jason, et al. "Chain-of-thought prompting elicits reasoning in large language models." *Advances in neural information processing systems* 35 (2022): 24824-24837.

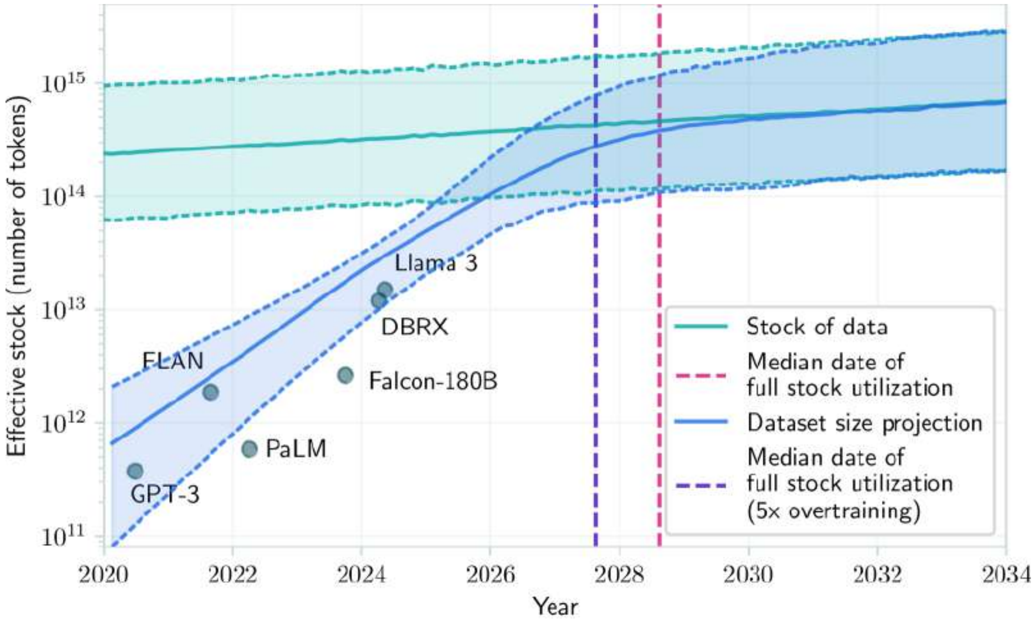


Figure 1: *The online available information constriction expected to exhaust novel LLM training data by 2028 .*[16]

Conservative projections indicate that LLMs will be trained on dataset sizes approximately equal to the total stock of textual data of human origin by 2032 at the latest. More recent analyses press that this will happen in 2028. The effect is that the reasoning capabilities of LLMs risk plateauing irrespective of the number of parameters upon which they are trained.[17] This is because LLMs are partially bounded by neural scaling laws that require increasing training dataset sizes for notable performance gains.[18]

Context Augmentation:

In evolving powerful models, specificity and contextual augmentation of reasoning and outputs must improve This has been incrementally achieved through retrieval-augmented generation (RAG), referring to a process in which queries to a language model are amended with relevant supplementary information from a database. RAG has improved the accuracy and relevance of language model outputs for knowledge-intensive prompts.[19] *Figure 2* presents the general workflow of most RAG systems and how they can be used to upskill a model with supplementary data.

Tools such as Text2SQL can convert natural language into specialized queries that allow more user-friendly retrieval of data. This allows models to interpret real-world representations of processes and objects and leads to more informed and accurate decisions. Interfacing the SLM architecture with databases that can be dynamically updated also creates something akin to a source of truth for the language model. Text2SQL unlocks real-time agentic access to B2B use cases where LLMs may have zero knowledge about internal product data. This domain-specific implementation of Text2SQL has found particular use in SLMs.[20]

[16] *Villalobos, Pablo, et al. "Will we run out of data? an analysis of the limits of scaling datasets in machine learning." arXiv preprint arXiv:2211.04325 (2022)*

[17] "Language Model Scaling Laws." Colab notebook, Google, https://colab.research.google.com/drive/1qv2-hUR5hPqw3OcfmLEhq6Tg15bf06Mj?usp=sharing.

[18] Hoffmann, Jordan, et al. "Training compute-optimal large language models." *arXiv preprint arXiv:2203.15556* (2022).

[19] Gao, Yunfan, et al. "Retrieval-augmented generation for large language models: A survey." *arXiv preprint arXiv:2312.10997* (2023).

[20] GitHub. "GitHub - Anindyadeep/text2sql: Text to SQL using only Small Language Models." GitHub, https://github.com/Anindyadeep/text2sql.

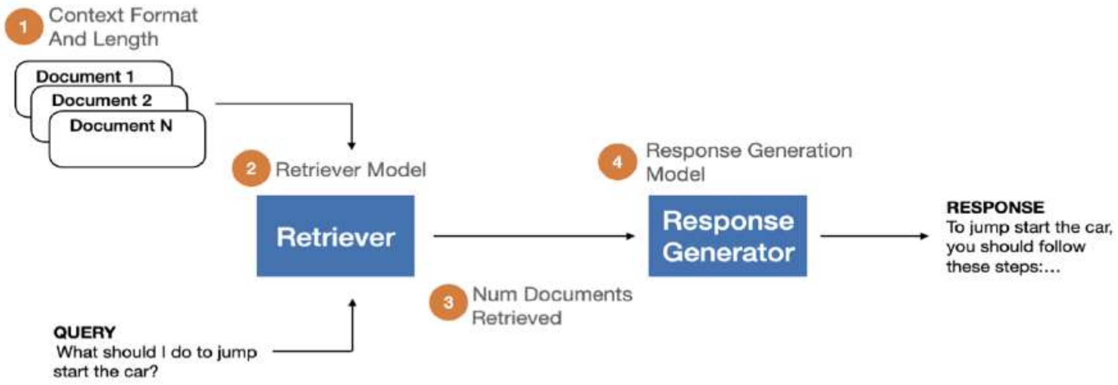
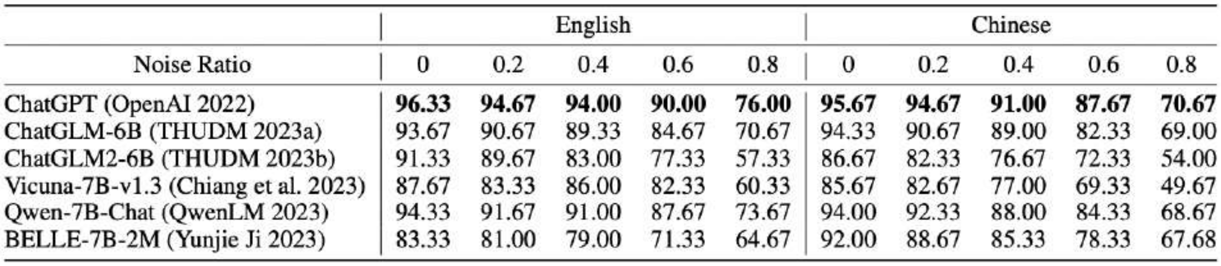

1 Context Format And Length
Document 1
Document 2
Document N
2 Retriever Model
Retriever
4 Response Generation Model
Response Generator
RESPONSE
To jump start the car, you should follow these steps:...
3 Num Documents Retrieved
QUERY
What should I do to jump start the car?

Figure 2: *A basic RAG system workflow with the dependent variable parameters.*[21]

While RAG can improve the responses of LLMs, the increasing noise rates pose persistent challenges. For example, when evaluating external documents for scanning relevant information, the generation of reasonable answers can be impacted by noise. Notably, when the noise ratio surpasses 80%, accuracy drops significantly at a 0.05 significance level. In the case of ChatGPT, accuracy was reduced by nearly 20%, while ChatGLM2-6B's collapsed more than 32%. Results of RAG employed on information sourced between English and Chinese documentation are shown for various models in *Figure 3*, below.



	English					Chinese				
Noise Ratio	0	0.2	0.4	0.6	0.8	0	0.2	0.4	0.6	0.8
ChatGPT (OpenAI 2022)	**96.33**	**94.67**	**94.00**	**90.00**	**76.00**	**95.67**	**94.67**	**91.00**	**87.67**	**70.67**
ChatGLM-6B (THUDM 2023a)	93.67	90.67	89.33	84.67	70.67	94.33	90.67	89.00	82.33	69.00
ChatGLM2-6B (THUDM 2023b)	91.33	89.67	83.00	77.33	57.33	86.67	82.33	76.67	72.33	54.00
Vicuna-7B-v1.3 (Chiang et al. 2023)	87.67	83.33	86.00	82.33	60.33	85.67	82.67	77.00	69.33	49.67
Qwen-7B-Chat (QwenLM 2023)	94.33	91.67	91.00	87.67	73.67	94.00	92.33	88.00	84.33	68.67
BELLE-7B-2M (Yunjie Ji 2023)	83.33	81.00	79.00	71.33	64.67	92.00	88.67	85.33	78.33	67.68

Figure 3: *Though RAG has improved LLM responses in noise robustness, increasing noise rates pose challenges.*[22]

Proponents of LLMs have also often defended the practice of self-training with AI-generated data as another means of resolving insufficient performance. Evidence shows that even this can degenerate model outputs when taken too far. Tools have been built into existing data pipelines to improve such tracking; [23] However, many practices, especially those contributing to the training of LLMs, bundle data without tracking, risking unreliable information.[24] It is also expected that new data added to training corpora in coming years will be of decreasing quality, incorporating less-reliable chat logs, private groups, or convoluted multimodal sources. Revisionary techniques, whether data exhaustion, re-training, or upskilling, continue to struggle with resolving technical LLM deficiencies.

[21] Friel, Robert, et al. "RAGBench: Explainable Benchmark for Retrieval-Augmented Generation Systems" *arXiv preprint arXiv:2407.11005*

[22] Chen, Jiawei, et al. "Benchmarking Large Language Models in Retrieval-Augmented Generation" arXiv preprint *arXiv: 2309.01431v2*

[23] "What is Data Provenance? | IBM." IBM, IBM, 23 July 2024, www.ibm.com/think/topics/data-provenance.

[24] Longpre, Shayne, et al. "The data provenance initiative: A large scale audit of dataset licensing & attribution in ai." *arXiv preprint arXiv:2310.16787* (2023).

1.1.2 Economic Risks of Sustained LLM Development

i. Financing Problems

Doubts are beginning to be raised about the profitability of these ventures. Apart from digital assistants in the form of Chatbots, the spending in generative AI has so far had few applications to show for it. Some have raised concerns that these levels of capital expenditure have created a bubble[25,26] that will prove counterproductive in meeting the highly domain-specific functional needs of the real economy. The AI industry is seeing diminishing returns on large investments as the utility of the existing models is beginning to be reappraised.[27] The gap between the demand of businesses for more specialized AI solutions and large investments in general models is widening.

Furthermore, while the performance of LLMs against standardized benchmarks is still improving, the costs are rising exponentially and will soon become prohibitive. Modern LLMs already have billions of parameters. While early LLMs could be trained for less than $1M, the cost of DeepMind's latest Gemini Ultra reaches into the nine figures.[28] All the while, key technical weaknesses of language models in real-world applications have not been sufficiently addressed. In response to these discrepancies, investment trends have been negative. By December 2024, mergers & acquisitions in AI companies, which greatly boosted corporate monopolies' AI development, dropped 37% from their all-time high. Additionally, venture capital firms have reduced AI deal flow volume by 44%.[29]

ii. Obstructive Practices

Financially and politically influential institutions back the majority of LLM development and have participated in monopolistic and extractive business practices. The centralization of the most powerful industrial tools and funding means that BigTech wields greater consumer control, leaving less diversity in the market. Institutional LLMs have even bundled with other centralized services, such as cloud computing, APIs, or analytics, further limiting consumer options and increasing the exploitation of user data. LLMs have incentivized invasive data collection practices, with companies reusing private user data without compensation. Work that goes directly or indirectly towards the development of better AI models ought to be rewarded. Finally, large incumbents commonly acquire or merge AI start-ups, assimilating tech and hampering both the free market and technological advancement.

iii. Profession & Job Security

LLMs, like generative pre-trained transformer (GPT) chat bots, are increasingly integrated into various fields of work and study with considerable impacts on professional livelihoods. As of Q3 of 2023, approximately 10% of tasks performed by 80% of the U.S. workforce have been negatively affected by AI, with a further 19% at risk of losing their jobs in at least 50% of their workloads to it. Rank-and-file labor, including specialized jobs in erudition, are at risk of replacement by general reasoning engines. Experimenting with new business models capable of LLMs' strategic supplementation as opposed to job

[25] Buchanan, Mark. "The laws of inflating the AI bubble." *Nature Physics* 20.9 (2024): 1362-1362.

[26] Floridi, Luciano. "Why the AI Hype is another Tech Bubble." *Philosophy & Technology* 37.4 (2024): 128.

[27] Thompson, Neil C., et al. "Deep learning's diminishing returns: The cost of improvement is becoming unsustainable." *Ieee Spectrum* 58.10 (2021): 50-55.

[28] Stanford HAI. "AI Index: State of AI in 13 Charts." Stanford HAI, Stanford University, 15 Apr. 2024, hai.stanford.edu/news/ai-index-state-ai-13-charts.

[29] Ma, Hangyu and Zheng, Emily. "AI investment sags as financing, intellectual property issues complicate deals". S&P Global. (2023), https://www.spglobal.com/marketintelligence/en/news-insights/latest-news-headlines/ai-investment-sags-as-financing-intellectual-property-issues-complicate-deals-79386589

replacement has been expensive and difficult to pursue. Despite many proponents of LLMs arguing that jobs will be safe, 10-25% of occupations globally are projected to be displaced within a decade.

1.2 The New AI Paradigm

1.2.1 Verticalized AI

SLMs, while similar in concept to Large Language Models (LLMs) like Chat-GPT or BERT, are designed to be more accurate, specialized, and efficient. By focusing on specific tasks and data sets, SLMs provide superior performance for niche applications, making them better suited for specialized use cases.

Key Differences between SLMs and LLMs:

- **Parameter Count:** LLMs typically contain billions of parameters, whereas SLMs have millions.
- **Training Data:** While LLMs are trained on vast, generic datasets, SLMs utilize smaller, specialized, and curated datasets tailored to specific domains or applications.
- **Customization:** SLMs are highly customizable, allowing for the creation of specialized models tailored to the unique needs of any business vertical.
- **Efficiency:** SLMs have lower latency and are more cost-effective due to lower operational costs and reduced energy consumption.
- **Data Privacy:** Customers may host SLMs in their own secure environments, such as decentralized storage solutions.

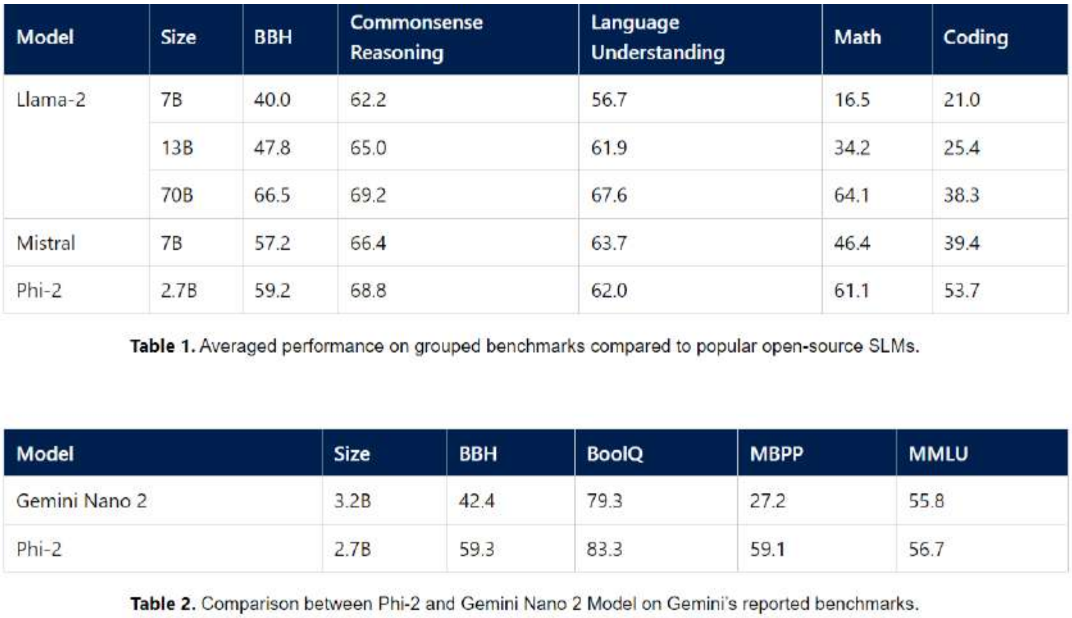
By utilizing tailored datasets and focusing on specific business needs, SLMs can deliver superior performance and situational adaptability at a fraction of the cost. This is also encouraging for open-source SLM building, where cheaper projects have previously developed SLMs with competitive accuracy to veteran LLMs at much lower costs. By the end of 2023, Mixtral 8x7B was an SLM multi-modal architecture that nearly matched the output quality of larger models.[30] Following this development, Zephyr-7B-β outperformed GPT-3.5 turbo, which was approximately 25 times larger.[31] Start-ups and Big Tech alike have begun backing smaller models, developing AI that can execute functions on significantly fewer parameters. Performances trumped those of larger models, such as Llama-2-70B and Gemini Nano 2, in logical analysis, mathematics, and language[32], as seen in *Figure 4*, below. SLMs are positioned to capture the highest ROI in the coming years if their market can reliably capture the breadth of reference ability that traditional LLMs introduce whilst exhibiting their vertical AI proficiencies. This can challenge a $600B industry sector.[33]

[30] Mistral AI. "Mixtral of experts." Mistral AI, 11 Dec. 2023, https://mistral.ai/news/mixtral-of-experts/

[31] Zephyr-7B Beta: An Alternative to ChatGPT." E2E Networks, www.e2enetworks.com/blog/zephyr-7b-beta-an-alternative-to-chatgpt.

[32] Teng, J. (2024) *Small models, big impact: Slms vs. llms*, *Small models, big impact: SLMs vs. LLMs - by Janelle Teng*. Available at: https://nextbigteng.substack.com/p/small-models-big-impact-slms-vs-llms

[33] Cahn, D. (2024) *Ai's $600B question*, *Sequoia Capital*. Available at: https://www.sequoiacap.com/article/ais-600b-question/).



Model	Size	BBH	Commonsense Reasoning	Language Understanding	Math	Coding
Llama-2	7B	40.0	62.2	56.7	16.5	21.0
	13B	47.8	65.0	61.9	34.2	25.4
	70B	66.5	69.2	67.6	64.1	38.3
Mistral	7B	57.2	66.4	63.7	46.4	39.4
Phi-2	2.7B	59.2	68.8	62.0	61.1	53.7

Table 1. Averaged performance on grouped benchmarks compared to popular open-source SLMs.

Model	Size	BBH	BoolQ	MBPP	MMLU
Gemini Nano 2	3.2B	42.4	79.3	27.2	55.8
Phi-2	2.7B	59.3	83.3	59.1	56.7

Table 2. Comparison between Phi-2 and Gemini Nano 2 Model on Gemini's reported benchmarks.

Figure 4: *Comparing SLMs to LLMs according to several performance benchmarks* [34]

1.2.2 Modular SLM Architectures

To address the limitations of LLM-based agents, advanced approaches have emerged involving multiple small language models (SLMs) working in collaborative agentic frameworks. By combining SLMs into agentic ensembles, users describe the problem, and a reasoning process engages various models to analyze, interpret, and provide the best possible solution. This allows AI to reduce the trade-off between general reasoning and in-depth functional solutions by hinging on distributed contextual reasoning across hybridized, domain-specific models. This offers both breadth and depth, whilst enabling selectivity between the participating models in the ensemble. Thus, practical and effective solutions can be explored for specialized and complex problems. Two core approaches are leveraged when developing AI agents from SLM ensembles L: (1) mixtures of experts (MoE) and (2) mixtures of agents (MoA).

Mixtures of Experts (MoE):

When combined in MoE ensembles, modern SLM reasoning can achieve enhanced learning flexibility without losing its capacity for functional problem solving. Ensemble learning can combine the reasoning skills of multiple smaller models, each specialized in different associated contexts, to solve complex problems.[35] This generates a hybrid comprehension that continues to allow the AI to deep-dive. Layers of experts can themselves be composed of MoEs, creating hierarchical structures to further buffer contextual complexity and problem solving proficiency. Such MoE layering is described below in *Figure 5*.

[34] Microsoft Research. "Phi-2: The surprising power of small language models." Microsoft, www.microsoft.com/en-us/research/blog/phi-2-the-surprising-power-of-small-language-models/.

[35] Islam, M.A. (2023) *The art of combining models: Understanding Ensemble Learning in depth*, *Medium*. Available at: https://mdahsanulhimel.medium.com/the-art-of-combining-models-understanding-ensemble-learning-in-depth-7493f8530c3e

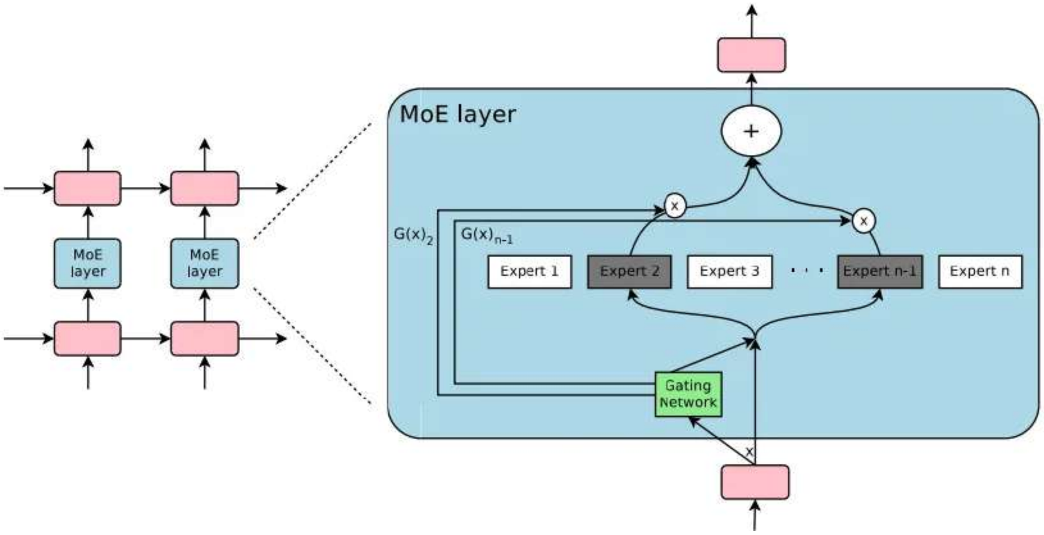


***Figure 5:** MoE layer embedded within a recurrent model in which sparse gating selects 2 experts for computation.*[36]

An MoE typically uses a sparse gating layer that dynamically selects among several parallel networks to give the most appropriate response to the prompt. To achieve more flexible responses, individual experts could be fine-tuned for code generation, translation, or sentiment analysis. More sophisticated MoE architectures may contain several such MoE layers in combination with other components. Like any typical language model architecture, the MoE gating layer operates on semantic tokens and requires training.

Mixtures of Agents (MoA):

When assembled into MoA architectures, SLMs enhance the selectivity of diversified reasoning ensembles, enabling AI to enact precise execution of a task with the required methodology. Agentic models are assembled in a consortium that layers execution protocols to improve efficiency and problem solving of complex tasks. The AI is therefore works in multi-domain scenarios. Teams of agents can work in sequence, iteratively improving upon previous results. MoA has previously significantly outperformed larger models, including GPT-4 Omni's 57.5% accuracy score on AlpacaEval 2.0, even in open-source models.[37]

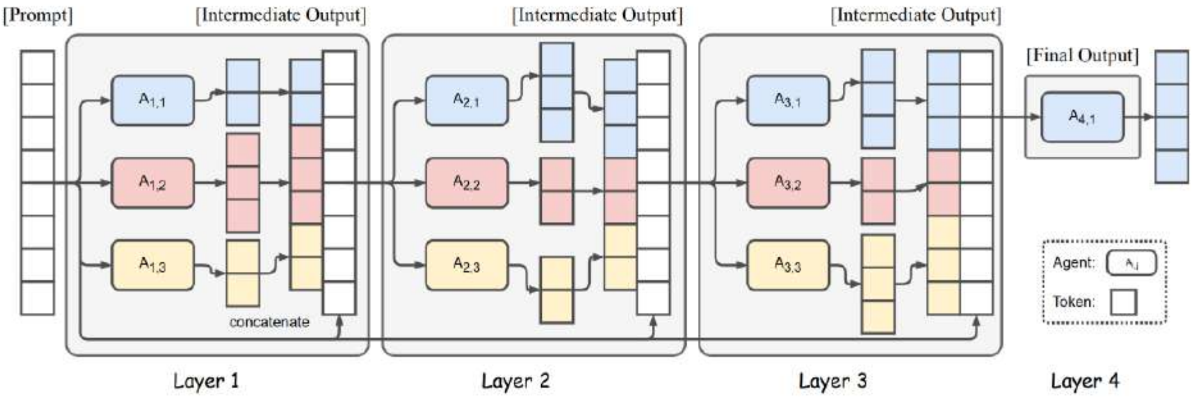


***Figure 6:** The layered structuring of MoA executing functions from prompt to an output consensus.*[38]

[36] *Fedus, William, Barret Zoph, and Noam Shazeer. "Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity." Journal of Machine Learning Research 23.120 (2022): 1-39.*

[37] Wang, Junlin, et al. "Mixture-of-Agents Enhances Large Language Model Capabilities." *arXiv preprint arXiv:2406.04692* (2024).

[38] *Wang, Junlin, et al. "Mixture-of-Agents Enhances Large Language Model Capabilities." arXiv preprint arXiv:2406.04692 (2024).*

A Mixture of Agents (MoA) operates on the level of model outputs, not semantic tokens.[39] It does not feature a gating layer but forwards the text prompt to all agents in a parallelized manner. This structuring is seen in *Figure 6*. Outputs of the MoA are also not aggregated by addition and normalization. Instead, they are concatenated and combined with a synthesize-and-aggregate prompt before being passed on to a separate model to produce the final output. The models are thus divided into "proposers" that compute diverse outputs and "aggregators" that integrate the results. Just like for MoE, several of these layers can be combined. The lack of gating layers makes this architecture an attractive choice because it becomes possible to portably aggregate smaller modules into complex ones.

2. Web3: The DeAI Gig Economy

2.1 Justification for DeAI Gig Economies

A Web3 economy of DeAI gig workers can solve problems associated with the traditional LLM business landscape by guarding jobs, creating new professional opportunities, and accelerating technological development. The decentralization of a contributor layer ensures transparency, accountability, and validation of meritorious achievement, rewarding participants in proportion to their offerings. The emerging vision is to revolutionize the development & deployment of artificial intelligence through this decentralized creation and community ownership of MoA SLMs. By fostering peer review, development collaboration, and data validation amongst vested contributors, human expertise can be pooled towards personalized modern AI with bespoke catering to important problems. Furthermore, contributors are compensated in reciprocation to their value add whilst retaining ownership of the AI, keeping them interested and financially vested in an expanding DeAI free market. This protects jobs, secures new professional opportunities, and creates an industry reliant on user expertise to accelerate AI advancement.

The technical specifications of SLMs, compounded with the modern, verticalized ensemble designs, make it possible to construct a sustainably growing, internal economy of scale. SLMs are smaller, domain-specific, and cheaper. They are therefore faster to build, revise, and deploy whilst now, also leveraging next-generation modular architecting techniques to expand their breadth. Human operators can now even elect to design their MoE and MoA organization, creating novel multi-modal SLM architectures. SLMs also perform in domain-specific use cases more effectively and benefit from human insight and adjustment. Therefore, human expertise can be used to tune models tailor-fit to a supply chain of increasingly complex problems that need answering. This new paradigm in verticalized and modular AI fragments the development space, expanding the surface area for opportunity.

2.2 Technical Industry Advantages

The emergence of MoA architectures for modern SLMs brings several direct, technical benefits to the industrial landscape that alleviate the lasting complications of traditional LLMs:

a) Problem-Solving Competence- SLMs are already better trained for domain-specific use cases and produce more functional, actionable content for problems. Modular architectures expand the breadth of their reasoning capabilities while maintaining or improving the depth of solution finding.

[39] Wang, Junlin, et al. "Mixture-of-Agents Enhances Large Language Model Capabilities." *arXiv preprint arXiv:2406.04692* (2024).

b) **Efficiency: Cost & Computation**- The costs associated with SLM development are substantially lower than for LLMs and have been retained even with mixtures of models. They are faster to produce, require less training data, can be revised more quickly, and are adaptable. Computational burdens are significantly reduced, and infrastructural dependencies are less than in LLMs, further mitigating power consumption-related costs.

c) **Adaptability**- The decentralized nature of blockchain platforms and services increases the opportunity pool for domain-specific catering, making SLMs well suited to them. On-chain processing is also more easily managed by smaller models than by LLMs. Whether Web2 or Web3, heterogeneous populations of nodes or different online protocols may have varied compatibility requirements. SLMs can be more readily produced and improved upon to cater to these niches.

d) **Data Privacy**- SLMs only require enough information with which to handle specialized problems. Datasets can be deployed locally, enabling greater control over data. This also limits their exposure to external threats, unlike LLMs, which must recruit data openly and generally have poor provenance, requiring the tracking of metadata cross-processing steps.

e) **Scalability**- Parallel computing, collaboration, and data curation are all streamlined when developing or running SLMs because of their less computational load, memory, and size. Operational efficiencies also reduce latency and make collaboration in real-time more possible. The high throughput and rapid go-to-market advantages make the scalability of an SLM-dependent ecosystem considerable.

f) **Environmental Sustainability**- Energy costs and the associated emissions with running smaller language models are greatly reduced. Smaller data packages are also less burdensome and easier to store or transfer, further emissions-related impact on the environment. Hardware shelf lives are extended when used for SLMs, which makes them last longer, while the constant manufacture of replacement devices can be reduced.

2.3 Added Coordination Power of Web3

Web3 business models and infrastructure introduce unique modes of coordination across services and devices that may uniquely buffer the value add of DeAI SLMs.

2.3.1 Infrastructure:

Another area of potential synergy between Web3 technology and AI is the decentralization of AI infrastructure.[40] Decentralized physical infrastructure (DePIN) is one of the dominant narratives of the current market cycle. It refers to the use of blockchain networks and other distributed computing technologies to manage compute resources. DePIN networks are on the path to creating an open and competitive market for these resources and breaking into the monopolies of established cloud service providers. The resulting efficiency and flexibility gains will become important as the training and execution of AI models are rapidly increasing energy demands. The recent diversification of Bitcoin miners into AI computation already demonstrates that similar economic incentives are at play.[41]

[40] Kersic, Vid, and Muhamed Turkanovic. "A review on building blocks of decentralized artificial intelligence." arXiv preprint arXiv:2402.02885 (2024).

[41] Butterfill, James et al. "CoinShares Bitcoin Mining Report Update: Our Insights at the 2024 Halving" CoinShares Research Blog, 2024, https://blog.coinshares.com/coinshares-mining-report-the-halving-and-its-impact-on-hash-rate-and-miners-cost-structures-8646835d88ac.

2.3.2 Information:

Besides problems with low-quality data sources, it is also difficult to ensure that the information contained in training data is veridical and up to date. There is no single source of truth that language models or AI agents can draw on to faithfully access facts. This is an inherent limitation of Web2 datasets, but one that could be solved by integrating blockchain databases into the design of AI systems.[42] However, this integration of blockchain oracles will require smaller and more modular systems that the currently dominant LLM architectures.

3. AssisterrAI: Technology & Ecosystem

AssisterrAI is the culmination of the two emerging trends that will shape the future of AI. The first is the transition from expensive, general-purpose LLMs, which are reaching an innovation plateau, to SLMs, their small domain-specific cousins. The second is the decentralization of training, reasoning and data ownership to create a fair and open AI gig economy. Besides these two core innovations, the Assisterr ecosystem will offer a framework to build both agentic AI and passive chatbots. *Figure 7* outlines the structure of the Assisterr platform from models' creation to use within the DeAI economy.

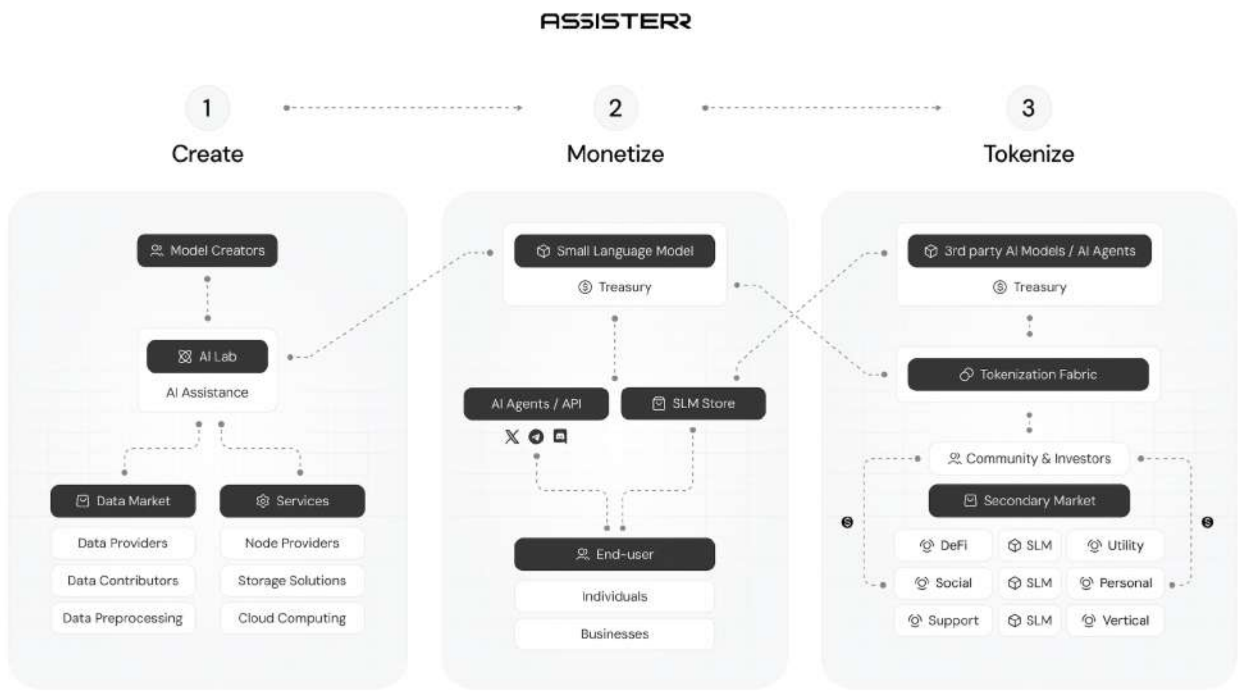


Figure 7: *High-level overview of the AssisterrAI ecosystem's DeAI facilities, illustrating the creation and financial utilization of SLMs and MoA techniques towards solving complex, functional solution-requiring problems.*

3.1 AI Lab

AssisterrAI provides a unified infrastructure pipeline to create, tokenize and distribute SLMs in a way that incentivizes all community contributions. Our *AI Lab* will allow users to contribute to models in their

[42] Soldatos, John, et al. "Blockchain based data provenance for trusted artificial intelligence." *Trusted Artificial Intelligence in Manufacturing: A Review of the Emerging Wave of Ethical and Human Centric AI Technologies for Smart Production* (2021): 1-29.

knowledge area. It will let them become both co-creators and co-owners of the AI. This is motivated by our belief that the AI gig worker ought to not only earn on a one-time, transactional basis but capture a wider value from the market; This will secure a better future and make people the beneficiaries of AI, not a victim of progress and automation.

In order to access the platform, users connect a browser-based Solana wallet as well as their X profile and Discord account. They can then create models through the *AI Lab* tab of the Assisterr user interface. It currently offers a simple form to specify key parameters, prompt templates, and the metadata of the model. It also allows the user to directly upload data that will be embedded in the model through retrieval augmented generation (RAG) and later through fine-tuning. Upon creation, the model can immediately be made public through the SLM store.

In the future, the AI Lab will be based on a modular, multi-model paradigm with a Mixture of Agents architecture and augmented retrieval strategies. We aim to solve real-world problems with deeper context and complex, domain-specific reasoning. Following this philosophy, it is Assisterr's approach to define real-world use cases and break them down into sub-tasks. These tasks are then used to set rules and workflows that ensure that the SLM-powered Agent delivers an end-to-end solution for each use case. A reasoning process engages various models to analyze, interpret, and provide the best possible solution.

AssisterrAI will use a modular SLM architecture to address the limitations of general-purpose LLMs in business applications. We believe that these limitations have to be tackled through carefully tailored, domain-specific SLMs which can be combined into modular, agentic frameworks that meet the needs of real-world applications. Various demonstrations of the reasoning capabilities of small models have already been made[43,44] and the effectiveness of contextual fine-tuning is well known.[45,46] SLMs also have lower latency, operational costs, and reduced energy consumption. Furthermore, owing to this smaller hardware footprint, it becomes feasible for customers to self-host models in their own execution environments and maintain data privacy.

3.2 SLM Store

Assisterr contributors will be rewarded for all steps in the genesis of an AI model, ranging from data contribution and model creation to validation and review. This revenue-sharing mechanism will be implemented through an SLM tokenization module. The *AI Lab* will also connect business use cases effectively with their required data and expertise. Once a model shows up in the SLM Store tab of the Assisterr interface, any user can query it through a chatbot interface. Currently, bots that assist with various niches in Web3 ecosystems, Healthcare, Software development, and Finance are available.

Every model in the SLM store comes with a treasury denominated in Assisterr's native token. This treasury is topped up from the respective user's balance upon every query that is made. However, queries cannot only be placed from the WebUI with a connected Solana wallet, but also through an API, which makes models from the SLM store accessible through other applications.

[43] Li, Liunian Harold, et al. "Symbolic chain-of-thought distillation: Small models can also" think" step-by-step." arXiv preprint arXiv:2306.14050 (2023).

[44] Magister, Lucie Charlotte, et al. "Teaching small language models to reason." arXiv preprint arXiv:2212.08410 (2022).

[45] Tinn, Robert, et al. "Fine-tuning large neural language models for biomedical natural language processing." Patterns 4.4 (2023).

[46] Thirunavukarasu, Arun James, et al. "Large language models in medicine." Nature medicine 29.8 (2023): 1930-1940.

Contributors will be able to create SLMs, assemble them to agents, and deploy them, all through a no-code interface. This gives creators a quick go-to-market period and a fast innovation cycle. It solves the distribution and monetization challenges faced by independent model creators and developers.

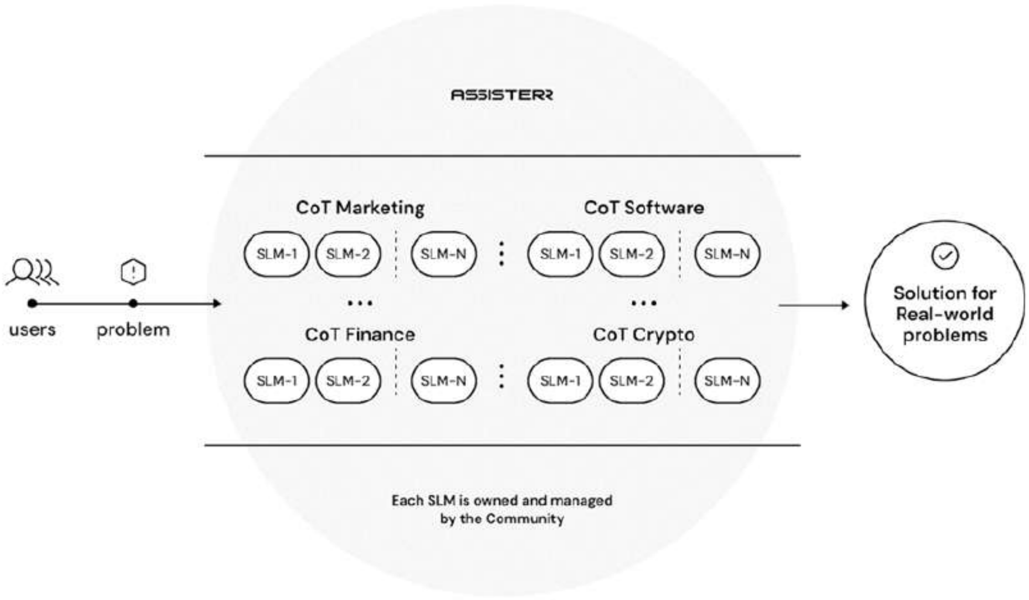


Figure 8: *Image Source: AssisterrAI. Examples of chain-of-thought applicability across marketing, software, finance, and crypto in SLMs created by the distributed participant base in response to growing real-world problems.*

As seen in *Figure 8*, each SLM deployed on the marketplace can participate in MoA architectures. Because these ensembles hybridize reasoning and problem solving proficiencies across multiple models, opportunities are compounded. This augments the potential rewards for contributors by allowing their creations to not only be stand-alone solutions, but parts of a whole with selective applicability.

3.3 Collaborative Elements

Through the Contribute and Earn tab, users also will be able to participate in iterative improvements to existing models from the SLM store by fulfilling data requests and validating performance metrics in exchange for management tokens (MTs) or the native Assisterr token. Due to this peer review, there will be constant evolution and increased throughput in model creation over time. In combination with features such as MoA, this will allow for cumulative progress and continuous bottom-up tinkering. Due to the modular and specialised nature of SLMs, models can also be rapidly integrated into existing work pipelines. In the future, any business or individual will be able to describe their problem and Assisterr's services will handle the involvement of a relevant pool of SLMs/Agents to find a solution.

There are also so-called use case validators. These will initially consist of a permissioned set of external experts or act on demand from SLM Creators. These subject matter experts verify or reject data to be used by SLMs. Furthermore, they actively check for model errors, define what additional data is required, and identify which parts of the dataset may have caused anomalous behavior. End users will also be incentivized to improve data integrity by submitting reports of hallucinations and other spurious outputs.

3.4 Assisterr Treasury Model

The native Assisterr token is the vehicle upon which AssisterrAI ecosystem operations are run. It is transacted in response to the validation of actions taken in fulfillment of smart contract protocols at each stage of the SLM development process. By leveraging the token, participants are able to engage with the facilities of the Assisterr ecosystem, such as accessing products, paying fees, and contributing to SLMs' creation, management, and monetization.

The interaction of the Assisterr token across the platform's facilities hinges on the Assisterr Treasury Model (ATM), described in *Figure 9*. It is designed for adaptability to diverse use cases, enabling flexible governance, a scalable treasury set-up, and a fair rewards system. The execution of this model is composed of 3 progressive stages through which the SLM lifecycle framework is managed:

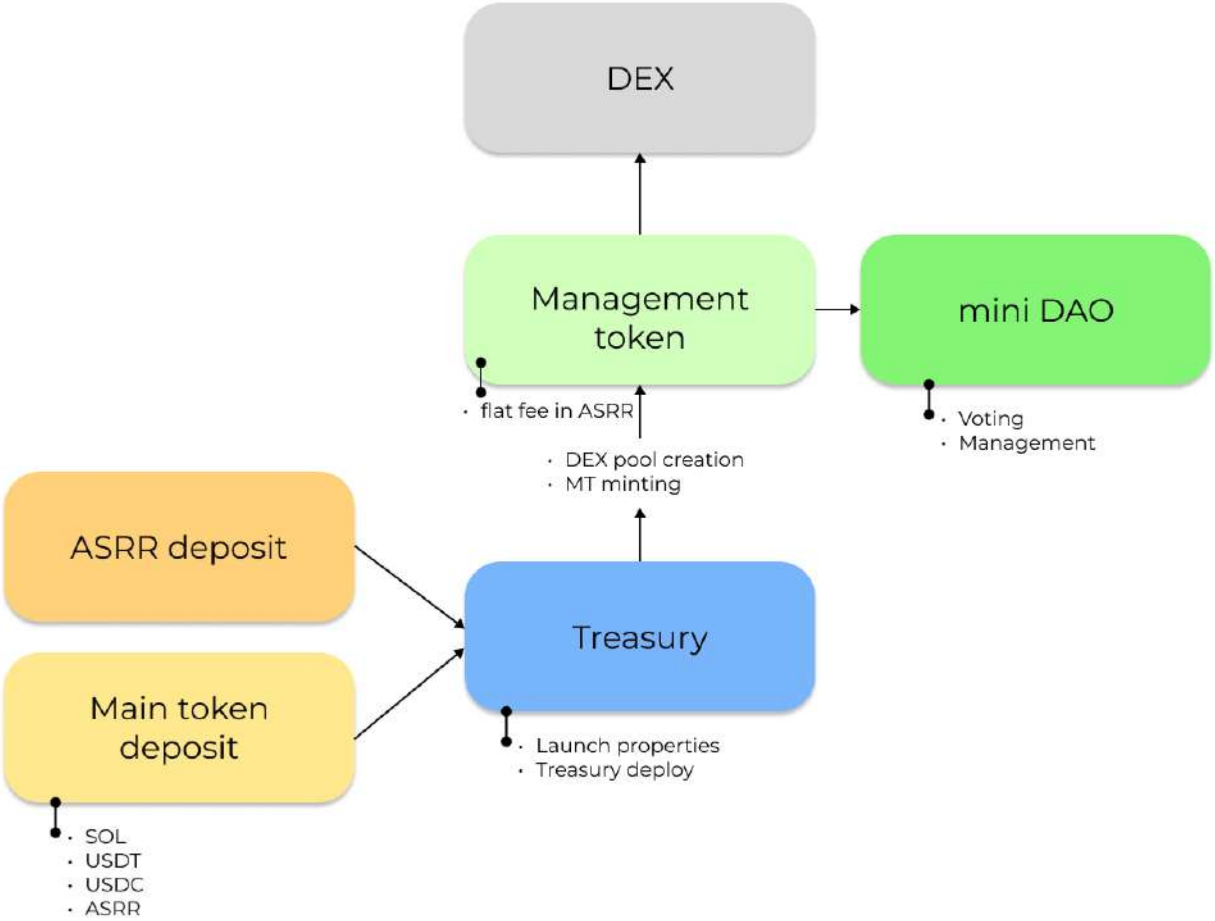


Figure 9: *The Assisterr Treasury Model illustrates ASRR native token transfer across the AssisterrAI ecosystem's facilities.*

Stage 1: Foundation Set-Up

Creators establish the seminal rules and conditions for the treasury, whose parameters define the governance, funding, and operational structure of their model.

Features:

i. Token Parameters- The name, hard supply, and initial liquidity of the SLM creator's Management Token (MT) for their respective model is conceived.

ii. Creator's Allocation- The creator's MT share is confirmed, ensuring initial control and motivation for the development process.

iii. Governance Rules- Critical voting parameters are set, such as the quorum percentage prerequisite for initializing treasury decisions.

iv. Initial Fee Payment- The native token set-up fee is paid, granting access to the ATM.

Stage 2: Crowdfunding & MT Mint

Decentralized participation in the SLM's evolution is enabled by unlocking its MTs for acquisition by interested contributors, thereby granting them co-ownership rights. This stake encourages continued involvement and enables the sharing of rewards and governance rights.

Features:

i. Crowdfunding Mechanism**-** Contributors will purchase MTs according to the predefined crowdfunding conditions set by the creator. This adds treasury liquidity.

ii. MT Allocation- Co-ownership, rewards, and governance rights are enabled.

iii. Decentralized Distribution- Fair and transparent MT access and transactions are ensured to create a trustworthy collaboration in the project's growth.

Stage 3: Collaboration and Development

Once fundraising has concluded, the treasury transitions to the management phase, during which MT holders may now collaborate on the development of the model.

Features:

i. Collaborative Management- MT holders form a mini-DAO specific to their SLM, allowing them to participate in governance on the growth, development, and promotion of the project.

ii. Voting- Governance is facilitated via a voting mechanism in accordance with quorum requirements.

iii. Reward Sharing- Value capture of the model generates returns in the treasury that is subsequently distributed amongst co-owners in proportion to the MT they hold.

iv. Secondary Market- MTs can be traded on secondary markets along with a fee payment in the native token.

3.5 Recent AssisterrAI Solutions in Review

AssisterrAI has several noteworthy use cases for its SLM and SLM-MoA projects that can be completed in a decentralized economic model.

. Decentralized Finance (DeFi) Management Agents:

Decentralized finance (DeFi) AI agents are a formidable emergence in the Web3 space. Moving ahead from general-purpose recommender systems and innovations in account abstraction, specialized AI that operates within safe, permissioned constraints can better optimize and automate financial portfolios. When agentic SLMs are regularly created to cater to especially rapid-transaction media, such as Solana DeFi protocols, lending/borrowing, perpetual trading, staking, and more can envision a future with better data curation, multimodal reasoning, and deep, functional analysis learned through SLM ensembles and executed via modern MoA consortia.

. Trading Agents:

Agents that have been verticalized towards complex trading scenarios, including the analysis of wallet clusters and price action trends, can be highly useful in the volatile DeFi market and traditional finance (TradFi) world. SLM-based MoA can be especially useful in quality data-referenced trading strategies where the medium and manner of execution matter.

. Autonomous Chat Agents:

The development of autonomous chat agents with higher degrees of learning and analytical proficiency is valuable in a world where human operators, across academic, social, and professional arenas require supported thinking. They can additionally be used as support proxies for a host of services, connecting to social networks and IT apps. By adding agentic functionality, actionable, conversational support models can work as liaisons that regularly implement functions with user feedback.

. Public-Facing Avatars:

Agents can be built as text-based, audio-based, or video-based proxies. This allows SLMs to produce avatars for deep-dive, public-facing works. Complex utilities, such as 3D avatars, generation of autonomous text-to-video, and livestream integrations on social platforms, where next-generation multimodal interactions may appear, may benefit from SLM-based MoA.

. Developer Relations- *a Compelling Case Study*:

The launch of a specialized Web3 Developer Relations (DevRel) proof of concept on the AssisterrAI platform provided indicators of strong market fit. A robust DevRel regime is vital to ensure that developers are engaged and given comprehensive support when adopting a technology stack; however, this comes with substantial costs. Salaries for DevRel roles alone range from $90k - $200k per year. Though these expenses suggest high-rigor and strenuous tasks, the majority of developer support requests are predictable and could easily be automated. Therefore, there is scope for increasing DevRel efficiency through the targeted use of SLMs.

Value Proposition:

AssisterrAI enables blockchain networks to develop SLMs tailored to predictable and routine DevRel tasks, which can result in the automation of up to 95% of support requests. Models are quick and easy to deploy, training on a range of sources:

. historical DevRel inquiries

. developer documentation
. on-chain data
. other relevant datasets

Once implemented, models can offer fully customizable utilities for UI/UX as preferred for a project's theme and interface.

Conclusion

AssisterrAI is transforming industrial AI proficiencies and economics by addressing pivotal issues stemming from centralization and large language model technicalities. These included data accessibility and equitable compensation. By providing no-code infrastructure for SLMs and an incentivization DeAI-based reciprocating remuneration, developers are enabled to develop, own, and benefit from the continuous evolution of next-generation SLM-based MoA projects as they cater to specific business cases.

As Assisterr continues to integrate with key blockchain and AI projects and expand its ecosystem, the platform will remain dedicated to delivering decentralized, open AI. We believe that open-sourcing AI is essential not only for realizing its immediate benefits but also for preventing monopolization by large technology firms and mitigating job risk while creating a future of new professional opportunities and erudition for society.

Acknowledgements

This v1 white paper was collaborated upon by the teams at AssisterrAI and the Cambridge Blockchain Labs (CBL). AssisterrAI has been developing in-depth technologies & enterprise. The core team provided profound domain expertise and their strategic vision, upon which the CBL conducted deep-dive research, technical analysis, and marshaled the business plan. -December 11th, 2024