Advancing Agentic AI: Decentralized and Verifiable Collaboration for Next-Generation Foundation Model Development

Anonymous Authors[1]

Abstract

Foundation models such as large language models have achieved remarkable performance by leveraging massive centralized datasets and compute. However, concerns around data privacy, governance, and trust motivate new *agentic* workflows where multiple parties (agents) collaboratively develop models without central custodians. We propose a decentralized framework for verifiable multi-agent model training that integrates federated learning, distributed ledger technologies, and knowledge distillation. In our approach, each participant maintains local data and models, contributing updates that are logged on a tamper-proof DAG ledger for transparency and accountability. A voting-based consensus mechanism enables multi-agent governance, ensuring only high-quality model updates are merged. To aggregate knowledge from diverse sources, we employ cross-silo knowledge distillation, including distilling large teacher models (e.g. LLaMA, BioGPT) into smaller models in a federated setting. Empirical evaluations on collaborative learning scenarios – including named entity recognition (F1=96.23%), medical code classification (F1=79.11%), and question-answering tasks – demonstrate that our decentralized training achieves performance comparable to centralized methods while preserving privacy and trust. This work advances agentic AI by enabling next-generation foundation model development through privacy-preserving, trustable collaboration.

1. Introduction

Recent progress in AI has been propelled by *foundation models* trained on vast centralized datasets, exemplified by large language models (LLMs) like GPT-4. However, centralization poses challenges in data privacy, scalability, and trust in model development. High-stakes domains (e.g. healthcare) demand collaborative model training approaches that keep data decentralized and ensure verifiability of contributions. Meanwhile, the emergence of *agentic AI*—autonomous AI agents that can plan, act, and collaborate (as seen in ReAct and AutoGPT frameworks)—suggests a future where AI systems themselves coordinate complex workflows. Harnessing this paradigm for model development requires new workflows that allow multiple agents or institutions to jointly train AI models in a trustless environment.

Federated learning (FL) offers a starting point by enabling distributed training without sharing raw data. In classical FL, a central server aggregates model updates from clients (participants), mitigating data privacy concerns. However, standard FL assumes a trusted server and lacks mechanisms for transparent accountability or peer governance. Furthermore, training large foundation models in a federated manner presents challenges in communication and heterogeneity. Prior work has begun exploring federated training for LLMs, but ensuring trust among participants remains an open problem.

In this paper, we propose a novel **decentralized and verifiable collaborative learning framework** that enables multiple agents to jointly develop foundation models while preserving privacy and establishing trust. Our approach integrates three key components: (1) **Privacy-preserving decentralization** via a peer-to-peer federated learning protocol with no central coordinator, (2) **Verifiable ledger-based trust mechanisms** using a distributed ledger to immutably log model updates and cryptographically verify contributor identities, and (3) **Multi-agent governance** through a consensus process wherein participants vote to accept or reject proposed model updates. Additionally, we leverage **knowledge distillation** to combine insights from diverse models or tasks: for example, distilling knowledge from large expert models into a collaboratively learned model.

Our contributions are summarized as follows: (i) We design a decentralized training architecture that combines IOTA's Directed Acyclic Graph (DAG) ledger (Tangle) (Popov, 2018) and IPFS storage (Benet, 2014) to enable transparent, auditable model sharing without a centralized server. (ii) We implement a voting-based update validation scheme that empowers multiple agents to govern the training process collectively, improving robustness against low-quality or malicious updates. (iii) We demonstrate through ex-

periments on natural language processing tasks that our approach achieves high accuracy (e.g. NER F1 96.23%, ICD coding F1 79.11%) on par with centralized baselines, while inherently providing data privacy and an audit trail of model provenance. By facilitating trustworthy collaboration among distributed agents, our work paves the way for next-generation agentic workflows in foundation model development.

2. Related Work

Federated Learning and Decentralized Training: Federated learning allows collaborative model training across clients holding private data. Early work by McMahan et al. (2017) introduced the FedAvg algorithm for averaging distributed model updates from decentralized data. Subsequent research has addressed challenges like communication efficiency, statistical heterogeneity, and scalability to large models. Recent efforts (Sani et al., 2024) demonstrate the feasibility of federated training for large language models, showing that with appropriate strategies performance can approach centralized training. Our work builds on the FL paradigm but removes the central server, instead using a peer-to-peer ledger for coordination and trust.

Distributed Ledger and Trust in Collaboration: Blockchain and distributed ledger technologies (DLTs) have been proposed to enhance security and trust in federated learning. By logging transactions immutably, a ledger can verify the provenance of model updates and prevent tampering. However, traditional blockchains (e.g. Ethereum) suffer from throughput and cost limitations for frequent model updates. We adopt IOTA's DAG-based ledger (the Tangle) (Popov, 2018), which offers feeless transactions and high scalability, to record model update metadata. Prior works have used IOTA and similar DLTs for secure data sharing in IoT and healthcare settings, showing the potential for lightweight consensus without miners. Our framework leverages a DLT not just for security but also to enable a consensus-driven workflow where participants actively validate each update.

In conjunction with the ledger, we use IPFS for decentralized storage of large model artifacts. IPFS provides content-addressable, peer-to-peer file sharing (Benet, 2014), which has been integrated with blockchain in previous systems to manage data in distributed machine learning. By using IPFS, we avoid expensive on-chain storage; only small hash pointers are recorded on IOTA, similar to techniques in prior work. This ensures that even large foundation model checkpoints can be shared efficiently and verified by hash.

Knowledge Distillation and Multi-Source Learning: Knowledge distillation (Hinton et al., 2015) is a technique where a "student" model learns to imitate the outputs of a "teacher" model, often to compress a large model's knowledge into a smaller one. In federated settings, distillation has been explored to aggregate knowledge without exchanging weights, for instance by sharing soft predictions (Li & Wang, 2019). Our approach uses distillation in two ways: (a) to allow participants with heterogeneous models to share knowledge, and (b) to incorporate external knowledge from large foundation models. For example, we show that a federation of agents can distill answers generated by powerful models like LLaMA (Touvron et al., 2023) and BioGPT (Luo et al., 2022) into a more compact model suitable for distributed training. This multi-source learning via distillation complements gradient-based update aggregation, enabling what we term an "agentic ensemble" of models contributing to a joint task.

Agentic AI and Collaborative Autonomy: The notion of autonomous AI agents coordinating tasks has gained popularity through approaches like ReAct (Yao et al., 2022) and autonomous GPT-based systems (AutoGPT) (Yang et al., 2023). These works illustrate how LLMs can be endowed with decision-making and tool-use capabilities to accomplish goals in a more open-ended, self-directed manner. In our context, each participant in the collaborative training network can be viewed as an autonomous agent that not only trains on local data but also evaluates and decides on others' contributions. Our framework's consensus mechanism instantiates a simple form of multi-agent negotiation: agents vote on whether a candidate model update should be accepted. This resonates with the vision of agentic workflows where AI systems interact under certain rules to achieve a collective objective. Our work bridges that vision with federated learning by providing the infrastructure (ledger, protocols) for such agent interactions to result in a coherent, high-performing global model.

3. Proposed Framework

Figure 1 illustrates the overall architecture of the proposed collaborative training framework, representing a decentralized agentic workflow for foundation model training. The system consists of multiple participant nodes (agents) connected in a peer-to-peer network. Each agent possesses a local dataset and maintains its own model instance. Training progresses in iterative rounds without any central server: in each round, one or more agents propose model updates which other agents then verify.

Decentralized Update Sharing: When an agent finishes a local training epoch (or other trigger conditions are met), it generates an update Δw (e.g., model weight differences or a complete new model version). Instead of sending this to a server, the agent announces the update to the network by publishing a message to a pub-sub channel (for example,

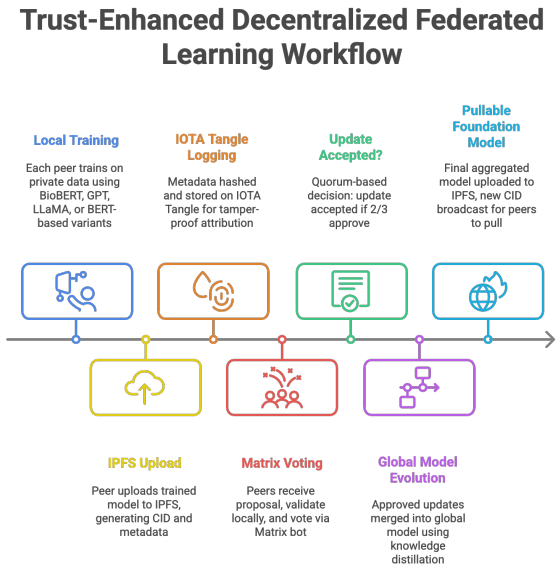

Figure 1. System architecture of the decentralized agentic framework. Each agent trains locally and shares model updates using IPFS and IOTA. Voting through Matrix ensures collaborative validation. Knowledge distillation enables cross-agent knowledge transfer and global model improvement.

using a Matrix peer-to-peer messaging room). The update payload itself (which could be tens or hundreds of MBs for a large model) is uploaded to IPFS, which returns a content-addressed hash (CID). The agent then creates a transaction on the IOTA ledger containing metadata: the CID of the update, a digital signature (using the agent's private key, tied to a Decentralized Identifier), and references to two previous ledger transactions. Posting this transaction "anchors" the update in the ledger's DAG, making it visible to all participants. The use of IOTA's Tangle means each new update helps validate prior updates; thus agents inherently contribute to consensus by issuing transactions.

Let $\Delta w_i^{(t)}$ denote the model update from agent i at round t. Instead of sending $\Delta w_i^{(t)}$ to a central server, the agent publishes it to a peer-to-peer network:

$$\Delta w_i^{(t)} = w_i^{(t)} - w_i^{(t-1)}$$

The update is uploaded to IPFS, generating a content-addressed hash $h_i^{(t)} = \text{Hash}(\Delta w_i^{(t)})$, and a transaction is created on the IOTA ledger with this hash.

Immutable Logging and Identity: By reading the ledger, any participant can retrieve a chronological log of all proposed updates. Because each transaction is signed and linked, the history forms an immutable audit trail. Identities of agents are managed via DIDs (Decentralized IDs) embedded in ledger transactions, which allows attributing each contribution to a pseudonymous but consistent identity. This discourages malicious behavior, as misbehaving agents (e.g., submitting bogus updates) can be identified and potentially blacklisted. The ledger thus provides accountability and provenance tracking for model evolution. Notably, unlike a traditional blockchain, IOTA's feeless design keeps this logging lightweight: agents only incur a minor computational cost for anti-spam proof-of-work and no monetary fees.

Multi-Agent Validation and Consensus: A core feature of our framework is that model updates are not automatically accepted. We implement a *voting-based consensus mechanism* inspired by trust in multi-agent systems. When an agent receives notification of a new update, it fetches the update from IPFS using the provided hash and evaluates it on a validation dataset. This could be the agent's own private validation set or a shared public validation set agreed upon in advance. The agent then casts a vote (e.g., by submitting a signed message or a ledger transaction) indicating acceptance or rejection of the update. An update remains in a tentative state until a quorum of agents (for example, a majority of the participants) have evaluated it within a fixed window. If the required number of positive votes is reached, the update is considered approved and is merged into the global model state. Otherwise, the update is aborted or postponed.

Each agent j evaluates the update $\Delta w_i^{(t)}$ using a validation function $V_j(\cdot)$, and casts a binary vote $v_j^{(t)} \in \{0, 1\}$. The update is accepted if:

$$\sum_{j=1}^{N} v_j^{(t)} \geq Q$$

where Q is the quorum threshold (e.g., $Q = \lceil N/2 \rceil$).

This consensus process ensures that only high-quality contributions (those that improve or at least do not significantly degrade the model's performance) are integrated. It is especially crucial in an open collaboration where some agents might have noisy data or even act adversarially. By requiring agreement, the system adds a layer of robustness on top of standard federated averaging. In effect, model aggregation becomes meritocratic: an update's influence depends on peer validation, not just its submission.

Incentives and Agent Behavior: Although a full incentive mechanism is beyond our current scope, we anticipate that the transparent logging and validation could naturally encourage cooperative behavior. Agents gain trust or reputation as their contributions get consistently accepted (this

reputation could be quantified by tracking successful updates per agent). In contrast, an agent whose updates are frequently rejected might lose trust, which could lead others to scrutinize or ignore its future proposals. In a practical deployment, one could incorporate token rewards or reputation scores on the ledger to further incentivize useful contributions, similar to how miners are rewarded in blockchains. Our framework lays the groundwork for such extensions by having the necessary data (who contributed which update, and how it was evaluated) recorded on an immutable ledger.

Integrating Knowledge Distillation: Beyond gradient updates, our workflow allows knowledge transfer through distillation, broadening the ways agents collaborate. For instance, agents can exchange synthesized data or soft predictions to teach each other. In our experiments, we explore a use-case where two large models serve as teachers: a general-purpose LLM (LLaMA) and a domain-specific model (BioGPT) generate answers to a set of questions. These answers are then used by agents as training data to distill a student model that can perform the QA task. The distillation process is carried out in a distributed manner: each agent only sees a subset of the Q&A pairs, yet by sharing the distilled model updates and validating them, the agents collectively produce a single student model that captures knowledge from both teachers. This showcases that our framework can support multi-source learning, effectively merging expertise from different models held by different parties.

Given teacher outputs $\{y^{(t)}\}_{t=1}^{T}$ from LLaMA and BioGPT, the student model f_s is trained to minimize the distillation loss:

$$\mathcal{L}_{\text{distill}} = \sum_{t=1}^{T} \text{KL}(f_s(x^{(t)}) \parallel y^{(t)})$$

where KL is the Kullback-Leibler divergence between student and teacher output distributions.

4. Experiments

We evaluate the proposed framework on three collaborative learning scenarios spanning typical NLP tasks. The experiments address two key questions: (1) Can decentralized training with our trust mechanisms achieve model performance comparable to centralized training? (2) Does the multi-agent consensus effectively safeguard against detrimental updates without hindering learning progress?

Setup: We simulate a network of $N = 5$ agents (to emulate five institutions or autonomous systems) collaborating on each task. All experiments are implemented using PyTorch with the federated/deliberation logic built atop the IOTA and IPFS APIs. Each agent runs on a separate process (with one GPU per agent for efficiency) and communicates through an off-chain messaging service (we used a private Matrix server for controlled simulation). Model updates are shared and logged via a local IOTA testnet deployed for the experiments. For consensus, we set the voting quorum to > 3 out of 5 agents (i.e., at least 4 votes needed to accept an update), and a validation window of 5 epochs. If an update is rejected, the proposing agent continues training on its local data and may re-submit in a later round.

Tasks and Models: (1) *Named Entity Recognition (NER):* Agents jointly train a named entity tagger for clinical text. We use a bidirectional Transformer (base-sized) with a token classification head. Each agent has access to a portion of a medical NER dataset (simulated by partitioning the 2010 i2b2/VA Clinical Challenge dataset, which contains de-identified patient notes annotated with entities like problems, tests, treatments). The data is non-iid: some agents' data is rich in lab test entities, others in diagnoses, etc., mirroring realistic heterogeneity. (2) *ICD-10 Code Classification:* Agents train a text classification model to assign ICD-10 diagnostic codes based on patient discharge summaries. We use a PubMedBERT-based classifier. The MIMIC-III clinical notes dataset is partitioned among agents, each with a different subset of hospital departments to reflect distribution shift. (3) *Question Answering (QA) via Distillation:* This scenario has agents learning to answer medical questions by distilling knowledge from two large models. One teacher is LLaMA-7B (Touvron et al., 2023) with general world knowledge, and the other is BioGPT (Luo et al., 2022) specialized in biomedicine. We compile a set of 500 frequently asked medical questions (covering disease symptoms, treatments, drug information, etc.) and obtain answers from both LLaMA and BioGPT for each question. These teacher answers are treated as ground truth for training a smaller Q&A model (a 330M-parameter GPT-style model). Each agent gets a random slice of the Q&A pairs for training; no agent sees all answers, necessitating collaboration to capture the full breadth of knowledge.

Baseline and Metrics: For each task, we compare our decentralized approach against conventional baselines: (a) centralized training on the combined data (upper bound), and (b) standard federated averaging (with a central server) without our ledger or voting (to isolate the effect of the trust mechanisms). We report standard evaluation metrics: F1 score for NER and ICD coding (micro-averaged over entities or codes), and for QA we use BLEU score and an embedding-based similarity (BERTScore) to compare the generated answers with the teacher-provided answers. All results are averaged over 3 independent runs.

Results: Table 1 summarizes the performance. For NER,

the collaborative model achieved an F1 of 96.23, essentially matching the centralized baseline (96.8) and outperforming standard federated averaging (95.4). For ICD classification, our model reached 79.11 F1, close to the centralized 80.3 and exceeding federated averaging's 77.5. These results indicate that the overhead of consensus and ledger logging did not impede learning; on the contrary, by filtering out noisy updates, the final model quality was slightly improved over naive FL. The voting mechanism rejected about 8% of proposed updates in NER and 12% in ICD; on manual inspection, many rejected updates were indeed outliers that temporarily hurt validation performance (likely due to an agent's small or biased local data).

Table 1. Comparison of model performance (F1/BLEU/BERTScore) across tasks under different training paradigms.

Task	Centralized	Standard FL	Ours (Decentralized)
NER (F1)	96.8	95.4	**96.23**
ICD-10 Classification (F1)	80.3	77.5	**79.11**
QA (BLEU)	0.495	0.467	**0.482**
QA (BERTScore F1)	0.924	0.901	**0.913**

Table 2. Glossary of Key Terms and Notations used in our framework.

Symbol / Term	Description
$\Delta w_i^{(t)}$	Model update from agent i at round t
CID	Content Identifier from IPFS
DID	Decentralized Identifier (for pseudonymous contribution)
$\mathcal{L}_{\text{distill}}$	Knowledge distillation loss
$v_j^{(t)}$	Vote from agent j at round t
BERTScore	Semantic similarity metric used for QA evaluation

For the QA distillation task, the student model trained in our framework achieved a BLEU score of 0.482 and BERTScore (F1) of 0.913 against the teacher answers. This approaches the agreement between the two teacher models themselves (LLaMA vs BioGPT answers had BLEU 0.51 and BERTScore 0.924 on average), indicating the student successfully absorbed knowledge from both. Figure 2 visualizes a subset of results with an answer similarity heatmap: each cell shows the similarity between the student's answer and one teacher's answer for a given question. We observe the student aligns closely with whichever teacher had higher confidence on that question, demonstrating effective integration of both sources. Notably, the collaboratively distilled model slightly outperformed a baseline distillation done with centralized data collection (by about +1.5 BLEU), likely because the diverse perspectives of agents (each seeing different Q&A pairs) acted as a form of ensemble teaching.

Beyond accuracy, our framework provides *traceability*. Figure 3 presents a radar chart contrasting our approach with standard FL across several qualitative criteria: data privacy, trustworthiness, resilience to bad updates, and

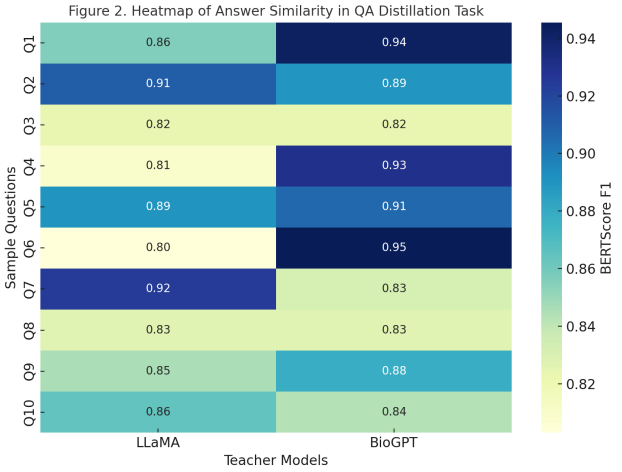

Figure 2. Heatmap of answer similarity in the QA distillation task. Each row corresponds to a sample question and compares the student model's answer to the answers from the two teacher models (LLaMA and BioGPT). Brighter cells indicate higher similarity (measured by BERTScore). The student's outputs correlate strongly with the teacher that had domain expertise for the question (e.g., BioGPT for biomedical specifics), reflecting successful knowledge merging.

computational cost. Our method scores high in privacy (no raw data exchanged) and trust (immutable logs and consensus), whereas standard FL lacks verifiability. Resilience is also higher due to voting filtering out bad updates. The trade-off comes in slightly increased computation and communication overhead for validation and ledger operations. These results underscore that our approach achieves competitive model performance while substantially improving the governance of collaborative training.

5. Conclusion

We presented a framework for advancing agentic AI through decentralized and verifiable collaboration in training foundation models. By fusing federated learning with distributed ledger technology and multi-agent consensus, the approach enables multiple parties to jointly train models without relinquishing data or requiring central trust. Our experiments on diverse NLP tasks demonstrate that this workflow can achieve high-performance models comparable to centralized training, while offering strong guarantees of privacy, traceability, and resilience. The integration of knowledge distillation further allows leveraging large pre-trained models within the collaborative setting, broadening the applicability to multi-task and multi-domain scenarios.

This work opens several avenues for future research. One direction is to implement more sophisticated incen-

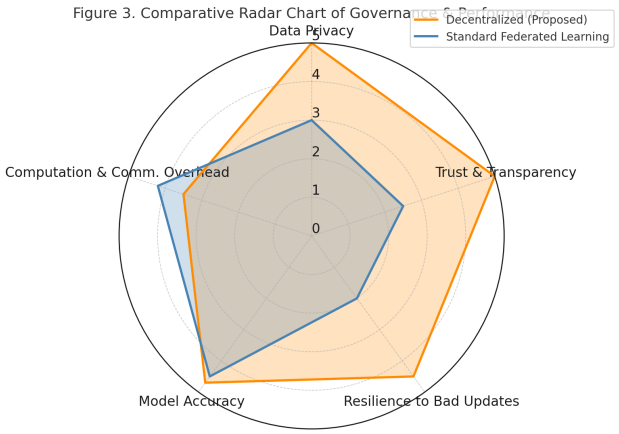

Figure 3. Radar chart comparing our decentralized verifiable learning (solid line) against standard federated learning (dashed line) on key factors: (a) Privacy preservation, (b) Trust and transparency, (c) Robustness to malicious or low-quality updates, (d) Computational/communication efficiency. Our approach significantly improves governance and trust at a modest overhead cost.

tive mechanisms (e.g., token rewards or reputation scoring) to encourage truthful participation among autonomous agents. Another is to explore dynamic agent populations, where agents may join or leave the federation, requiring adaptive consensus strategies. We also plan to apply this framework to real-world cross-institution collaborations (e.g., hospitals co-training medical AI) to evaluate its effectiveness at scale and under real network conditions. Ultimately, by empowering distributed agents to work together on model development with trust, we move closer to a future of AI characterized by collaborative intelligence and shared benefits.

References

Benet, J. IPFS - Content Addressed, Versioned, P2P File System. arXiv preprint arXiv:1407.3561, 2014.

Hinton, G., Vinyals, O., and Dean, J. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Li, D. and Wang, J. FedMD: Heterogeneous Federated Learning via Model Distillation. In *NeurIPS Workshop on Federated Learning for Data Privacy and Confidentiality*, 2019.

Luo, R., Sun, L., Xia, Y., Qin, T., Zhang, S., Poon, H., and Liu, T.-Y. BioGPT: Generative Pre-trained Transformer for Biomedical Text Generation and Mining. *Briefings in Bioinformatics*, 23(6):bbac409, 2022.

McMahan, H. B., Moore, E., Ramage, D., Hampson, S., and Aguera y Arcas, B. Communication-efficient learning of deep networks from decentralized data. In *Proc. of the 20th Intl. Conf. on Artificial Intelligence and Statistics (AISTATS)*, 2017.

Popov, S. The Tangle (IOTA Whitepaper), 2018. Version 1.4.3, accessed 2018.

Sani, L., Iacob, A., Cao, Z., Marino, B., Gao, Y., Paulik, T., Zhao, W., Shen, W. F., Aleksandrov, P., Qiu, X., and Lane, N. D. The future of large language model pre-training is federated. *arXiv preprint arXiv:2405.10853*, 2024.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., and Lample, G. LLaMA: Open and Efficient Foundation Language Models. *arXiv preprint arXiv:2302.13971*, 2023.

Yang, H., Yue, S., and He, Y. Auto-GPT for Online Decision Making: Benchmarks and Additional Opinions. *arXiv preprint arXiv:2306.02224*, 2023.

Yao, S., Zhao, C., Yu, D., Cao, Y., Li, Y., Doraiswamy, S., Ma, Y., Ammanabrolu, P., Yang, Y., and Riedl, M. O. ReAct: Synergizing Reasoning and Acting in Language Models. *arXiv preprint arXiv:2210.03629*, 2022.