agent framework for abductive, deductive, and inductive reasoning in language models. arXiv
tion, arXiv preprint arXiv:2502.02464, 2025. URL https://arxiv.org/abs/2502.02464v3.
long sequences, arXiv preprint arXiv:2411.17116, 2024. URL https://arxiv.org/abs/2411.

conversations and tool use? coalm: A unified conversational agentic language model, arXiv preprint arXiv:2502.08820, 2025. URL https://arxiv.org/abs/2502.08820v3.
futuredirections,arXivpreprintarXiv:2504.16939,2025. URLhttps://arxiv.org/abs/2504.
synthetic corpus generation for knowledge-enhanced language model pre-training. arXiv preprint,
restful proxy for model context protocol servers, arXiv preprint arXiv:2504.08999, 2025. URL
[17] Adel Al-Jumaily. Multi-agent system concepts theory and application phases. arXiv preprint, 2006.
alibi extrapolation. arXiv preprint, 2023.
developing generative ai apps: An empirical study, arXiv preprint arXiv:2506.16453, 2025. URL

arXiv preprint arXiv:2410.01690v1, 2024. URL https://arxiv.org/abs/2410.01690v1.
[25] Xavier Amatriain. Prompt design and engineering: Introduction and advanced methods, arXiv preprint arXiv:2401.14423, 2024. URL https://arxiv.org/abs/2401.14423v4.
Designing distributed agents in a worldwide network, arXiv preprint arXiv:2410.22339, 2024. URL
answering with retrieval augmented generation. arXiv preprint arXiv:2406.13372, 2024.

Arigraph: Learning knowledge graph world models with episodic memory for llm agents, arXiv preprint arXiv:2407.04363, 2024. URL https://arxiv.org/abs/2407.04363v3.
of building effective llm-based multi agent systems, arXiv preprint arXiv:2504.01963, 2025. URL
beled learning, arXiv preprint arXiv:2502.12565, 2025. URL https://arxiv.org/abs/2502.
shared prefixes in llms, arXiv preprint arXiv:2403.08845, 2024. URL https://arxiv.org/abs/
adaptive cognitive-inspired sketching. arXiv preprint, 2025.
Foundational models in medical imaging: A comprehensive survey and future vision, arXiv preprint arXiv:2310.18689, 2023. URL https://arxiv.org/abs/2310.18689v1.

perspective, arXiv preprint arXiv:2405.16640v2, 2024. URL https://arxiv.org/abs/2405.
arXiv preprint arXiv:2212.08073, 2022. URL https://arxiv.org/abs/2212.08073.
arXiv preprint arXiv:2404.04442, 2024. URL https://arxiv.org/abs/2404.04442v1.
for evaluating llms on nested sequences of api calls, arXiv preprint arXiv:2409.03797, 2024. URL

attributedgraphswithbigtex,arXivpreprintarXiv:2504.12474,2025. URLhttps://arxiv.org/
Overflow prevention enhances long-context recurrent llms. arXiv preprint, 2025.
[64] M. Benna and Stefano Fusi. Computational principles of biological memory, arXiv preprint arXiv:1507.07580, 2015. URL https://arxiv.org/abs/1507.07580v1.
ing, arXiv preprint arXiv:2505.24726, 2025. URL https://arxiv.org/abs/2505.24726v1.
orchestratemultipleagents?,arXivpreprintarXiv:2503.13577,2025. URLhttps://arxiv.org/

arXiv preprint arXiv:2506.06287, 2025. URL https://arxiv.org/abs/2506.06287v1.
[88] Vicent Botti. Agentic ai and multiagentic: Are we reinventing the wheel?, arXiv preprint arXiv:2506.01463, 2025. URL https://arxiv.org/abs/2506.01463v1.
survey of methods and datasets, arXiv preprint arXiv:2504.20119, 2025. URL https://arxiv.
traffic structure. arXiv preprint, 2017.

Amodei. Language models are few-shot learners, arXiv preprint arXiv:2005.14165, 2020. URL
retrieval augmented generation, arXiv preprint arXiv:2503.02922, 2025. URL https://arxiv.

arXiv preprint arXiv:2311.07491, 2023. URL https://arxiv.org/abs/2311.07491v1.
arXiv preprint arXiv:2505.19683, 2025. URL https://arxiv.org/abs/2505.19683v1.
graph-based retrieval-augmented generation for design space exploration. arXiv preprint, 2024.
multi-turn agentic planning, arXiv preprint arXiv:2505.16986, 2025. URL https://arxiv.org/

guarantees for multi-agent llm planning, arXiv preprint arXiv:2503.11951, 2025. URL https:
for generalizing to longer contexts, arXiv preprint arXiv:2502.14280, 2025. URL https://arxiv.
arXiv preprint arXiv:2411.11531, 2024. URL https://arxiv.org/abs/2411.11531v2.
acceleration, arXiv preprint arXiv:2410.10165, 2024. URL https://arxiv.org/abs/2410.
Yang. Pathrag: Pruning graph-based retrieval augmented generation with relational paths, arXiv preprint arXiv:2502.14902, 2025. URL https://arxiv.org/abs/2502.14902v1.
attention mechanisms on diverse hardware platforms, arXiv preprint arXiv:2502.15349, 2025. URL
generative tasks, arXiv preprint arXiv:2504.17261v1, 2025. URL https://arxiv.org/abs/

compound al systems, arXiv preprint arXiv:2506.04565, 2025. URL https://arxiv.org/abs/
Edgeinfinite: A memory-efficient infinite-context transformer for edge devices, arXiv preprint arXiv:2503.22196, 2025. URL https://arxiv.org/abs/2503.22196v1.
core: Multi-agent,iterative,coarse-to-finerefinementforreasoning,arXivpreprintarXiv:2409.12147,
Judgelrm: Large reasoning models as a judge, arXiv preprint arXiv:2504.00050, 2025. URL https:
of thought: A reasoning boundary framework to quantify and optimize chain-of-thought, arXiv preprint arXiv:2410.05695, 2024. URL https://arxiv.org/abs/2410.05695.
of-thought for reasoning large language models, arXiv preprint arXiv:2503.09567, 2025. URL
thought for reasoning large language models, arXiv preprint arXiv:2503.09567, 2025. URL https:
learning and chain-of-thought in large language model, arXiv preprint arXiv:2502.03325, 2025.
Wanxiang Che. Ai4research: A survey of artificial intelligence for scientific research, arXiv preprint arXiv:2507.01903, 2025. URL https://arxiv.org/abs/2507.01903.

of large language models via positional interpolation, arXiv preprint arXiv:2306.15595, 2023. URL
smalllanguagemodels,arXivpreprintarXiv:2505.07460,2025. URLhttps://arxiv.org/abs/
by step. arXiv preprint arXiv:2312.14033, 2023.
Mindsearch: Mimicking human minds elicits deep ai searcher, arXiv preprint arXiv:2407.20183,
multi-hop instruction datasets? insights and best practices, arXiv preprint arXiv:2409.01893, 2025.
JundongLi. Asurveyofscalinginlargelanguagemodelreasoning,arXivpreprintarXiv:2504.02181,

ation, arXiv preprint arXiv:2503.10677, 2025. URL https://arxiv.org/abs/2503.10677v2.
with knowledge graph for complex problem solving, arXiv preprint arXiv:2503.06567, 2025. URL
intelligentagents: Definitions,methods,andprospects,arXivpreprintarXiv:2401.03428,2024. URL
benchmark for solving complex tasks with reinforcement learning, arXiv preprint arXiv:2502.10550,
production-ready ai agents with scalable long-term memory, arXiv preprint arXiv:2504.19413, 2025.
for pharmacovigilance, arXiv preprint arXiv:2408.01869, 2024. URL https://arxiv.org/abs/
arXiv preprint arXiv:2503.11733, 2025. URL https://arxiv.org/abs/2503.11733v1.

Cui, Longfei Li, Junqing Zhou, and Sheng Li. Data-centric financial large language models, arXiv preprint arXiv:2310.17784, 2023. URL https://arxiv.org/abs/2310.17784v2.
llm-basedagents,arXivpreprintarXiv:2311.09618,2024. URLhttps://arxiv.org/abs/2311.
attribution in large language models, arXiv preprint arXiv:2502.09604, 2025. URL https://
and reproducible evaluation sandbox for deep research, arXiv preprint arXiv:2505.19253, 2025.
[184] Erica Coppolillo. Injecting knowledge graphs into large language models, arXiv preprint arXiv:2505.07554, 2025. URL https://arxiv.org/abs/2505.07554v1.
Scaling reasoning without scaling models, arXiv preprint arXiv:2504.18116, 2025. URL https:

arXiv preprint arXiv:2209.03859, 2022. URL https://arxiv.org/abs/2209.03859v1.
hancing tool learning in large language models with hierarchical error checklists, arXiv preprint arXiv:2506.00042, 2025. URL https://arxiv.org/abs/2506.00042v1.
graph pattern comprehension, arXiv preprint arXiv:2410.05298v2, 2024. URL https://arxiv.
[197] Fatemeh Daneshfar and H. Bevrani. Multi-agent systems in control engineering: a survey. arXiv
collaboration via evolving orchestration, arXiv preprint arXiv:2505.19591, 2025. URL https:
Larimar: Large language models with episodic memory control, arXiv preprint arXiv:2403.11901,
[203] Adrian de Wynter, Xun Wang, Qilong Gu, and Si-Qing Chen. On meta-prompting, arXiv preprint arXiv:2312.06562, 2023. URL https://arxiv.org/abs/2312.06562v3.

Asano. Self-supervised open-ended classification with small visual language models, arXiv preprint arXiv:2310.00500, 2023. URL https://arxiv.org/abs/2310.00500v2.
Qian. Trail: Trace reasoning and agentic issue localization, arXiv preprint arXiv:2505.08638, 2025.
systems: A scoping survey, arXiv preprint arXiv:2312.17601, 2023. URL https://arxiv.org/
arXiv preprint arXiv:2502.02046, 2025. URL https://arxiv.org/abs/2502.02046v2.
languagemodels,arXivpreprintarXiv:2502.11404,2025. URLhttps://arxiv.org/abs/2502.

Longnet: Scaling transformers to 1, 000, 000, 000 tokens. arXiv preprint, 2023.
An algorithmic survey, arXiv preprint arXiv:2312.00678, 2023. URL https://arxiv.org/abs/
reinforcement learning. arXiv preprint, 2025.
llms truly act? an empirical evaluation of agentic capabilities in llm compression, arXiv preprint arXiv:2505.19433, 2025. URL https://arxiv.org/abs/2505.19433v2.
support document, arXiv preprint arXiv:2410.23452, 2024. URL https://arxiv.org/abs/
arXiv preprint arXiv:2412.14352, 2024. URL https://arxiv.org/abs/2412.14352v1.

retrieval-augmented language models training efficiency, arXiv preprint arXiv:2505.14309, 2025.
allocationovermulti-agentsystems,arXivpreprintarXiv:2401.15607,2024. URLhttps://arxiv.
systems: Techniques, challenges and future directions, arXiv preprint arXiv:2402.01968, 2024. URL
mixture-of-memories, arXiv preprint arXiv:2502.13685, 2025. URL https://arxiv.org/abs/
comprehensive benchmark for deep research agents, arXiv preprint arXiv:2506.11763, 2025. URL
survey on the optimization of large language model-based agents, arXiv preprint arXiv:2503.12434,
retrieval,andsynthesisinquestionanswering,arXivpreprintarXiv:2402.16288,2024. URLhttps:

for multi-gnns?, arXiv preprint arXiv:2410.16822, 2024. URL https://arxiv.org/abs/2410.
refinement for design critique generation, arXiv preprint arXiv:2412.16829, 2024. URL https:
cognitive tools, arXiv preprint arXiv:2506.12115, 2025. URL https://arxiv.org/abs/2506.
rag approach to query-focused summarization. arXiv preprint arXiv:2404.16130, 2024.
knowledge graphs and vector database for accreditation reporting assistance, arXiv preprint arXiv:2405.15436, 2024. URL https://arxiv.org/abs/2405.15436v1.
protocol (a2a), and agent network protocol (anp), arXiv preprint arXiv:2505.02279, 2025. URL
prompting techniques for large language models: A practitioner's guide. arXiv preprint, 2024.
ramanian, Parsa Hosseini, and S. Feizi. Gaming tool preferences in agentic llms, arXiv preprint arXiv:2505.18135, 2025. URL https://arxiv.org/abs/2505.18135v1.
llms, arXiv preprint arXiv:2503.23514, 2025. URL https://arxiv.org/abs/2503.23514v1.

augmentedmultimodalagentforvideounderstanding,arXivpreprintarXiv:2403.11481,2024. URL
self-supervised learning for language understanding, arXiv preprint arXiv:2005.12766, 2020. URL
should identify and mitigate third-party safety risks in mcp-powered agent systems, arXiv preprint arXiv:2506.13666, 2025. URL https://arxiv.org/abs/2506.13666v1.
reasoning path aggregation. arXiv preprint, 2024.
instruction optimization for llm agents via tool play, arXiv preprint arXiv:2503.14432, 2025. URL
multimodal comprehension and generation on graphs. arXiv preprint, 2025.
Towardsconversationalaiforhuman-machinecollaborativemlops,arXivpreprintarXiv:2504.12477,
from scratch. arXiv preprint, 2025.
retrieval-augmented chatbots, arXiv preprint arXiv:2403.01193, 2024. URL https://arxiv.

replay,arXivpreprintarXiv:2505.17716,2025. URLhttps://arxiv.org/abs/2505.17716v1.
Jinxin Chi, and Wanjun Zhong. Retool: Reinforcement learning for strategic tool use in llms, arXiv preprint arXiv:2504.11536, 2025. URL https://arxiv.org/abs/2504.11536v2.
attention block, arXiv preprint arXiv:2306.12599, 2023. URL https://arxiv.org/abs/2306.
foundation model, arXiv preprint arXiv:2503.01203v1, 2025. URL https://arxiv.org/abs/
comprehensive review, arXiv preprint arXiv:2504.19678, 2025. URL https://arxiv.org/abs/
lems for llms, arXiv preprint arXiv:2503.22732, 2025. URL https://arxiv.org/abs/2503.

arXiv preprint arXiv:2407.09450, 2024. URL https://arxiv.org/abs/2407.09450.
comprehensive evaluation benchmark of multi-modal llms in video analysis. arXiv preprint, 2024.
evaluation of llm agents in real-world environments. arXiv preprint, 2025.
languagemodels,arXivpreprintarXiv:2502.18845,2025. URLhttps://arxiv.org/abs/2502.
[300] Stefano Fusi. Memory capacity of neural network models, arXiv preprint arXiv:2108.07839, 2021.

augmented generation. arXiv preprint, 2025.
edge, arXiv preprint arXiv:2505.05177, 2025. URL https://arxiv.org/abs/2505.05177v1.
arXiv preprint arXiv:2307.14984, 2025. URL https://arxiv.org/abs/2307.14984.
[306] Hang Gao and Yongfeng Zhang. Memory sharing for large language model based agents, arXiv preprint arXiv:2404.09982, 2024. URL https://arxiv.org/abs/2404.09982v2.
tion, arXiv preprint arXiv:2403.17698, 2024. URL https://arxiv.org/abs/2403.17698v1.
knowledge graphs and llms for ai research idea generation, arXiv preprint arXiv:2503.08549, 2025.
benchmark for evaluating tool use capabilities in large language models. arXiv preprint, 2025.
arXiv preprint arXiv:2312.10997, 2023. URL https://arxiv.org/abs/2312.10997v5.

into lego-like reconfigurable frameworks, arXiv preprint arXiv:2407.21059, 2024. URL https:
reasoning: A systematic review, arXiv preprint arXiv:2504.15909, 2025. URL https://arxiv.
large language model capabilities, arXiv preprint arXiv:2310.01441, 2023. URL https://arxiv.

arXiv preprint arXiv:2506.08074, 2025. URL https://arxiv.org/abs/2506.08074v1.
[335] D. Ghica. Function interface models for hardware compilation: Types, signatures, protocols, arXiv preprint arXiv:0907.0749, 2009. URL https://arxiv.org/abs/0907.0749v1.
[337] In Gim, Seung seob Lee, and Lin Zhong. Asynchronous llm function calling, arXiv preprint arXiv:2412.07017, 2024. URL https://arxiv.org/abs/2412.07017v1.
dialogue agents via targeted human judgements, arXiv preprint arXiv:2209.14375, 2022. URL
underwater. arXiv preprint, 1975.
analysis through prompt engineering for llms, arXiv preprint arXiv:2409.14879, 2024. URL https:
[343] E. Gordon and B. Logan. Managing goals and resources in dynamic environments. arXiv preprint,

structured matrices and orthogonal transformations, arXiv preprint arXiv:2506.02818, 2025. URL
[349] C. Gros. Complex and adaptive dynamical systems, arXiv preprint arXiv:0807.4838, 2008. URL
improving in-context learning, arXiv preprint arXiv:2401.16184, 2024. URL https://arxiv.
inference by steering attention on reused contexts, arXiv preprint arXiv:2411.13009, 2024. URL
llm-based agents for multi-turn conversations: A survey, arXiv preprint arXiv:2503.22458, 2025.
descriptions of graphs with tokenized topological modeling, arXiv preprint arXiv:2406.13250, 2024.

arXiv preprint arXiv:2407.21075, 2024. URL https://arxiv.org/abs/2407.21075v1.
structured data ? an empirical evaluation and benchmarking, arXiv preprint arXiv:2305.15066,
Empowering working memory for large language model agents, arXiv preprint arXiv:2312.17259,
graphs with reinforcement learning. arXiv preprint arXiv:2505.18499, 2025.
arXiv preprint arXiv:2506.08972, 2025. URL https://arxiv.org/abs/2506.08972v1.
augmented generation, arXiv preprint arXiv:2410.05779, 2024. URL https://arxiv.org/abs/

codebases with llm agents, arXiv preprint arXiv:2406.12276, 2024. URL https://arxiv.org/
Aloe: A family of fine-tuned open healthcare llms, arXiv preprint arXiv:2405.01886, 2024. URL
and related systems. arXiv preprint, 1966.
ai application leveraging a2a protocol, arXiv preprint arXiv:2504.16902, 2025. URL https://
preference alignment, arXiv preprint arXiv:2505.23634, 2025. URL https://arxiv.org/abs/
attention recalibration, arXiv preprint arXiv:2504.09402, 2025. URL https://arxiv.org/abs/

graphs (graphrag), arXiv preprint arXiv:2501.00309, 2025. URL https://arxiv.org/abs/
budget-aware llm reasoning, arXiv preprint arXiv:2412.18547, 2024. URL https://arxiv.org/
Cowan. Evaluating the sensitivity of llms to prior context, arXiv preprint arXiv:2506.00069, 2025.
understanding, arXiv preprint arXiv:2504.00409, 2025. URL https://arxiv.org/abs/2504.
tive agents, arXiv preprint arXiv:2409.18538, 2024. URL https://arxiv.org/abs/2409.
agents: Theoretical foundations, architectural components, and cognitive integration, arXiv preprint arXiv:2504.06943, 2025. URL https://arxiv.org/abs/2504.06943v2.
dynamic retrieval for improving generation quality in rag models, arXiv preprint arXiv:2504.19436,

out fine-tuning, arXiv preprint arXiv:2407.04997, 2024. URL https://arxiv.org/abs/2407.
context fine-tuning for large language model, arXiv preprint arXiv:2506.11103, 2025. URL https:
agent systems, arXiv preprint arXiv:2505.24201, 2025. URL https://arxiv.org/abs/2505.
route: Automatic mode switching via capability estimation for efficient reasoning. arXiv preprint,
abilities of llms via data structures. arXiv preprint, 2025.
Camelot: Towards large language models with training-free consolidated associative memory, arXiv preprint arXiv:2402.13449, 2024. URL https://arxiv.org/abs/2402.13449v1.
of shared and separate context approaches, arXiv preprint arXiv:2504.07303, 2025. URL https:

[406] Thomas Hoang. Gnn: Graph neural network and large language model for data discovery, arXiv preprint arXiv:2408.13609, 2024. URL https://arxiv.org/abs/2408.13609v2.
experience replay for lifelong language learning, arXiv preprint arXiv:2009.04891, 2020. URL
programming for multi-agent collaborative framework. arXiv preprint, 2023.
arXiv preprint arXiv:2308.00352, 2024. URL https://arxiv.org/abs/2308.00352.
rization with context-aware fine-grained graph rag. arXiv preprint, 2025.

models via instruction-aware contextual compression, arXiv preprint arXiv:2408.15491, 2024. URL
multimodalllmsforsurgicalvqaviaself-containedinquiry,arXivpreprintarXiv:2411.10937v1,2024.
security threats, and future research directions, arXiv preprint arXiv:2503.23278, 2025. URL
[426] Marc W Howard and M. Kahana. A distributed representation of temporal context. arXiv preprint,
llmswithdatabasesastheirsymbolicmemory,arXivpreprintarXiv:2306.03901,2023. URLhttps:
arXiv preprint arXiv:2502.11147, 2025. URL https://arxiv.org/abs/2502.11147v2.
nano surge approach for code reasoning efficiency, arXiv preprint arXiv:2504.15989, 2025. URL
model, arXiv preprint arXiv:2408.09559, 2024. URL https://arxiv.org/abs/2408.09559.

models, arXiv preprint arXiv:2310.08582, 2024. URL https://arxiv.org/abs/2310.08582.
task automation, arXiv preprint arXiv:2505.23885, 2025. URL https://arxiv.org/abs/2505.
case study with claude 3.5 computer use. arXiv preprint, 2024.
guage generation, arXiv preprint arXiv:2309.06759, 2023. URL https://arxiv.org/abs/
diversity of llm generated ideas, arXiv preprint arXiv:2410.14255, 2024. URL https://arxiv.
models' ability on understanding graph data, arXiv preprint arXiv:2310.04944, 2023. URL https:
in multimodal llms, arXiv preprint arXiv:2408.01417v1, 2024. URL https://arxiv.org/abs/
From feature-based, generative to agentic paradigms, arXiv preprint arXiv:2504.16420, 2025. URL
in the era of multimodal large language models, arXiv preprint arXiv:2503.16734, 2025. URL

level tool-use preference alignment training framework with fine-grained evaluation, arXiv preprint arXiv:2505.20016, 2025. URL https://arxiv.org/abs/2505.20016v1.
arXiv preprint arXiv:2406.06110, 2024. URL https://arxiv.org/abs/2406.06110v1.
binding protocol (acnbp), arXiv preprint arXiv:2506.13590, 2025. URL https://arxiv.org/
openmulti-agentframework,arXivpreprintarXiv:2505.18105,2025. URLhttps://arxiv.org/
error scenarios, arXiv preprint arXiv:2506.13977, 2025. URL https://arxiv.org/abs/2506.
knowledge in structured data with large language models, arXiv preprint arXiv:2408.12188, 2024.
enhanced llms for structured knowledge, arXiv preprint arXiv:2502.18125, 2025. URL https:
Integratinglargelanguagemodelsforinteractiverecommendations,arXivpreprintarXiv:2308.16505,
RuimingTang,andEnhongChen. Understandingtheplanningofllmagents: Asurvey,arXivpreprint
arXiv:2402.02716, 2024. URL https://arxiv.org/abs/2402.02716v1.

language models: A comprehensive survey, arXiv preprint arXiv:2311.12351, 2023. URL https:
Visualtoolagent (vista): A reinforcement learning framework for visual tool selection, arXiv preprint arXiv:2505.20289, 2025. URL https://arxiv.org/abs/2505.20289v1.
box: Transparent and interactive memory management for conversational agents, arXiv preprint arXiv:2308.01542, 2023. URL https://arxiv.org/abs/2308.01542.
arXiv preprint arXiv:2312.10256, 2023. URL https://arxiv.org/abs/2312.10256v2.
tonomous llm-powered multi-agent architectures. arXiv preprint, 2023.
rag: A systems approach to question answering, arXiv preprint arXiv:2412.06832, 2024. URL
and G. Karypis. Efficient and effective training of language and graph neural network models, arXiv preprint arXiv:2206.10781, 2022. URL https://arxiv.org/abs/2206.10781v1.
ing, arXiv preprint arXiv:2410.15639, 2024. URL https://arxiv.org/abs/2410.15639v5.

cal ventilation, arXiv preprint arXiv:2505.04645, 2025. URL https://arxiv.org/abs/2505.
integration of large language models for message passing in graph neural networks, arXiv preprint arXiv:2407.14996, 2024. URL https://arxiv.org/abs/2407.14996v1.
[483] R. Janik. Aspects of human memory and large language models, arXiv preprint arXiv:2311.03839,
A.Ozcan. Learningtoremember,forgetandignoreusingattentioncontrolinmemory,arXivpreprint arXiv:1809.11087, 2018. URL https://arxiv.org/abs/1809.11087v1.
autonomous agents, arXiv preprint arXiv:2506.01804, 2025. URL https://arxiv.org/abs/

autonomously learn without external supervision, arXiv preprint arXiv:2406.00606, 2024. URL
scale universal user representation with sparse mixture of experts, arXiv preprint arXiv:2207.04648,
in-context learning capabilities of large language models, arXiv preprint arXiv:2503.22401, 2025.
Narasimhan. Swe-bench: Can language models resolve real-world github issues?, arXiv preprint
arXiv:2310.06770, 2024. URL https://arxiv.org/abs/2310.06770.

based agents for software engineering: A survey of current, challenges and future, arXiv preprint arXiv:2408.02479, 2024. URL https://arxiv.org/abs/2408.02479v2.
large language model inference, arXiv preprint arXiv:2404.09336, 2024. URL https://arxiv.
perspectives. arXiv preprint, 2025.
Self-evolving code agents via iterative generation-verification, arXiv preprint arXiv:2506.11442,

arXiv preprint arXiv:2310.02172, 2023. URL https://arxiv.org/abs/2310.02172v1.
based conversational agents, arXiv preprint arXiv:2504.02891, 2025. URL https://arxiv.org/
arXiv preprint arXiv:2504.08148, 2025. URL https://arxiv.org/abs/2504.08148v1.
generation via streaming algorithm and k-means clustering. arXiv preprint, 2024.
[521] Jiazheng Kang, Mingming Ji, Zhe Zhao, and Ting Bai. Memory os of ai agent, arXiv preprint arXiv:2506.06326, 2025. URL https://arxiv.org/abs/2506.06326v1.
Purtorab, and Andy Toulis. Lm2: Large memory models, arXiv preprint arXiv:2502.06049, 2025.
models for function calling, arXiv preprint arXiv:2505.10570, 2025. URL https://arxiv.org/

autonomously learn without external supervision, arXiv preprint arXiv:2406.00606, 2024. URL
scale universal user representation with sparse mixture of experts, arXiv preprint arXiv:2207.04648,
in-context learning capabilities of large language models, arXiv preprint arXiv:2503.22401, 2025.
Narasimhan. Swe-bench: Can language models resolve real-world github issues?, arXiv preprint
arXiv:2310.06770, 2024. URL https://arxiv.org/abs/2310.06770.

mystifying ai agents and test-time scaling from an ai infrastructure perspective, arXiv preprint arXiv:2506.04301, 2025. URL https://arxiv.org/abs/2506.04301v1.
in-context learning by meta-learning transformers, arXiv preprint arXiv:2212.04458, 2022. URL
llm for ultra-long context horizons, arXiv preprint arXiv:2506.01963, 2025. URL https://arxiv.
telemetry-aware in-ide ai application development using the model context protocol (mcp), arXiv preprint arXiv:2506.11019, 2025. URL https://arxiv.org/abs/2506.11019v1.

multimodal agents on realistic visual web tasks, arXiv preprint arXiv:2401.13649, 2024. URL
usageoflargelanguagemodel-basedagentsinreal-worldsystems,arXivpreprintarXiv:2311.11315,
[562] Oliver Kramer. Cognitive prompts using guilford’s structure of intellect model. arXiv preprint, 2025.
arXiv preprint arXiv:2502.01901, 2025. URL https://arxiv.org/abs/2502.01901v1.
for transparency and dedicated metrics for energy consumption, arXiv preprint arXiv:2502.17903,
Hisham Cholakkal. Multi-modal generation via cross-modal in-context learning, arXiv preprint
arXiv:2405.18304v1, 2024. URL https://arxiv.org/abs/2405.18304v1.

utilization for personalized assistance, arXiv preprint arXiv:2505.16348, 2025. URL https://
language models for local life services, arXiv preprint arXiv:2506.02720, 2025. URL https:
logue agents, arXiv preprint arXiv:2409.04617, 2024. URL https://arxiv.org/abs/2409.
[580] Pak Kin Lau and Stuart Michael McManus. Mining asymmetric intertextuality, arXiv preprint arXiv:2410.15145, 2024. URL https://arxiv.org/abs/2410.15145v1.
example generation for machine translation, arXiv preprint arXiv:2506.00507, 2025. URL https:
long-term memory, arXiv preprint arXiv:2409.11192, 2024. URL https://arxiv.org/abs/

prunedattention,arXivpreprintarXiv:2406.09827,2024. URLhttps://arxiv.org/abs/2406.
optimization for contextual understanding, arXiv preprint arXiv:2506.01274v1, 2025. URL https:
enhanced collaborative llm agents for drug discovery, arXiv preprint arXiv:2502.17506, 2025. URL
arXiv preprint, 2024.
Towards improving performance of large language models on structured data, arXiv preprint arXiv:2407.02750, 2024. URL https://arxiv.org/abs/2407.02750v1.
Optimal representations of structured data in prompting large language models, arXiv preprint arXiv:2402.14195, 2024. URL https://arxiv.org/abs/2402.14195v1.
correction via linear representations and latent concepts, arXiv preprint arXiv:2505.11924, 2025.
conversations, arXiv preprint arXiv:2505.23121v1, 2025. URL https://arxiv.org/abs/2505.

arXiv preprint arXiv:2506.09820, 2025. URL https://arxiv.org/abs/2506.09820v2.
arXiv preprint arXiv:2405.16376, 2024. URL https://arxiv.org/abs/2405.16376v2.
Yu, and Sanjiv Kumar. Large language models with controllable working memory, arXiv preprint arXiv:2211.05110, 2022. URL https://arxiv.org/abs/2211.05110.
[604] Daniel Li and Lincoln Murr. Humaneval on latest gpt models - 2024. arXiv preprint, 2024.
developing large language models for behavior tree generation, arXiv preprint arXiv:2401.08089,
lucinations: A case study on domain-specific queries in private knowledge-bases, arXiv preprint arXiv:2403.10446, 2024. URL https://arxiv.org/abs/2403.10446v1.

sentence-level self-evolution, arXiv preprint arXiv:2503.01695, 2025. URL https://arxiv.org/
Yu Qiao. Videochat: Chat-centric video understanding, arXiv preprint arXiv:2305.06355v2, 2023.
intelligence, arXiv preprint arXiv:2506.10157, 2025. URL https://arxiv.org/abs/2506.
current surveys, arXiv preprint arXiv:2409.18991, 2024. URL https://arxiv.org/abs/2409.
for scalable agent systems, arXiv preprint arXiv:2505.03864, 2025. URL https://arxiv.org/
editing via memory-augmented modality, arXiv preprint arXiv:2503.02701v1, 2025. URL https:

for efficient model editing, arXiv preprint arXiv:2505.22156, 2025. URL https://arxiv.org/
models can self-improve in long-context reasoning, arXiv preprint arXiv:2411.08147, 2024. URL
A survey of personalization: From rag to agent, arXiv preprint arXiv:2504.10147, 2025. URL
arXiv preprint arXiv:2504.21776, 2025. URL https://arxiv.org/abs/2504.21776v1.
collaboration, arXiv preprint arXiv:2410.18032v4, 2024. URL https://arxiv.org/abs/2410.
iors for llm-based task-oriented coordination via collaborative generative agents, arXiv preprint
arXiv:2310.06500, 2023. URL https://arxiv.org/abs/2310.06500.
self-information-based content filtering, arXiv preprint arXiv:2304.12102, 2023. URL https:

anisotropy in graph neural networks with language semantics, arXiv preprint arXiv:2504.01429,
operatingsystemformemory-augmentedgeneration(mag)inlargelanguagemodels,arXivpreprint
arXiv:2505.22101, 2025. URL https://arxiv.org/abs/2505.22101v1.
llm,arXivpreprintarXiv:2505.18110v2,2025. URLhttps://arxiv.org/abs/2505.18110v2.
uncertainty enhanced long-context modeling for retrieval-augmented generation, arXiv preprint
arXiv:2410.02719, 2024. URL https://arxiv.org/abs/2410.02719v1.

arXiv preprint arXiv:2505.16120, 2025. URL https://arxiv.org/abs/2505.16120v1.
arXiv preprint arXiv:2506.10408, 2025. URL https://arxiv.org/abs/2506.10408v1.
Li. Scm: Enhancing large language model with self-controlled memory framework, arXiv preprint arXiv:2304.13343, 2023. URL https://arxiv.org/abs/2304.13343v4.
agents with reflective and memory-augmented abilities, arXiv preprint arXiv:2409.00872, 2024.
self-improvement,arXivpreprintarXiv:2503.19271,2025.URLhttps://arxiv.org/abs/2503.
llm inference via algorithm-hardware co-design, arXiv preprint arXiv:2505.03745, 2025. URL
scratch through an iterative self-enhancement paradigm. arXiv preprint, 2024.
approach for t5 using knowledge graphs to address complex tasks, arXiv preprint arXiv:2502.16484,
tems, arXiv preprint arXiv:1711.02634, 2017. URL https://arxiv.org/abs/1711.02634v1.

Efficient llm service for long context with distattention and distributed kvcache, arXiv preprint arXiv:2401.02669, 2024. URL https://arxiv.org/abs/2401.02669.
arXiv preprint arXiv:2505.17648, 2025. URL https://arxiv.org/abs/2505.17648v2.
arXiv preprint arXiv:2410.16392, 2024. URL https://arxiv.org/abs/2410.16392v2.
device language models via function masking, arXiv preprint arXiv:2410.04587, 2024. URL https:
Yun-Nung Chen. Llm inference enhanced by external knowledge: A survey, arXiv preprint arXiv:2505.24377, 2025. URL https://arxiv.org/abs/2505.24377v1.

On the intrinsic self-correction capability of llms: Uncertainty and latent concept, arXiv preprint arXiv:2406.02378, 2024. URL https://arxiv.org/abs/2406.02378v2.
framework for knowledge graph question answering, arXiv preprint arXiv:2406.01145, 2024. URL arXiv preprint arXiv:2504.19838, 2025. URL https://arxiv.org/abs/2504.19838v2.
Acps: Agent collaboration protocols for the internet of agents, arXiv preprint arXiv:2505.13523,
ing Lou. Large language model-based agents for software engineering: A survey, arXiv preprint arXiv:2409.02977, 2024. URL https://arxiv.org/abs/2409.02977v1.
Think-in-memory: Recalling and post-thinking enable llms with long-term memory. arXiv preprint,
Think-in-memory: Recalling and post-thinking enable llms with long-term memory, arXiv preprint arXiv:2311.08719, 2023. URL https://arxiv.org/abs/2311.08719.
agent: A memory enhanced architecture with fine-tuning of large language models, arXiv preprint arXiv:2401.02777, 2024. URL https://arxiv.org/abs/2401.02777v2.

chical multi-agent multimodal retrieval augmented generation. arXiv preprint, 2025.
arXiv preprint arXiv:2409.00920, 2024. URL https://arxiv.org/abs/2409.00920v1.
agent usability is agentic roi, arXiv preprint arXiv:2505.17767, 2025. URL https://arxiv.org/
Bo Jiang, Aimin Zhou, and Liang He. Mathematical language models: A survey, arXiv preprint
arXiv:2312.07622, 2023. URL https://arxiv.org/abs/2312.07622v4.
with temporal episodic memory, arXiv preprint arXiv:2502.16090, 2025. URL https://arxiv.
Huang, and Bryan Hooi. Efficient inference for large reasoning models: A survey, arXiv preprint
arXiv:2503.23077, 2025. URL https://arxiv.org/abs/2503.23077v2.

task-oriented agent collaboration, arXiv preprint arXiv:2310.02170, 2023. URL https://arxiv.
knowledge input beyond context windows of llms via multi-agent collaboration, arXiv preprint arXiv:2505.21471, 2025. URL https://arxiv.org/abs/2505.21471v1.
reasoning through the lens of neuroscience. arXiv preprint, 2025.
institutions, arXiv preprint arXiv:2503.13524, 2025. URL https://arxiv.org/abs/2503.
Wang, Lifeng Shang, and Qun Liu. Self: Self-evolution with language feedback, arXiv preprint arXiv:2310.00533, 2023. URL https://arxiv.org/abs/2310.00533v4.
Memochat: Tuning llms to use memos for consistent long-range open-domain conversation, arXiv preprint arXiv:2308.08239, 2023. URL https://arxiv.org/abs/2308.08239.
api-first llm-based agents, arXiv preprint arXiv:2409.17140, 2025. URL https://arxiv.org/
long-context llms through multi-objective partitioning, arXiv preprint arXiv:2409.00997, 2024. URL

sentations with message passing on hierarchical relational graphs, arXiv preprint arXiv:2109.04223,
Scalemcp: Dynamic and auto-synchronizing model context protocol tools for llm agents, arXiv preprint arXiv:2505.06416, 2025. URL https://arxiv.org/abs/2505.06416v1.
transformer: Optimizing intermediate memory for long sequences training, arXiv preprint arXiv:2407.15892, 2024. URL https://arxiv.org/abs/2407.15892v4.
short reasoning for efficient large language models, arXiv preprint arXiv:2505.22662, 2025. URL
Zhang. Enhance graph alignment for large language models, arXiv preprint arXiv:2410.11370v1,
generation with hypergraph-structured knowledge representation. arXiv preprint, 2025.
Dacheng Tao. O1-pruner: Length-harmonizing fine-tuning for o1-like reasoning pruning. arXiv
survey on methodology, applications and challenges, arXiv preprint arXiv:2503.21460, 2025. URL
for small language models, arXiv preprint arXiv:2506.07712, 2025. URL https://arxiv.org/

question answering systems that use tools, arXiv preprint arXiv:2505.16113, 2025. URL https:
A benchmark for deep research shopping agents, arXiv preprint arXiv:2506.02839, 2025. URL
framework for large language models, arXiv preprint arXiv:2409.03155, 2024. URL https://
language model based agents: A survey and perspective, arXiv preprint arXiv:2402.00262, 2024.
augmented large language models, arXiv preprint arXiv:2305.14283, 2023. URL https://arxiv.

scaling law: Agent rl with spontaneous code execution for mathematical problem solving, arXiv preprint arXiv:2505.07773, 2025. URL https://arxiv.org/abs/2505.07773v2.
tion learning with large language models: A comprehensive survey of techniques, arXiv preprint arXiv:2402.05952v1, 2024. URL https://arxiv.org/abs/2402.05952v1.
[749] Sophia Maria. Compass-v2 technical report, arXiv preprint arXiv:2504.15527, 2025. URL https:
between knowledge graphs and llms for complex reasoning, arXiv preprint arXiv:2505.24478, 2025.
architectures for reasoning, planning, and tool calling: A survey, arXiv preprint arXiv:2404.11584,
netlist-to-schematic conversion. arXiv preprint, 2024.

languagemodels,arXivpreprintarXiv:2506.09342,2025. URLhttps://arxiv.org/abs/2506.
operatingsystem,arXivpreprintarXiv:2403.16971,2024. URLhttps://arxiv.org/abs/2403.
Fine-grainedsafegenerationwithspecializedrepresentationrouter,arXivpreprintarXiv:2410.02684,
Cheng. a1: Steep test-time scaling law via environment augmented generation, arXiv preprint arXiv:2504.14597, 2025. URL https://arxiv.org/abs/2504.14597.
"malicious": Being careful about hallucinations of large language models' jailbreak, arXiv preprint arXiv:2406.11668, 2025. URL https://arxiv.org/abs/2406.11668.

reasoningwithlogic-enhancedlanguagemodelagents,arXivpreprintarXiv:2408.16081,2024. URL
[773] G. M. Mensink and J. Raaijmakers. A model for interference and forgetting. arXiv preprint, 1988.
computer systems, arXiv preprint arXiv:2504.04485, 2025. URL https://arxiv.org/abs/
Gaia: a benchmark for general ai assistants, arXiv preprint arXiv:2311.12983, 2023. URL https:
arXiv preprint arXiv:2312.15234, 2023. URL https://arxiv.org/abs/2312.15234v1.
[780] JacobMiller,GuillaumeRabusseau,andJohnTerilla. Tensornetworksforlanguagemodeling. arXiv
and domain expertise, arXiv preprint arXiv:2410.19811, 2024. URL https://arxiv.org/abs/

read-writememoryforlargelanguagemodels,arXivpreprintarXiv:2305.14322,2024. URLhttps:
[786] Behnam Mohammadi. Pel, a programming language for orchestrating ai agents, arXiv preprint arXiv:2505.13453, 2025. URL https://arxiv.org/abs/2505.13453v2.
[789] Dimitri Coelho Mollo and Raphael Milliere. The vector grounding problem, arXiv preprint arXiv:2304.01481, 2023. URL https://arxiv.org/abs/2304.01481v2.
A. Gholami. Efficient and scalable estimation of tool representations in vector space, arXiv preprint arXiv:2409.02141, 2024. URL https://arxiv.org/abs/2409.02141v1.
models with ensemble of critics. arXiv preprint, 2023.
to dialogue: Building kg-rag enhanced ai assistants, arXiv preprint arXiv:2502.15237, 2025. URL
training elicits concise reasoning in large language models, arXiv preprint arXiv:2502.20122, 2025.

infinite context transformers with infini-attention, arXiv preprint arXiv:2404.07143, 2024. URL
Browser-assisted question-answering with human feedback, arXiv preprint arXiv:2112.09332, 2022.
pixel-level semantic knowledge in diffusion models, arXiv preprint arXiv:2401.11739, 2024. URL
ment engine for the downstream uses of llms. arXiv preprint, 2024.
objective directional prompting, arXiv preprint arXiv:2504.18722, 2025. URL https://arxiv.

agents to solve tasks using large tool libraries. arXiv preprint, 2024.
personal assistants, arXiv preprint arXiv:2505.06328, 2025. URL https://arxiv.org/abs/
using a retrieval-augmented language model, arXiv preprint arXiv:2503.14103, 2025. URL https:
[816] A. Orhan. Recognition, recall, and retention of few-shot memories in large language models, arXiv preprint arXiv:2303.17557, 2023. URL https://arxiv.org/abs/2303.17557v1.
of llms for test case generation. arXiv preprint, 2024.
Memgpt: Towards llms as operating systems, arXiv preprint arXiv:2310.08560, 2023. URL https:
Gonzalez. Memgpt: Towards llms as operating systems, arXiv preprint arXiv:2310.08560, 2024.
learning and inference-time scaling law, arXiv preprint arXiv:2505.02665, 2025. URL https:
in gated llms, arXiv preprint arXiv:2504.21239, 2025. URL https://arxiv.org/abs/2504.

long chain-of-thought in language models without distillation, arXiv preprint arXiv:2502.03860,
arXiv preprint arXiv:2303.09014, 2023. URL https://arxiv.org/abs/2303.09014v1.
llmself-playingandself-improving,arXivpreprintarXiv:2503.03967,2025. URLhttps://arxiv.
connectedwithmassiveapis,arXivpreprintarXiv:2305.15334,2023. URLhttps://arxiv.org/
automated reasoning, arXiv preprint arXiv:2504.00428, 2025. URL https://arxiv.org/abs/
arXiv preprint, 2024.

Chen. A survey of useful llm evaluation, arXiv preprint arXiv:2406.00936, 2024. URL https:
JonathanJ.Halcrow. Letyourgraphdothetalking: Encodingstructureddataforllms,arXivpreprint
arXiv:2402.05862, 2024. URL https://arxiv.org/abs/2402.05862v1.
ment: A survey, arXiv preprint arXiv:2407.17030, 2024. URL https://arxiv.org/abs/2407.
with sequence order recall tasks, arXiv preprint arXiv:2410.08133, 2024. URL https://arxiv.
language models, a survey, arXiv preprint arXiv:2503.23037, 2025. URL https://arxiv.org/

Rafailov. Agent q: Advanced reasoning and learning for autonomous ai agents, arXiv preprint arXiv:2408.07199, 2024. URL https://arxiv.org/abs/2408.07199v1.
arXiv preprint arXiv:2405.17846, 2024. URL https://arxiv.org/abs/2405.17846v1.
via self-evolving online curriculum reinforcement learning, arXiv preprint arXiv:2411.02337, 2024.
agents for software development, arXiv preprint arXiv:2307.07924, 2024. URL https://arxiv.
developing agents, arXiv preprint arXiv:2405.04219, 2024. URL https://arxiv.org/abs/
and Heng Ji. Toolrl: Reward is all tool learning needs, arXiv preprint arXiv:2504.13958, 2025. URL
management via query-guided activation refilling, arXiv preprint arXiv:2412.12486, 2024. URL

arXiv preprint arXiv:2409.05591, 2025. URL https://arxiv.org/abs/2409.05591.
autonomous driving, arXiv preprint arXiv:2505.15298, 2025. URL https://arxiv.org/abs/
memory, arXiv preprint arXiv:2505.22006, 2025. URL https://arxiv.org/abs/2505.
inspired hierarchical video decomposition with transformative representations, arXiv preprint arXiv:2204.10105, 2022. URL https://arxiv.org/abs/2204.10105v3.
and future directions, arXiv preprint arXiv:2208.13629, 2022. URL https://arxiv.org/abs/
Improving zero-shot chain-of-thought reasoning across languages, arXiv preprint arXiv:2310.14799,
understanding, arXiv preprint arXiv:2406.10505, 2024. URL https://arxiv.org/abs/2406.

Afreelunchforhandlingunlimitedsequencelengthsinlargelanguagemodels. arXivpreprint,2024.
Training-free agent distillation with generalizable mcp boxes, arXiv preprint arXiv:2506.14728,
multimodal historical reasoning: Histbench and histagent, arXiv preprint arXiv:2505.20246, 2025.
models: Language, multimodality, and beyond, arXiv preprint arXiv:2503.21614, 2025. URL
Verhoef. Memory-augmented generative adversarial transformers, arXiv preprint arXiv:2402.19218,

and S. Roukos. Self-refinement of language models from external proxy metrics feedback, arXiv preprint arXiv:2403.00827, 2024. URL https://arxiv.org/abs/2403.00827v1.
[893] Sumedh Rasal. An artificial neuron for enhanced problem solving in large language models, arXiv preprint arXiv:2404.14222, 2024. URL https://arxiv.org/abs/2404.14222v1.
A review of trust, risk, and security management in llm-based agentic multi-agent systems, arXiv preprint arXiv:2506.04133, 2025. URL https://arxiv.org/abs/2506.04133v2.
[895] Jing Ren and Feng Xia. Brain-inspired artificial intelligence: A comprehensive review, arXiv preprint arXiv:2408.14811, 2024. URL https://arxiv.org/abs/2408.14811v1.
intelligence: A survey of llm-based scientific agents, arXiv preprint arXiv:2503.24047, 2025. URL

automation, arXiv preprint arXiv:2001.03543, 2020. URL https://arxiv.org/abs/2001.
Fimp: Foundation model-informed message passing for graph neural networks, arXiv preprint arXiv:2210.09475v5, 2022. URL https://arxiv.org/abs/2210.09475v5.
developpromptingproficiencyinarchitecturalai-generatedimages,arXivpreprintarXiv:2504.13948, arXiv preprint arXiv:2502.00757, 2025. URL https://arxiv.org/abs/2502.00757v3.
algorithms for collective behavior: a structural taxonomy, arXiv preprint arXiv:1803.05464, 2018.
for collective behavior: A structural and application-focused atlas, arXiv preprint arXiv:2103.11067,

attention transformers, arXiv preprint arXiv:2306.13501, 2023. URL https://arxiv.org/abs/
planning and tool usage, arXiv preprint arXiv:2308.03427, 2023. URL https://arxiv.org/
Waseem Alshikh. Writing in the margins: Better inference pattern for long context retrieval, arXiv preprint arXiv:2408.14906, 2024. URL https://arxiv.org/abs/2408.14906v1.
arXiv preprint arXiv:2411.13157, 2024. URL https://arxiv.org/abs/2411.13157v2.
edge graphs, arXiv preprint arXiv:2410.18251, 2024. URL https://arxiv.org/abs/2410.
arXiv preprint arXiv:2402.07927, 2024. URL https://arxiv.org/abs/2402.07927v2.
sine Benajiba. Meminsight: Autonomous memory augmentation for llm agents, arXiv preprint arXiv:2503.21760, 2025. URL https://arxiv.org/abs/2503.21760.
smart spaces, arXiv preprint arXiv:2505.00472, 2025. URL https://arxiv.org/abs/2505.

[931] S.Santhanam. Contextbasedtext-generationusinglstmnetworks,arXivpreprintarXiv:2005.00048,
conceptual taxonomy, applications and challenges, arXiv preprint arXiv:2505.10468, 2025. URL
designpatterncentricreview,arXivpreprintarXiv:2506.05364,2025. URLhttps://arxiv.org/
pretraining on diverse table data tasks, arXiv preprint arXiv:2310.00789, 2023. URL https:

scientific collections. arXiv preprint, 2025.
integrated reinforcement learning. arXiv preprint, 2025.
languagemodels,arXivpreprintarXiv:2504.02441,2025. URLhttps://arxiv.org/abs/2504.
Query,learning,andapplications,arXivpreprintarXiv:2404.14809v2,2024. URLhttps://arxiv.
arXiv preprint arXiv:2310.10158, 2023. URL https://arxiv.org/abs/2310.10158.
by scaling test-time interaction, arXiv preprint arXiv:2506.07976, 2025. URL https://arxiv.
Qwenlong-cprs: Towards ∞-llms with dynamic context optimization. arXiv preprint, 2025.

[954] Zhuocheng Shen. Llm with tools: A survey, arXiv preprint arXiv:2409.18807, 2024. URL https:
Wang, Y. Jiang, and Wangchunshu Zhou. Taskcraft: Automated generation of agentic tasks, arXiv
preprint arXiv:2506.10055, 2025. URL https://arxiv.org/abs/2506.10055v2.
languagemodels,arXivpreprintarXiv:2407.03600,2024. URLhttps://arxiv.org/abs/2407.
Howfararewe?,arXivpreprintarXiv:2409.12682,2024. URLhttps://arxiv.org/abs/2409.

arXiv preprint arXiv:2504.20070, 2025. URL https://arxiv.org/abs/2504.20070v1.
Khoei. Exploring prompt engineering: A systematic review with swot analysis, arXiv preprint arXiv:2410.12843, 2024. URL https://arxiv.org/abs/2410.12843v1.
forllmsviareinforcementlearning,arXivpreprintarXiv:2505.01441,2025. URLhttps://arxiv.
commit history, arXiv preprint arXiv:2506.11060, 2025. URL https://arxiv.org/abs/2506.
heterophilic node classification, arXiv preprint arXiv:2503.05763, 2025. URL https://arxiv.
thorn. Contextually entangled gradient mapping for optimized llm comprehension, arXiv preprint arXiv:2502.00048, 2025. URL https://arxiv.org/abs/2502.00048v1.
actions, arXiv preprint arXiv:2310.03720, 2024. URL https://arxiv.org/abs/2310.03720.

[980] Manthankumar Solanki. Efficient document retrieval with g-retriever. arXiv preprint, 2025.
arXiv preprint, 2025.
long-context processing in transformers, arXiv preprint arXiv:2506.01215, 2025. URL https:
calling and routing, arXiv preprint arXiv:2501.05255, 2025. URL https://arxiv.org/abs/
arXiv preprint arXiv:2410.16464, 2025. URL https://arxiv.org/abs/2410.16464.

agents to solve tasks using large tool libraries. arXiv preprint, 2024.
personal assistants, arXiv preprint arXiv:2505.06328, 2025. URL https://arxiv.org/abs/
using a retrieval-augmented language model, arXiv preprint arXiv:2503.14103, 2025. URL https:
[816] A. Orhan. Recognition, recall, and retention of few-shot memories in large language models, arXiv preprint arXiv:2303.17557, 2023. URL https://arxiv.org/abs/2303.17557v1.
of llms for test case generation. arXiv preprint, 2024.
Memgpt: Towards llms as operating systems, arXiv preprint arXiv:2310.08560, 2023. URL https:
Gonzalez. Memgpt: Towards llms as operating systems, arXiv preprint arXiv:2310.08560, 2024.
learning and inference-time scaling law, arXiv preprint arXiv:2505.02665, 2025. URL https:
in gated llms, arXiv preprint arXiv:2504.21239, 2025. URL https://arxiv.org/abs/2504.

graph. arXiv preprint, 2023.
llms and knowledge graphs, arXiv preprint arXiv:2404.07677, 2024. URL https://arxiv.org/
tion of large language models and knowledge graphs, arXiv preprint arXiv:2410.12298, 2024. URL
Yang Shen. Multi-agent coordination across diverse applications: A survey, arXiv preprint
arXiv:2502.14743, 2025. URL https://arxiv.org/abs/2502.14743v2.
arXiv preprint, 2024.
arXiv preprint arXiv:2506.01659, 2025. URL https://arxiv.org/abs/2506.01659v1.

and Ruslan Salakhutdinov. Training a generally curious agent, arXiv preprint arXiv:2502.17543,
intelligence, arXiv preprint arXiv:2503.13754, 2025. URL https://arxiv.org/abs/2503.
base question answering with in-context learning, arXiv preprint arXiv:2305.13972, 2023. URL
Enhancing large language models reasoning with structured data. arXiv preprint, 2024.
multimodal large language models, arXiv preprint arXiv:2410.16983v1, 2024. URL https://
surveyon(m)llm-basedguiagents,arXivpreprintarXiv:2504.13865,2025. URLhttps://arxiv.
large language models on graph computation. arXiv preprint arXiv:2407.00379, 2024.

library in large language models improves chemical reasoning, arXiv preprint arXiv:2501.06590,
work, arXiv preprint arXiv:2201.08975, 2022. URL https://arxiv.org/abs/2201.08975v1.
dynamic token pruning for large language models, arXiv preprint arXiv:2504.04514, 2025. URL
els, arXiv preprint arXiv:2402.01812, 2024. URL https://arxiv.org/abs/2402.01812v1.
Typhoon t1: An open thai reasoning model, arXiv preprint arXiv:2502.09042, 2025. URL https:
An efficient data augmentation strategy for long-context pre-training in language models, arXiv
preprint arXiv:2409.04774, 2024. URL https://arxiv.org/abs/2409.04774v1.
video reasoning. arXiv preprint, 2025.

xr with llm-powered conversational agents, arXiv preprint arXiv:2504.05527, 2025. URL https:
Nguyen. Multi-agent collaboration mechanisms: A survey of llms, arXiv preprint arXiv:2501.06322,
code, arXiv preprint arXiv:2503.12188, 2025. URL https://arxiv.org/abs/2503.12188v1.
multi-modality learning, arXiv preprint arXiv:2402.08086v2, 2024. URL https://arxiv.org/

and-copy attention heads for multiple-choice qa, arXiv preprint arXiv:2410.02343, 2024. URL
heterogeneous executors for offensive security. arXiv preprint, 2025.
warning system investments, arXiv preprint arXiv:2504.05104, 2025. URL https://arxiv.org/
verificationusingllmagentssystem,arXivpreprintarXiv:2409.03440,2024. URLhttps://arxiv.
tagent: Adapting multimodal web agents with few-shot learning from human demonstrations, arXiv
preprint arXiv:2411.13451, 2024. URL https://arxiv.org/abs/2411.13451v1.
generation from structured data, arXiv preprint arXiv:2404.15604, 2024. URL https://arxiv.

[1073] James Vo. Sparseaccelerate: Efficient long-context inference for mid-range gpus, arXiv preprint arXiv:2412.06198, 2024. URL https://arxiv.org/abs/2412.06198v1.
Exploring knowledge graph-large language model synergies. arXiv preprint, 2025.
duction, arXiv preprint arXiv:2505.15784, 2025. URL https://arxiv.org/abs/2505.15784.
with multi turns, arXiv preprint arXiv:2506.13356, 2025. URL https://arxiv.org/abs/2506.
meta-reinforcement learning with self-supervised trajectory contrastive learning, arXiv preprint arXiv:2103.06386, 2021. URL https://arxiv.org/abs/2103.06386v1.
Zhoujun Li. Scm: Enhancing large language model with self-controlled memory framework, arXiv preprint arXiv:2304.13343, 2025. URL https://arxiv.org/abs/2304.13343.
arXiv preprint arXiv:2405.17234, 2024. URL https://arxiv.org/abs/2405.17234v6.
Anima Anandkumar. Voyager: An open-ended embodied agent with large language models, arXiv preprint arXiv:2305.16291, 2023. URL https://arxiv.org/abs/2305.16291.

engineering?, arXiv preprint arXiv:2411.02093, 2024. URL https://arxiv.org/abs/2411.
Evaluation and methodology. arXiv preprint arXiv:2507.07999, 2025.
Tuning llama model with chinese medical knowledge, arXiv preprint arXiv:2304.06975, 2023. URL
arXiv preprint arXiv:2502.14477, 2025. URL https://arxiv.org/abs/2502.14477v1.
Wong. Toward a theory of agents as tool-use decision-makers, arXiv preprint arXiv:2506.00886,
[1096] Jingjin Wang. Proprag: Guiding retrieval with beam search over proposition paths, arXiv preprint
arXiv:2504.18070, 2025. URL https://arxiv.org/abs/2504.18070v1.

thought machines (ctm) and model context protocol (mcp), arXiv preprint arXiv:2505.19339, 2025.
language model for the aviation domain, arXiv preprint arXiv:2311.17686, 2023. URL https:
Dongmei Zhang, and Qi Zhang. Large action models: From inception to implementation, arXiv preprint arXiv:2412.10047, 2025. URL https://arxiv.org/abs/2412.10047.
Symmetry of agents and interplay with prompts, arXiv preprint arXiv:2311.07076, 2023. URL
tracking,arXivpreprintarXiv:2306.00434,2023. URLhttps://arxiv.org/abs/2306.00434.
based on segment-wise inference, arXiv preprint arXiv:2405.17755, 2024. URL https://arxiv.

Y. Jiang, and Wangchunshu Zhou. Weaver: Foundation models for creative writing, arXiv preprint arXiv:2401.17268, 2024. URL https://arxiv.org/abs/2401.17268v1.
Hong-Jun Yoon, M. Wahib, and J. Gounley. Ultra-long sequence distributed transformer, arXiv preprint arXiv:2311.02382, 2023. URL https://arxiv.org/abs/2311.02382v2.
lingual cross-modal ambiguity resolution for multimodal large language models, arXiv preprint arXiv:2506.17046v1, 2025. URL https://arxiv.org/abs/2506.17046v1.
retrieval via reversible compression. arXiv preprint, 2025.
ommendation, arXiv preprint arXiv:2308.14296, 2024. URL https://arxiv.org/abs/2308.

arXiv preprint arXiv:2409.09030, 2024. URL https://arxiv.org/abs/2409.09030v2.
Limin Wang. Internvideo2.5: Empowering video mllms with long and rich context modeling. arXiv
explanations, arXiv preprint arXiv:2505.22823, 2025. URL https://arxiv.org/abs/2505.
languagemodels,arXivpreprintarXiv:2402.04624,2024. URLhttps://arxiv.org/abs/2402.
arXiv preprint, 2025.
improved structure modeling. arXiv preprint, 2024.
and Dusit Niyato. Internet of agents: Fundamentals, applications, and challenges, arXiv preprint
arXiv:2505.07176, 2025. URL https://arxiv.org/abs/2505.07176v1.

comprehensive benchmark and investigation, arXiv preprint arXiv:2502.18771v1, 2025. URL
WenhaoHuang. Mio: Afoundationmodelonmultimodaltokens,arXivpreprintarXiv:2409.17692v3,
a survey from the language model perspective, arXiv preprint arXiv:2403.15452, 2024. URL
Honeygpt: Breaking the trilemma in terminal honeypots with large language model, arXiv preprint
arXiv:2406.01882, 2024. URL https://arxiv.org/abs/2406.01882v2.
[1144] Zora Zhiruo Wang, Jiayuan Mao, Daniel Fried, and Graham Neubig. Agent workflow memory, arXiv
preprint arXiv:2409.07429, 2024. URL https://arxiv.org/abs/2409.07429.
and simulation of storage performance, arXiv preprint arXiv:2401.00381, 2023. URL https:
learning and its advancements on large language models: A holistic survey, arXiv preprint
arXiv:2212.08966, 2022. URL https://arxiv.org/abs/2212.08966v5.
multi-turn reinforcement learning. arXiv preprint, 2025.

vulnerability detection. arXiv preprint, 2024.
arXiv preprint, 2025.
[1153] Danny Weyns and F. Oquendo. An architectural style for self-adaptive multi-agent systems, arXiv preprint arXiv:1909.03475, 2019. URL https://arxiv.org/abs/1909.03475v1.
multilingual continual learning, arXiv preprint arXiv:2305.16252, 2023. URL https://arxiv.
long-term interactions, arXiv preprint arXiv:2505.23662, 2025. URL https://arxiv.org/abs/
and recent trends in multimodal mobile agents: A survey, arXiv preprint arXiv:2411.02006, 2024.
information seeking agency, arXiv preprint arXiv:2505.22648, 2025. URL https://arxiv.org/
Pengjun Xie, and Fei Huang. Webwalker: Benchmarking llms in web traversal, arXiv preprint arXiv:2501.07572, 2025. URL https://arxiv.org/abs/2501.07572v2.
deep research, arXiv preprint arXiv:2502.04644, 2025. URL https://arxiv.org/abs/2502.

form: Reasoning large language model for communication system formulation, arXiv preprint arXiv:2506.08551, 2025. URL https://arxiv.org/abs/2506.08551v2.
Enabling next-gen llm applications via multi-agent conversation, arXiv preprint arXiv:2308.08155,
tion answering, arXiv preprint arXiv:2506.00232, 2025. URL https://arxiv.org/abs/2506.
ing: A survey, arXiv preprint arXiv:2407.13193, 2024. URL https://arxiv.org/abs/2407.
for evaluating knowledge editing of llms, arXiv preprint arXiv:2308.09954, 2023. URL https:
llm-as-a-meta-judge. arXiv preprint, 2024.
models through thinking intervention, arXiv preprint arXiv:2503.24370, 2025. URL https://
through structured data, arXiv preprint arXiv:2412.10654, 2024. URL https://arxiv.org/

llms, arXiv preprint arXiv:2504.15965, 2025. URL https://arxiv.org/abs/2504.15965v2.
diversity tradeoff in adaptive multi-agent systems, arXiv preprint arXiv:2502.16565, 2025. URL
Huan, and Tao Gui. The rise and potential of large language model based agents: A survey, arXiv preprint arXiv:2309.07864, 2023. URL https://arxiv.org/abs/2309.07864v3.
memory test benchmark for language models, arXiv preprint arXiv:2502.03358, 2025. URL https:
language models, arXiv preprint arXiv:2504.12345v3, 2025. URL https://arxiv.org/abs/
ation, arXiv preprint arXiv:2506.05690, 2025. URL https://arxiv.org/abs/2506.05690v1.

Chaochao Jia, Dahai Li, and Maosong Sun. Minicpm4: Ultra-efficient llms on end devices, arXiv preprint arXiv:2506.07900, 2025. URL https://arxiv.org/abs/2506.07900v1.
Reasoning refinement for efficient and effective test-time scaling, arXiv preprint arXiv:2505.19187,
ing path: Distilling effective guidance for llm reasoning with knowledge graphs, arXiv preprint arXiv:2506.10508, 2025. URL https://arxiv.org/abs/2506.10508v1.
arXiv preprint arXiv:2410.09675, 2024. URL https://arxiv.org/abs/2410.09675v1.
anisms for stable context representationin large language models, arXiv preprint arXiv:2505.22921,
language agents for retrieval-augmented generation. arXiv preprint, 2025.
autonomous agents, arXiv preprint arXiv:2407.08516, 2024. URL https://arxiv.org/abs/
els, arXiv preprint arXiv:2505.16957, 2025. URL https://arxiv.org/abs/2505.16957v1.
arXiv preprint arXiv:2505.16067, 2025. URL https://arxiv.org/abs/2505.16067v1.

KaiYu. Reducingtoolhallucinationviareliabilityalignment,arXivpreprintarXiv:2412.04141,2024.
Test-time reinforcement learning-based model caching and inference offloading, arXiv preprint arXiv:2501.14205, 2025. URL https://arxiv.org/abs/2501.14205v1.
language games, arXiv preprint arXiv:2505.18218, 2025. URL https://arxiv.org/abs/2505.
Noderag: Structuring graph-based rag with heterogeneous nodes, arXiv preprint arXiv:2504.11544,
[1210] Wenrui Xu and Keshab K. Parhi. A survey of attacks on large language models, arXiv preprint arXiv:2505.12567, 2025. URL https://arxiv.org/abs/2505.12567v1.
memory for llm agents. arXiv preprint, 2025.
memoryforllmagents,arXivpreprintarXiv:2502.12110,2025. URLhttps://arxiv.org/abs/

question synthesis for thinking-centric fine-tuning?, arXiv preprint arXiv:2503.09499, 2025. URL
modelpipelinerefinementandoptimizationleveragingllmexperts,arXivpreprintarXiv:2502.18530,
arXiv preprint arXiv:2409.01392, 2024. URL https://arxiv.org/abs/2409.01392v2.
Li. Beyond self-talk: A communication-centric survey of llm-based multi-agent systems. arXiv
[1220] Tianqiang Yan and Tiansheng Xu. Refining the responses of llms by themselves, arXiv preprint
arXiv:2305.04039, 2023. URL https://arxiv.org/abs/2305.04039v1.
Inftythink: Breaking the length limits of long-context reasoning in large language models, arXiv
preprint arXiv:2503.06692, 2025. URL https://arxiv.org/abs/2503.06692v3.

guage models, arXiv preprint arXiv:2311.04879, 2023. URL https://arxiv.org/abs/2311.
arXiv preprint, 2023.
Multimodal large diffusion language models, arXiv preprint arXiv:2505.15809v1, 2025. URL
sparse attention, arXiv preprint arXiv:2502.14866, 2025. URL https://arxiv.org/abs/2502.
Evaluating large language models on pest management in agriculture. arXiv preprint, 2024.
text, deeper thinking: Uncovering the role of long-context ability in reasoning, arXiv preprint
arXiv:2505.17315, 2025. URL https://arxiv.org/abs/2505.17315v1.

multiagent reinforcement learning, arXiv preprint arXiv:2011.00583, 2020. URL https://arxiv.
languagemodelsforpredictivetabulartasksindatascience,arXivpreprintarXiv:2403.20208,2024.
financialdomaininstructiontuning,arXivpreprintarXiv:2309.13064,2023. URLhttps://arxiv.
of ai agent protocols, arXiv preprint arXiv:2504.16736, 2025. URL https://arxiv.org/abs/
language model os, arXiv preprint arXiv:2409.01495, 2024. URL https://arxiv.org/abs/
Meta-pathguidedretrievalandin-graphtextforrag-equippedllm,arXivpreprintarXiv:2503.00309,
Comal: Collaborative multi-agent large language models for mixed-autonomy traffic, arXiv preprint
arXiv:2410.14368, 2024. URL https://arxiv.org/abs/2410.14368v2.
transformer, arXiv preprint arXiv:2408.16978, 2024. URL https://arxiv.org/abs/2408.

React: Synergizingreasoningandactinginlanguagemodels,arXivpreprintarXiv:2210.03629,2023.
user interaction in real-world domains. arXiv preprint, 2024.
arXiv preprint arXiv:2308.02151, 2024. URL https://arxiv.org/abs/2308.02151.
query-driven benchmark for evaluating large language models in multi-hop tool use, arXiv preprint arXiv:2501.02506, 2025. URL https://arxiv.org/abs/2501.02506v4.

Prompt alchemy: Automatic prompt refinement for enhancing code generation, arXiv preprint arXiv:2503.11085, 2025. URL https://arxiv.org/abs/2503.11085v1.
Shmueli-Scheuer. Survey on evaluation of llm-based agents, arXiv preprint arXiv:2503.16416, 2025.
data synthesis and distillation via graph translation, arXiv preprint arXiv:2503.07826, 2025. URL

Dong Yu. Teaching llms to refine with tools, arXiv preprint arXiv:2412.16871, 2024. URL https:
llm agents: Threats and countermeasures, arXiv preprint arXiv:2503.09648, 2025. URL https:
efficient mathematical reasoning in large models, arXiv preprint arXiv:2506.10716, 2025. URL
of llava in visual question answering, arXiv preprint arXiv:2411.10950v2, 2024. URL https:
Adaptivereasoningwithinferenceawareoptimization,arXivpreprintarXiv:2501.17974,2025. URL
visual tool reinforcement learning, arXiv preprint arXiv:2505.08617, 2025. URL https://arxiv.
language model agents to reflect via iterative self-training. arXiv preprint, 2025.
arXiv preprint, 2025.
self-evolutionary reinforcement learning, arXiv preprint arXiv:2505.12370, 2025. URL https:
[1290] Murong Yue. A survey of large language model agents for question answering, arXiv preprint arXiv:2503.19213, 2025. URL https://arxiv.org/abs/2503.19213v1.

arXiv preprint arXiv:2203.14465, 2022. URL https://arxiv.org/abs/2203.14465v2.
self-improving code generation, arXiv preprint arXiv:2310.02304, 2023. URL https://arxiv.
arXiv preprint arXiv:2405.11299, 2024. URL https://arxiv.org/abs/2405.11299v2.
arXiv preprint arXiv:2412.15266, 2024. URL https://arxiv.org/abs/2412.15266v1.
arXiv preprint arXiv:2501.09766, 2025. URL https://arxiv.org/abs/2501.09766v4.
optimization, arXiv preprint arXiv:2502.05605, 2025. URL https://arxiv.org/abs/2502.

Chao Du, et al. Ufo2: The desktop agentos. arXiv preprint arXiv:2504.14603, 2025.
throughactiveself-reflection,arXivpreprintarXiv:2502.14932,2025. URLhttps://arxiv.org/
guage models, arXiv preprint arXiv:2406.05678, 2024. URL https://arxiv.org/abs/2406.
memoryinjection,arXivpreprintarXiv:2404.03565,2025.URLhttps://arxiv.org/abs/2404.

Jing Mai, Bin Gu, and Zhi Jin. Computational thinking reasoning in large language models, arXiv preprint arXiv:2506.02658, 2025. URL https://arxiv.org/abs/2506.02658v2.
where, and how well?, arXiv preprint arXiv:2503.24235, 2025. URL https://arxiv.org/abs/
complementary strengths of large and small llms, arXiv preprint arXiv:2502.07942, 2025. URL
Coordfield: Coordination field for agentic uav task allocation in low-altitude urban scenarios, arXiv preprint arXiv:2505.00091, 2025. URL https://arxiv.org/abs/2505.00091v3.
tion at test time, arXiv preprint arXiv:2506.06254, 2025. URL https://arxiv.org/abs/2506.
reward: Which is better for agentic rag reinforcement learning, arXiv preprint arXiv:2505.14069,
erarchicalmulti-agentframeworkforgeneral-purposetasksolving,arXivpreprintarXiv:2506.12508,
video discovery: Agentic search with tool use for long-form video understanding, arXiv preprint arXiv:2505.18079, 2025. URL https://arxiv.org/abs/2505.18079v2.

autonomous multi-agent system for web task execution with strategic exploration, arXiv preprint arXiv:2408.15978, 2024. URL https://arxiv.org/abs/2408.15978.
llms to master multi-api planning, arXiv preprint arXiv:2310.04474, 2023. URL https://arxiv.
verificationandwronginformation,arXivpreprintarXiv:2410.04463,2024. URLhttps://arxiv.
Zhang. Evaluating and steering modality preferences in multimodal large language model, arXiv preprint arXiv:2505.20977v1, 2025. URL https://arxiv.org/abs/2505.20977v1.
chain-of-action generation into reasoning models, arXiv preprint arXiv:2503.06580, 2025. URL
Ji-Rong Wen. A survey on the memory mechanism of large language model based agents, arXiv preprint arXiv:2404.13501, 2024. URL https://arxiv.org/abs/2404.13501v1.
personal assistants, arXiv preprint arXiv:2409.20163, 2024. URL https://arxiv.org/abs/
Robust sequence-to-sequence learning via self-supervised input representation, arXiv preprint arXiv:2204.07837, 2022. URL https://arxiv.org/abs/2204.07837v2.

agents are experiential learners, arXiv preprint arXiv:2308.10144, 2024. URL https://arxiv.
arXiv preprint arXiv:2402.19473, 2024. URL https://arxiv.org/abs/2402.19473v6.
intelligence agents, arXiv preprint arXiv:2309.14365, 2023. URL https://arxiv.org/abs/
learning enhancement. arXiv preprint, 2024.
reasoningviallm-generatedinferencepathsoverknowledgegraphs,arXivpreprintarXiv:2502.12029,
grapheuleriantransformer,arXivpreprintarXiv:2401.00529v3,2023. URLhttps://arxiv.org/
ing? revisiting long-cot compression with capability in mind for better reasoning, arXiv preprint
arXiv:2505.14582, 2025. URL https://arxiv.org/abs/2505.14582v1.

deliberative and adaptive reasoning over foundational capabilities, arXiv preprint arXiv:2503.17979,
rag for high efficiency and effectiveness. arXiv preprint, 2025.
Ma. Lifelongagentbench: Evaluatingllmagentsaslifelonglearners,arXivpreprintarXiv:2505.11942,
prompting with memory for computer control, arXiv preprint arXiv:2306.07863, 2024. URL
DachengTao,L.V.Gool,andXumingHu. Mllmsaredeeplyaffectedbymodalitybias,arXivpreprint
arXiv:2505.18657v1, 2025. URL https://arxiv.org/abs/2505.18657v1.
Deepresearcher: Scalingdeepresearchviareinforcementlearninginreal-worldenvironments,arXiv
preprint arXiv:2504.03160, 2025. URL https://arxiv.org/abs/2504.03160v4.

large language models with long-term memory, arXiv preprint arXiv:2305.10250, 2023. URL
and Emine Yilmaz. Trustrag: Enhancing robustness and trustworthiness in rag, arXiv preprint
arXiv:2501.00879, 2025. URL https://arxiv.org/abs/2501.00879.
languageagents,arXivpreprintarXiv:2309.07870,2023. URLhttps://arxiv.org/abs/2309.
unified framework, arXiv preprint arXiv:2503.04338, 2025. URL https://arxiv.org/abs/

arXiv preprint arXiv:2409.10102, 2024. URL https://arxiv.org/abs/2409.10102v1.
long-sequenceprocessingusinglargelanguagemodels,arXivpreprintarXiv:2410.09342,2024. URL
arXiv preprint arXiv:2506.15841, 2025. URL https://arxiv.org/abs/2506.15841v1.
language models. arXiv preprint, 2023.
evaluation of llm-based ai agents: A comprehensive survey, arXiv preprint arXiv:2506.11102, 2025.
Exploringadvancedtrainingandtest-timerecipesforopen-sourcemultimodalmodels,arXivpreprint
arXiv:2504.10479v3, 2025. URL https://arxiv.org/abs/2504.10479v3.
sequentialvisualinputreasoningandpredictioninmultimodallargelanguagemodels,arXivpreprint
arXiv:2310.13473v1, 2023. URL https://arxiv.org/abs/2310.13473v1.

knowledge and memory, arXiv preprint arXiv:2305.17144, 2023. URL https://arxiv.org/
management for prefix prefilling in llm inference, arXiv preprint arXiv:2505.21919, 2025. URL
Wang. Conversational crowdsensing: A parallel intelligence powered novel sensing approach, arXiv preprint arXiv:2402.06654, 2024. URL https://arxiv.org/abs/2402.06654v1.
approaches and functionalities in retrieval-augmented generation: A comprehensive survey, arXiv preprint arXiv:2504.10499, 2025. URL https://arxiv.org/abs/2504.10499v1.
knowledge grounding, arXiv preprint arXiv:2402.16671, 2024. URL https://arxiv.org/abs/
system prompts against prompt extraction attacks, arXiv preprint arXiv:2505.11459, 2025. URL

of benchmarking multimodal in-context learning. arXiv preprint, 2024.
Jiang, and Philip S. Yu. A survey on large language model based human-agent systems. arXiv languagemodels,arXivpreprintarXiv:2506.10943,2025. URLhttps://arxiv.org/abs/2506.