HERAKLES: Hierarchical Skill Compilation for Open-ended LLM Agents

Anonymous Author(s)

1 Introduction

Recent advances in AI have yielded agents with human-level performance in vision and language tasks, driven by foundation models trained on large-scale internet data [25, 1, 9]. However, these static datasets limit progress toward general intelligence [22]. In contrast, humans acquire diverse skills continuously via open-ended interaction with their environment. Replicating this ability is a central goal of AI: to build autotelic agents that self-generate goals and learn without fixed datasets [4, 17, 6]. Recent systems like MineDojo [7], Voyager [23], OMNI [26], and ACES [15] leverage foundation models for open-ended learning. These agents autonomously generate or select goals, prioritize learning progress [8], and adapt via curricula. Yet, as goal complexity increases, the combinatorial growth in required subskills slows learning [20, 3, 14]. Humans mitigate this through hierarchical learning: decomposing complex skills into reusable subskills [19]. Inspired by this, AI research has adopted hierarchical structures in vision [12] and reinforcement learning [18, 16], increasingly incorporating language for goal decomposition [2, 11]. However, most approaches assume predefined skills, limiting adaptability in open-ended settings. We introduce **HERAKLES** (**HiERA**rchic**A**l s**K**ill compi**L**ation for open-**E**nded agent**S**): a method for training autotelic agents with jointly learned high-level (HL) and low-level (LL) policies in dynamic goal spaces. HERAKLES uses an LLM-based HL policy to select mastered subgoals and guide the LL policy, which compiles skills into an efficient, executable form. Both components co-evolve via an online curriculum without requiring pre-trained skills. We evaluate HERAKLES in the Crafter environment [10]. HERAKLES simultaneously learns π^{HL} and π^{LL}, respectively the HL and LL policies (see Figure 1). π^{HL} is a pre-trained LLM, fine-tuned using RL. It samples skills from the set of skill $\tilde{G}_k$, the skill space constructed at each step, using constrained decoding. π^{LL} is a small, not pre-trained, neural network also trained using RL. As π^{HL} masters a goal g, it is distilled into π^{LL}. π^{HL} can then use π^{LL} to reach g, for example inside a trajectory to achieve a more complex goal g'.

2 Experiments

We evaluate HERAKLES in Crafter [10] (modify similarly to [5]), a 2D Minecraft-like environment with procedural generation and partial observability. Goals are organized in an achievement tree, often requiring the reuse of previously acquired artifacts (e.g., crafting a pickaxe requires a table). We assume goals are externally generated and focus on efficient hierarchical learning. Agents in Crafter are encouraged to master a wide range of goals from an achievement tree. Goal difficulty is state-dependent; e.g., `place table` is easier with wood in the inventory. To select goals that maximize learning progress, we use MAGELLAN [8], which estimates competence and learning progress online. MAGELLAN further improves generalization by leveraging semantic relations between goals. We instantiate π^{HL} with Mistral 7B, trained using the POAD on-policy RL algorithm [24], sampling skills via constrained decoding. π^{LL} is a 2M-parameter ResNet adapted from [13] and trained with the AWR off-policy algorithm. We compare HERAKLES against: textscPOAD[24], using only π^{HL} with the action space restricted to primitives A, to isolate the impact of hierarchy, FUN[21], a standard HRL baseline where subgoals are sampled in a learned embedding space. To match our setup, we make FUN's HL policy goal-conditioned by providing goal embeddings generated by the same LLM used in HERAKLES. Crafter features a heterogeneous and compositional goal space, where more difficult goals require chaining an increasing number of elementary actions. To assess how HERAKLES scales with goal difficulty, we evaluate sample efficiency by training agents for 30,000 high-level steps and measuring progress with the Crafter score [10]:$S_c = \exp\left(\frac{1}{N}\sum_{i=1}^{N}\ln(1+sr_i)\right) - 1$, where $sr_i \in [0, 100]$ is the success rate for goal i, and $N = 10$ is the total number of goals. This metric emphasizes rare and difficult achievements via geometric averaging. As shown in Figure 2, HERAKLES rapidly accumulates successful goals, leveraging compiled skills to accelerate learning of more complex tasks. In contrast, POAD plateaus at $S_c = 5$, under-

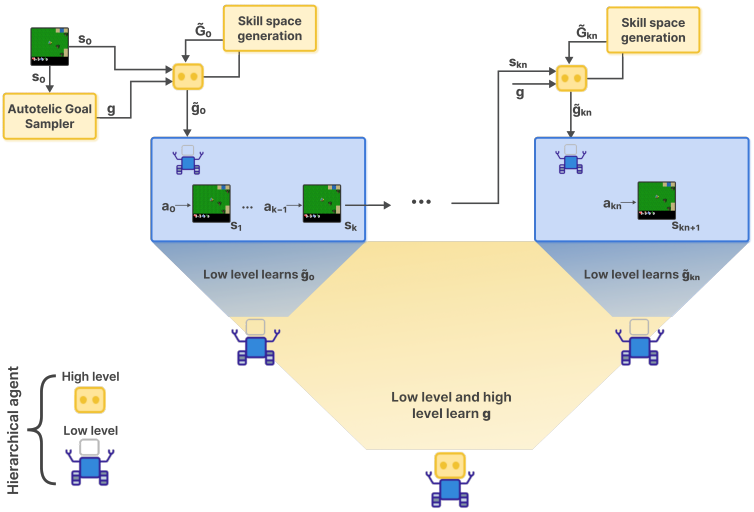

Figure 1: **Skill learning and compilation in HERAKLES**: Given a goal g and initial state s_0, the high-level policy π^{HL} constructs a skill space $\tilde{G}_0$ and samples a skill $\tilde{g}_0 \in \tilde{G}_0$. The low-level policy π^{LL} then executes k primitive actions to reach $\tilde{g}_0$, resulting in state s_k. This process iterates: π^{HL} samples a new skill $\tilde{g}_k$ given (s_k, g), and π^{LL} attempts to achieve it. The interaction continues until the goal g is reached or a step limit is exceeded, yielding:1)A high-level trajectory: the sequence of sampled skills 2) A set of low-level trajectories: one per skill, conditioned on reaching that skill. π^{HL} is trained on its trajectory to improve skill selection. π^{LL} is trained on all low-level segments, conditioned on their respective subgoals $\tilde{g}$.**Skill compilation** is performed by additionally training π^{LL} on the concatenated low-level trajectory, conditioned directly on g. This enables π^{LL} to gradually internalize full skill sequences, allowing direct goal execution without high-level intervention.

	HERAKLES	POAD	FUN
Original goals	49.4	3.3	27.3
Synonyms	41.5 (**-16%**)	2.5 (**-24%**)	19.9 (**-27%**)

Table 1: Generalization for synonyms goals.

performing even a random policy. Its success concentrates on trivial goals (e.g., `go to tree` with $sr = 1.0 \pm 0.0$), with minimal progress on complex ones (e.g., `make wood pickaxe` with $sr = 3.9 \pm 6.8 \times 10^{-3}$). FUN shows slow improvement, slightly above random, hindered by its inability to reuse mastered goals for skill composition.

We assess the generalization performance of HERAKLES, POAD, and FUN on a set of synonym-based goals. For each original goal, such as *"collect wood"*, we define a synonym set by selecting five alternative formulations (e.g., *"gather wood"*, *"harvest wood"*, *"procure wood"*, *"acquire wood"*, and *"amass wood"*), and compute the average Crafter score across these variants. Table 1 aggregates these instantaneous measurements into a single averaged metric over the entire training period. HERAKLES experiences only a 16% drop in average score relative to the original goal space, POAD and FUN exhibit more substantial decreases of 24% and 27%, respectively. These results highlight the superiority of HERAKLES in handling semantic variability in goal specification.

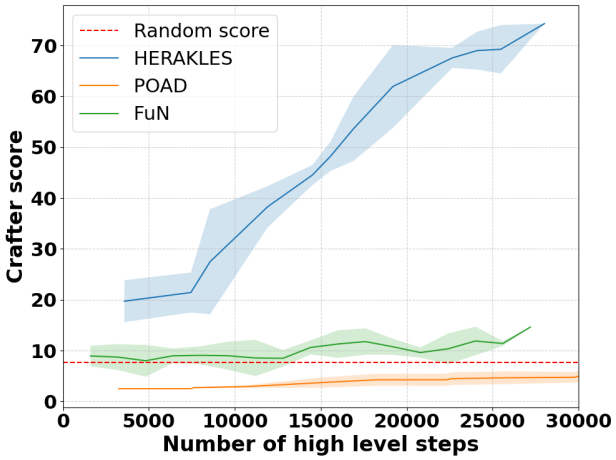

Figure 2: Number of goals reached as a function of the number of high-level steps. Shaded area denotes standard deviation over 4 seeds. HERAKLES is the only method with approximately linear goal acquisition over time.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo

Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2024. URL `https://arxiv.org/abs/2303.08774`.

[2] Arun Ahuja, Kavya Kopparapu, Rob Fergus, and Ishita Dasgupta. Hierarchical reinforcement learning with natural language subgoals, 2023. URL `https://arxiv.org/abs/2309.11564`.

[3] Jakob Bauer, Kate Baumli, Satinder Baveja, Feryal Behbahani, Avishkar Bhoopchand, Nathalie Bradley-Schmieg, Michael Chang, Natalie Clay, Adrian Collister, Vibhavari Dasagi, Lucy Gonzalez, Karol Gregor, Edward Hughes, Sheleem Kashem, Maria Loks-Thompson, Hannah Openshaw, Jack Parker-Holder, Shreya Pathak, Nicolas Perez-Nieves, Nemanja Rakicevic, Tim Rocktäschel, Yannick Schroecker, Jakub Sygnowski, Karl Tuyls, Sarah York, Alexander Zacherl, and Lei Zhang. Human-timescale adaptation in an open-ended task space, 2023. URL `https://arxiv.org/abs/2301.07608`.

[4] Cédric Colas, Tristan Karch, Clément Moulin-Frier, and Pierre-Yves Oudeyer. Language and culture internalization for human-like autotelic ai. *Nature Machine Intelligence*, 4(12):1068–1076, December 2022. ISSN 2522-5839. doi: 10.1038/s42256-022-00591-4. URL `http://dx.doi.org/10.1038/s42256-022-00591-4`.

[5] Yuqing Du, Olivia Watkins, Zihan Wang, Cédric Colas, Trevor Darrell, Pieter Abbeel, Abhishek Gupta, and Jacob Andreas. Guiding pretraining in reinforcement learning

with large language models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 8657–8677. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/du23f.html.

[6] Hughes Edward, Dennis Michael D, Parker-Holder Jack, Behbahani Feryal, Mavalankar Aditi, Shi Yuge, Schaul Tom, and Rocktäschel Tim. Position: Open-endedness is essential for artificial superhuman intelligence. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 20597–20616. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/hughes24a.html.

[7] Linxi Fan, Guanzhi Wang, Yunfan Jiang, Ajay Mandlekar, Yuncong Yang, Haoyi Zhu, Andrew Tang, De-An Huang, Yuke Zhu, and Anima Anandkumar. Minedojo: Building open-ended embodied agents with internet-scale knowledge. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 18343–18362. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/74a67268c5cc5910f64938cac4526a90-Paper-Datasets_and_Benchmarks.pdf.

[8] Loris Gaven, Thomas Carta, Clément Romac, Cédric Colas, Sylvain Lamprier, Olivier Sigaud, and Pierre-Yves Oudeyer. Magellan: Metacognitive predictions of learning progress guide autotelic llm agents in large goal spaces. In *International Conference on Machine Learning (ICML)*, 2025. URL https://arxiv.org/abs/2502.07709.

[9] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin

Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL https://arxiv.org/abs/2501.12948.

[10] Danijar Hafner. Benchmarking the spectrum of agent capabilities, 2022. URL https://arxiv.org/abs/2109.06780.

[11] YiDing Jiang, Shixiang (Shane) Gu, Kevin P Murphy, and Chelsea Finn. Language as an abstraction for hierarchical deep reinforcement learning. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/0af787945872196b42c9f73ead2565c8-Paper.pdf.

[12] Seyed Hamidreza Kasaei, Ana Maria Tomé, and Luís Seabra Lopes. Hierarchical object representation for open-ended object category learning and recognition. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/file/299a23a2291e2126b91d54f3601ec162-Paper.pdf.

[13] Seungyong Moon, Junyoung Yeom, Bumsoo Park, and Hyun Oh Song. Discovering hierarchical achievements in reinforcement learning via contrastive learning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 63674–63686. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/c919a2b5ec1de69f2629f9119676e336-Paper-Conference.pdf.

[14] Gabriel Poesia, David Broman, Nick Haber, and Noah D. Goodman. Learning formal mathematics from intrinsic motivation. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 43032–43057. Curran Associates, Inc., 2024. URL https://proceedings.

neurips.cc/paper_files/paper/2024/file/4b8001fc75f0532827472ea5a16af9ca-Paper-Conference.pdf.

[15] Guillaume Pourcel, Thomas Carta, Grgur Kovač, and Pierre-Yves Oudeyer. Autotelic LLM-based exploration for goal-conditioned RL. In *IMOL@NeurIPS 2024 - Intrinsically Motivated Open-ended Learning Workshop at NeurIPS 2024*, Vancouver, Canada, December 2024. URL https://inria.hal.science/hal-04861896.

[16] Doina Precup and Richard S. Sutton. *Temporal abstraction in reinforcement learning*. PhD thesis, University of Massachusetts, 2000. AAI9978540.

[17] Olivier Sigaud, Gianluca Baldassarre, Cedric Colas, Stephane Doncieux, Richard Duro, Pierre-Yves Oudeyer, Nicolas Perrin-Gilbert, and Vieri Giuliano Santucci. A definition of open-ended learning problems for goal-conditioned agents, 2024. URL https://arxiv.org/abs/2311.00344.

[18] Richard S. Sutton, Doina Precup, and Satinder Singh. Between mdps and semi-mdps: A framework for temporal abstraction in reinforcement learning. *Artif. Intell.*, 112 (1-2):181–211, 1999. URL http://dblp.uni-trier.de/db/journals/ai/ai112.html#SuttonPS99.

[19] Jonathan S Tsay, Hyosub E Kim, Samuel D McDougle, Jordan A Taylor, Adrian Haith, Guy Avraham, John W Krakauer, Anne GE Collins, and Richard B Ivry. Fundamental processes in sensorimotor learning: Reasoning, refinement, and retrieval. *eLife*, 13:e91839, aug 2024. ISSN 2050-084X. doi: 10.7554/eLife.91839. URL https://doi.org/10.7554/eLife.91839.

[20] Karthik Valmeekam, Sarath Sreedharan, Matthew Marquez, Alberto Olmo, and Subbarao Kambhampati. On the planning abilities of large language models (a critical investigation with a proposed benchmark), 2023. URL https://arxiv.org/abs/2302.06706.

[21] Alexander Sasha Vezhnevets, Simon Osindero, Tom Schaul, Nicolas Heess, Max Jaderberg, David Silver, and Koray Kavukcuoglu. Feudal networks for hierarchical reinforcement learning, 2017. URL https://arxiv.org/abs/1703.01161.

[22] Pablo Villalobos, Anson Ho, Jaime Sevilla, Tamay Besiroglu, Lennart Heim, and Marius Hobbhahn. Will we run out of data? limits of llm scaling based on human-generated data, 2024. URL https://arxiv.org/abs/2211.04325.

[23] Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models, 2023. URL https://arxiv.org/abs/2305.16291.

[24] Muning Wen, Ziyu Wan, Jun Wang, Weinan Zhang, and Ying Wen. Reinforcing llm agents via policy optimization with action decomposition. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 103774–103805. Curran Associates, Inc., 2024. URL https://proceedings.

neurips.cc/paper_files/paper/2024/file/bc09efb501c801ed92e181e26a885c2d-Paper-Conference.pdf.

[25] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models, 2022. URL https://arxiv.org/abs/2205.01917.

[26] Jenny Zhang, Joel Lehman, Kenneth Stanley, and Jeff Clune. Omni: Open-endedness via models of human notions of interestingness. In *The Twelfth International Conference on Learning Representations*, 2023.