Are Humans Evolved Instruction Followers? An Underlying Inductive Bias Enables Rapid Instructed Task Learning

Anjishnu Kumar
anjishnu.kr@gmail.com

Abstract

Human adults can often perform a novel task correctly on the first attempt after only receiving verbal or written instructions. This rapid instructed task learning (RITL) is a hallmark of human cognitive flexibility, yet its mechanisms and parallels in artificial systems remain under-explored across disciplines. In this position paper, we argue that humans possess an *evolved instruction-following bias*–an inductive bias shaped by evolution to interpret and execute linguistic instructions–which critically enables fast generalization of behavior from language. This bias functions analogously to the way large language models (LLMs) leverage instruction tuning to achieve zero-shot task performance. We synthesize evidence from cognitive science, neuroscience, and machine learning research to support this hypothesis. While instruction-following in AI is currently achieved via specialized training protocols, we posit that in humans it arises as an innate cognitive architecture feature. We outline testable predictions and call for more interdisciplinary research to investigate Instruction-Following as a unifying mechanism enabling rapid task learning in both natural and artificial neural networks.

1 Introduction

Humans uniquely excel at learning new tasks from instructions. With minimal or no direct experience, a person can often understand a set of written or spoken directions and immediately carry out a novel task correctly [Cole et al., 2013a]. For example, an adult might successfully operate an unfamiliar software application or play a new board game after reading a manual just once. In laboratory settings, cognitive scientists term this ability **Rapid Instructed Task Learning (RITL)**– "The ability to rapidly reconfigure our minds to perform new tasks based on instruction". RITL allows first-trial execution of novel rule sets without iterative practice, a capacity that is highly advantageous in daily life and distinguishes human learning from standard reinforcement learning (which typically requires repeated trial-and-error) [Sutton and Barto, 2018]. Indeed, non-human animals generally cannot do the same: they require numerous reinforced trials to acquire a new behavioral task, whereas humans often succeed by instruction on the first attempt [Asaad et al., 1998].

Cross-Disciplinary Gap. Research on instruction following is siloed. Cognitive and developmental scientists have studied how humans learn from pedagogy and language, while neuroscientists probe the brain networks enabling flexible task sets. In parallel, AI researchers have developed instruction-tuned language models that exhibit striking abilities to perform new tasks described in natural language [Ouyang et al., 2022]. However, these communities use different terminologies and frameworks, and a unifying perspective is lacking. Cognitive neuroscience studies the topic mainly tangentially, whereas AI views instruction-following largely as merely a useful model training step. We identify a critical gap in existing theoretical work: is there an underlying principle of instruction-based generalization common to both human brains and artificial models? And if so, have humans effectively been "instruction-tuned" by evolution?

Central Hypothesis. We propose that humans possess an evolved instruction-following bias that enables rapid generalization from linguistic instructions, analogous to how instruction tuning endows LLMs with broad task flexibility. In other words, the human mind may come prepared with inductive biases and neural machinery specialized for processing inputs as instructions in a way that is different to other species. We argue that this instruction-following bias in humans is likely an evolutionary adaptation–forged by the advantages of cultural transmission and social learning–and is a key enabler of humans' unparalleled capacity for fast instructed learning [Csibra and Gergely, 2011, Rendell et al., 2010]. Understanding this bias can inform the design of AI systems, and conversely, insights from instruction-tuned models can shed light on cognitive and neural mechanisms of RITL. We organize the paper as follows: (i) review relevant work from cognitive science, neuroscience, and machine learning on instruction following and rapid task generalization; (ii) articulate the proposed analogy between human and AI instruction-following, highlighting evolution in humans versus explicit training in machines; (iii) compare and contrast with animal studies and (iv) discuss testable implications for both neuroscience and AI.

2 Background and Related Work

2.1 Human Cognitive Capacity for Instruction Following

Work on human development suggests that infants are born students, equipped with biases to learn from communicative cues i.e. "natural pedagogy". According to this theory, infants innately recognize signals such as infant-directed speech as indicators that they are being taught, and they preferentially process information presented in this pedagogical context, enabling fast social learning of cognitively opaque cultural knowledge. Human children appear predisposed to follow instructions and demonstrations, allowing rapid uptake of knowledge that ensures cultural transmission. In adults, the pinnacle of instruction-following ability is captured by RITL: learning a novel task rule or procedure from an instruction and executing it correctly on the first try [Cole et al., 2013a]. Everyday examples include learning to use a new software after reading a guide, or mentally running through steps of a new recipe from a list. Experimental psychology demonstrates that humans can achieve high first-trial performance on arbitrary tasks when those tasks are described in advance by instructions–for example, implementing arbitrary mappings between novel stimuli and responses immediately and with high accuracy [Meiran et al., 2012, Wolfensteller and Ruge, 2012]. Such behavior would be impossible without instructions–imagine learning the rules of chess by pure trial-and-error versus having them explained verbally. RITL dramatically reduces the search space of learning by directly imposing a task set via language.

Comparative evidence underscores that RITL is a rare trait. Some primates can follow extremely simple instructions. Macaques, for example, often require many practice trials to learn even simple stimulus-response associations [Asaad et al., 1998]. In contrast, human participants facing similar novel mapping tasks can be highly accurate on the very first trial [Cole et al., 2013a]. Our closest ape relatives show a bit more prowess–the bonobo Kanzi could execute dozens of simple English commands (e.g., "Put your ball on the pine needles") with substantial accuracy [Savage-Rumbaugh et al., 1993]. But even Kanzi's performance is roughly equivalent to that of a human toddler. Classic experiments by Horner and Whiten [2005] showed that human children faithfully reproduced arbitrary actions they saw an adult perform while completing a simple task, in contrast, chimpanzees in the same setup skipped the irrelevant steps and went straight for the goal-efficient actions. This difference is profound: it suggests humans are intrinsically wired to follow instructions/observations from others with high fidelity, whereas apes emulate the goal but filter out what seems unnecessary. Follow-up studies have found overimitation to be robust across cultures, with verbal instructions enhancing the effect. [Papa et al., 2021]. The benefits of such a capacity are evident: being able to share new skills or behaviors through communication confers adaptive value. Computational tournaments of social learning strategies show how instruction-like social learning enhances group success and the spread of knowledge [Rendell et al., 2010]. This likely created selective pressure for an instruction-following bias in the human lineage [Csibra and Gergely, 2011].

Dogs as heterospecific instruction followers. Domestic dogs provide a natural experiment on whether an instruction-following bias (IF-bias) can be shaped by selection. In imperative object–choice tasks, dogs reliably use human pointing and gaze to select the requested item, whereas chimpanzees often perform at chance without extensive training [Hare et al., 2002, Kirchhofer et al.,

2012]. Dogs pups are able to outperform wolf pups of similar age on similar tasks [Virányi et al., 2008]. Taken together, the canids illustrate that selective pressure (human domestication and social hierarchy) can create a robust IF-bias for human signals — without the broad, compositional, language-mediated task recomposition that characterizes human RITL.

2.2 Neural Mechanisms: Task-Set Reconfiguration and Cognitive Control

What happens in the brain when we convert an instruction into action? Cognitive neuroscience points to the frontoparietal cognitive control network–especially lateral prefrontal cortex (LPFC)–as key for implementing novel task sets on the fly [Duncan, 2010]. Neuroimaging of RITL finds that when people encode new instructions, dorsolateral/anterior PFC represents the intended task even before any practice, and then interacts with parietal and task-relevant areas to guide execution [Cole et al., 2011, 2013b]. This aligns with task-set reconfiguration: when faced with a new instructed task, the brain rapidly configures a network that implements the mapping from stimuli to responses as specified by the rule; the pattern of frontoparietal connectivity shifts flexibly with current demands and can predict which task a person is performing, consistently across novel and practiced tasks [Cole et al., 2013b]. These "flexible hubs" are thought to underwrite generalized cognitive control.

There are indications that humans' proficiency at instruction-based task-set formation has distinct neural underpinnings compared to other species. Anatomically, anterior prefrontal cortex (approx. Brodmann area 10) is disproportionately enlarged in humans relative to other primates, and among non-human apes the bonobo shows a relatively large area 10 [Semendeferi et al., 2001]. This suggests evolutionary expansion of prefrontal circuitry supporting abstract, instruction-derived rules. Consistently, classic theory casts PFC as an instruction parser/buffer transforming linguistic/symbolic instruction into an executable task set via top-down control [Miller and Cohen, 2001]. Overall, the evidence suggests a general-purpose task engine configurable by instruction.

2.3 Instruction Tuning in AI and Zero-Shot Task Generalization

Mirroring the human ability to generalize, modern large language models have recently demonstrated remarkable zero-shot and few-shot capabilities. Base pretrained models (e.g., GPT-3) can solve new tasks given a few examples or prompts, but unable to follow arbitrary user instructions out-of-the-box [Brown et al., 2020].

To bridge this, instruction tuning fine-tunes models on many natural-language (instruction, response) pairs spanning diverse tasks (translation, QA, arithmetic, etc.), explicitly teaching models to follow instructions. Through supervised learning (often with alignment via human feedback), parameters are adjusted so the model interprets the instruction and generates an appropriate completion [Ouyang et al., 2022]. The outcome is strong zero-shot performance on unseen tasks–sometimes matching or exceeding larger non-tuned models or few-shot baselines simply by understanding the instruction [Wei et al., 2022, Chung et al., 2022]. In effect, instruction tuning confers task flexibility akin to human RITL. These developments motivate viewing instruction tuning as a machine analog of human instruction following.

3 Evolved Instruction-Following Bias: Linking Human and Machine

We hypothesize that the human brain comes equipped with a specialized inductive bias for interpreting and applying instructions, selected through the evolution of communication and culture. By default, when a human hears or reads a directive (especially in a pedagogical context), the cognitive system attempts to integrate that information as a guide for behavior–an innate expectation that language conveys actionable, generalized knowledge [Csibra and Gergely, 2011]. This is analogous to how an instruction-tuned LLM exhibits an internal policy (learned via fine-tuning) to treat inputs as instructions and produce appropriate outputs [Sanh et al., 2022, Ouyang et al., 2022]. For the LLM, the skill is externally added by additional training; for humans, we propose it is deeply ingrained, a result of our neurological "hardware" and inherited cognitive architecture. From an evolutionary perspective, individuals who could rapidly share survival skills or tool-use techniques via language would have an advantage [Rendell et al., 2010]. Over generations, this could favor brains adept at rapidly mapping linguistic input to goal-directed actions. Natural pedagogy in infants and the jump in instruction-following capacities from monkeys to apes to humans both point to evolutionary

tuning toward instruction-based learning. In short, what instruction tuning is to GPT-3, evolution (and cultural evolution) is to *Homo sapiens*: a process endowing the system with a general skill of following instructions.

This perspective reframes instruction following as a core computational strategy underlying human cognitive flexibility. It suggests that the brain's default prior (especially in cooperative social contexts) is that new behaviors can be acquired by listening and understanding, not solely by direct experience. Just as an instruction-tuned model combines pretraining knowledge with the prompt to produce the solution, a human combines prior world knowledge with the content of instructions to synthesize a plan for a novel task. The hypothesis further implies that certain neural mechanisms may be dedicated to, or disproportionately used for, parsing and implementing instructions (e.g., specialized representations of instructions in PFC) [Miller and Cohen, 2001, Cole et al., 2011]. Recent work by [Goldstein et al., 2024] shows alignment between language-model embeddings and human neural representations during language processing, and work by [Aw et al., 2023] demonstates that instruction-tuning of LLMs improves alignment between language-model and brain representations.

4 Implications and Future Directions

If humans indeed have an evolved instruction-following bias analogous to AI instruction tuning, several testable predictions and research directions emerge:

Neural Representation of Instructions. The hypothesis predicts that the brain maintains explicit representations of intended task instructions, likely in frontal cortex, detectable during encoding and predictive of later performance. During the instruction-encoding phase (before execution), one should observe stable LPFC patterns corresponding to the rule, which then configure broader task networks. Identifying alignments between instruction semantics and neural state geometry would support a dedicated instruction-processing mechanism [Goldstein et al., 2024].

Flexible-Hub Dynamics. If instruction following relies on flexible hubs in frontoparietal networks, perturbations should selectively impair instruction-based setup while sparing well-learned tasks. For instance, transient disruption of LPFC during instruction presentation should reduce the ability to configure new task sets, paralleling classic evidence that LPFC disruption impairs top-down control [Miller and Cohen, 2001].

Comparative and Developmental Studies. The bias should be present early in development and rudimentary or absent in non-human animals. Experiments can test toddlers' one-shot instruction following in novel tasks, tracking growth in parallel with language ability, and compare with primates using simplified gestural/symbolic instructions.

Parallels in AI Behavior. Instruction-tuned models offer a tool to simulate and predict human instruction following. By presenting matched novel tasks to humans and tuned LLMs; comparing divergences can reveal where human inductive priors differ from current models.

5 Conclusion

We argue that RITL should be regarded as a foundational aspect of human intelligence, fostered by evolution and cultural transmission. This instruction-following bias in humans plays a role analogous to instruction tuning in artificial language models. Recognizing this analogy provides a conceptual bridge between cognitive neuroscience and AI. For AI researchers, it has been key to creating flexible, human-like learners. For neuroscientists, it highlights instruction-following as a distinct cognitive function worthy of further focused investigation, with its own neural signatures and evolutionary history. Interdisciplinary collaboration leading better understanding instruction-following may prove key to a unifying theory for rapid generalization.

References

Wael F. Asaad, Gregor Rainer, and Earl K. Miller. Neural activity in the primate prefrontal cortex during associative learning. *Neuron*, 21(6):1399–1407, 1998. doi: 10.1016/S0896-6273(00)80658-3.

Khai Loong Aw, Syrielle Montariol, Badr AlKhamissi, Martin Schrimpf, and Antoine Bosselut. Instruction-tuning aligns llms to the human brain. *arXiv preprint arXiv:2312.00575*, 2023.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. URL https://arxiv.org/abs/2005.14165.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022. URL https://arxiv.org/abs/2210.11416.

Michael W. Cole, Joset A. Etzel, Jessica M. Zacks, Walter Schneider, and Todd S. Braver. Rapid transfer of abstract rules to novel contexts in human lateral prefrontal cortex. *Frontiers in Human Neuroscience*, 5:142, 2011. doi: 10.3389/fnhum.2011.00142. URL https://www.frontiersin.org/articles/10.3389/fnhum.2011.00142/full.

Michael W. Cole, Patryk Laurent, and Andrea Stocco. Rapid instructed task learning: A new window into the human brain's unique capacity for cognitive control. *Frontiers in Human Neuroscience*, 7:207, 2013a. doi: 10.3389/fnhum.2013.00207. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC3557598/.

Michael W. Cole, John R. Reynolds, Jonathan D. Power, Grega Repovš, Alan Anticevic, and Todd S. Braver. Multi-task connectivity reveals flexible hubs for adaptive task control. *Proceedings of the National Academy of Sciences*, 110(28):10061–10066, 2013b. doi: 10.1073/pnas.1302916110.

Gergely Csibra and György Gergely. Natural pedagogy as evolutionary adaptation. *Philosophical Transactions of the Royal Society B*, 366(1567):1149–1157, 2011. doi: 10.1098/rstb.2010.0319. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC3049090/.

John Duncan. The multiple-demand (md) system of the primate brain: mental programs for intelligent behaviour. *Trends in Cognitive Sciences*, 14(4):172–179, 2010. doi: 10.1016/j.tics.2010.01.004.

Ariel Goldstein, Avigail Grinstein-Dabush, Mariano Schain, Haocheng Wang, Zhuoqiao Hong, Bobbi Aubrey, Samuel A. Nastase, Zaid Zada, Eric Ham, Amir Feder, Harshvardhan Gazula, Eliav Buchnik, et al. Alignment of brain embeddings and artificial contextual embeddings in natural language points to common geometric patterns. *Nature Communications*, 15(1), 2024. doi: 10.1038/s41467-024-46631-y. URL https://ideas.repec.org/a/nat/natcom/v15y2024i1d10.1038_s41467-024-46631-y.html.

B. Hare, M. Brown, C. Williamson, and M. Tomasello. The domestication of social cognition in dogs. *Science*, 298(5598):1636–1639, 2002. doi: 10.1126/science.1072702.

Victoria Horner and Andrew Whiten. Causal knowledge and imitation/emulation switching in chimpanzees (pan troglodytes) and children (homo sapiens). *Animal cognition*, 8(3):164–181, 2005.

K. C. Kirchhofer, F. Zimmermann, J. Kaminski, and M. Tomasello. Dogs (canis familiaris), but not chimpanzees (pan troglodytes), understand imperative pointing. *PLOS ONE*, 7(2):e30913, 2012. doi: 10.1371/journal.pone.0030913.

Nachshon Meiran, Maayan Pereg, Yoav Kessler, Michael W. Cole, and Todd S. Braver. When planning results in loss of control: intention-based reflexivity and working memory. *Frontiers in Human Neuroscience*, 6:295, 2012. doi: 10.3389/fnhum.2012.00295.

Earl K. Miller and Jonathan D. Cohen. An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience*, 24:167–202, 2001. doi: 10.1146/annurev.neuro.24.1.167.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. URL `https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf`.

Aliki Papa, Mioara Cristea, Nicola McGuigan, and Monica Tamariz. Effects of verbal instruction vs. modelling on imitation and overimitation. *Humanities and Social Sciences Communications*, 8(1):1–12, 2021.

Luke Rendell, Robert Boyd, Daniel Cownden, Magnus Enquist, Kimmo Eriksson, Marcus W. Feldman, Laurel Fogarty, Stefano Ghirlanda, Timothy Lillicrap, and Kevin N. Laland. Why copy others? insights from the social learning strategies tournament. *Science*, 328(5975):208–213, 2010. doi: 10.1126/science.1184719. URL `https://pmc.ncbi.nlm.nih.gov/articles/PMC2989663/`.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. Multitask prompted training enables zero-shot task generalization. In *International Conference on Learning Representations (ICLR)*, 2022. URL `https://arxiv.org/abs/2110.08207`.

E. Sue Savage-Rumbaugh, Jeannine Murphy, Rose A. Sevcik, Karen E. Brakke, Shelly L. Williams, and Duane M. Rumbaugh. *Language Comprehension in Ape and Child*, volume 58 of *Monographs of the Society for Research in Child Development*. Wiley, 1993. URL `https://www.bonobohope.org/pdf/1993-Language-Comprehension-Ape.pdf`.

Katerina Semendeferi, Estel Armstrong, Axel Schleicher, Karl Zilles, and Gary W. Van Hoesen. Prefrontal cortex in humans and apes: A comparative study of area 10. *American Journal of Physical Anthropology*, 114(3):224–241, 2001. doi: 10.1002/1096-8644(200103)114:3<224::AID-AJPA1022>3.0.CO;2-I. URL `https://cs.brown.edu/people/tdean/projects/cortex/course/suggested_reading_list/supplements/documents/SemendeferietalAJPA-01.pdf`.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction (2nd ed.)*. MIT Press, 2018. URL `http://incompleteideas.net/book/the-book-2nd.html`.

Zsófia Virányi, Márta Gácsi, Enikő Kubinyi, József Topál, Beatrix Belényi, Dorottya Ujfalussy, and Ádám Miklósi. Comprehension of human pointing gestures in young human-reared wolves (canis lupus) and dogs (canis familiaris). *Animal cognition*, 11(3):373–387, 2008.

Jason Wei, Maarten Bosma, Vincent Zhao, et al. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations (ICLR)*, 2022. URL `https://arxiv.org/abs/2109.01652`.

Uta Wolfensteller and Hannes Ruge. Frontostriatal mechanisms in instruction-based learning. *Frontiers in Psychology*, 3:192, 2012. doi: 10.3389/fpsyg.2012.00192. URL `https://www.frontiersin.org/articles/10.3389/fpsyg.2012.00192/full`.