Response to Reviewer #1

We thank you for your constructive feedback, and pointing out several converging parallels between our work and others in NLP. We have added a discussion in our manuscript. Thank you also for your nuanced observation on using a paired t-test instead of a stricter 2 samples t-test, since we compare within a subject. **We have stronger results to report**—we now observe that *dynamic analysis*, one of the four key hand-selected code properties we evaluated (static, dynamic, control-flow, data type), is more accurately decoded by MD than the language system, and is decoded more accurately than the *static analysis* property by the MD. We discuss the implication of these results below, where we address your question on the motivation of our work.

Motivation. • Our work aims to examine whether the information encoded in ML models trained to understand code is similar to what we humans encode when we comprehend code; all ML communities modeling cognitive tasks like language, vision, *etc.* care about aligning their work with how the brain works.

• Further, it is unclear what mechanisms drive our cognition which enable us to comprehend code. If we find one class of ML models (say, masked-LMs) to be more predictive than another (say, autoencoders), then it is reasonable to suspect that our brains optimize objectives more similar to that of masked-LMs than that of autoencoders for comprehending code. *In essence, we show how ML models can serve as tools to reverse engineer our cognitive processes.* See also Caucheteux et al. at ICML 2021 [1], who adopt a similar paradigm to learn more about the language system.

• As a corollary, a poor correspondence between the information encoded by our brains and ML models suggests the possibility of unexplored neural architectures and objectives which may better model our cognition, which in turn may outperform extant ML models. In light of our revised results, we present evidence to reconsider the design of current neural architectures for code understanding, most of which do not accommodate dynamic (runtime) information, one of the core features encoded by the MD system. Our results suggest code model performance might benefit from mimicking both the MD and language systems more explicitly. We will extend our introduction section to include this discussion.

Other questions. Thank you for your note on *variable names*–we agree; *baselines*–the figures indeed showed a common theoretical baseline. We now present each empirical baseline; *auditory cortex*–yes, code vs. sentence in auditory is likely related to silent reading during sentence comprehension (Perrone-Bertolotti et al., 2012); *statistical testing*–we provide some details on testing and FDR in Appendix B currently, but have expanded these explanations, and have more explicitly added them to the methods section and figure captions. We omit reproducing those details here due to space constraints. *partial regressions*–Visual+MD vs.Visual can be compared directly as each presents with an equivalent null distribution, and is a standard ablation test to determine feature importance. We will highlight all these details in our draft.

Response to Reviewer #2

Thank you for your thoughtful feedback. We will ensure to find the right balance in presenting information in the main draft and the appendix–the current balance is a result of feedback we received from multiple proof-readers with different backgrounds.

Why code comprehension. We focus on code comprehension because very little of this important skill has been analyzed from a cognitive neuroscience perspective while steady advances are being made in training ML models to understand code and increase programmer productivity. Unlike in vision, ML models for understanding programming are direct adoptions of the state-of-the-art in language research. However, recent works we document have shown that comprehending programs does not share the same neural bases as natural language comprehension. Do code models then mimic human cognition of programs? Do language models mimic? If not, can we think of other objectives which are directly inspired by results from neuroscience? These are the open questions we want the community to address. Our results highlight the importance of modeling the MD system, which among other things significantly correlates to dynamic (runtime) information. We thus present evidence to rethink the design and optimization of current neural code models.

Other questions. *Participants*–yes, they all had ∼3 years of Python experience; *brain systems*–yes, the general class of stimuli these 4 systems respond to, and the ways to locate each of these system in any individual are well established. However, their roles in specific cognitive skills are actively studied, and our results refine our understanding of the representational content of each system. We use the auditory and visual systems as baselines–they account for noise fluctuations in a neural signal, and low-level stimulus properties like shapes, colors, *etc.* respectively; *interpretability*–the large voxel feature space is best interpreted with linear models, as we employ here, and which is typical in the probing literature; *ROIs*–Grouping by brain systems, which we reuse from Ivanova *et al.*, is the best granularity for our questions. Evaluating finer-grained ROIs is possible, but these subsets are highly correlated, and their specific differences relative to their composite systems haven't been consistently mapped to interpretable functions or mechanisms.

Response to Reviewer #4

Note: We were notified of a major update made by Reviewer #4 with less than 24 hours until the rebuttal dead-

[1] https://icml.cc/virtual/2021/spotlight/9272

line. We are hence going over the one-page limit to respond to Reviewer #4; we did not have much to respond to their previous set of terse comments.

Thank you for the comments.

• We do not understand the reviewer's claim of our work or related works being *controversial*. The claim is misplaced and the implications of reading code sequentially are irrelevant to ours and the work by Ivanova *et. al.* (and by extension, Liu *et. al.*). Peer reviews are publicly available on eLife - none of them refers to such 'controversy'.

• Related, the reviewer claims that our work "uses a majority of recycled references from that (Ivanova *et. al.*) paper." There are >200 citations between the two papers, and only 14 are shared. This reinforces how different our questions are from what Ivanova *et. al.* analyze.

• The reviewer makes multiple references to *notebook style* writing. It would be helpful if they could articulate what in the writing style they do not find appealing, and concrete instances of what they think is *notebook style*. We are unaware of this style and seek to learn and rectify.

• Thanks for referring to Marek *et. al.*, 2022; we are well aware of this work. They discuss predicting phenotype behavior using fMRI signals alone. **That is neither the outcome nor the goal of our work–we do not predict one's *ability* to comprehend computer programs or other such psychometric evaluations by looking at their fMRI responses.** In fact, the section "*Importance of small-sample neuroimaging*", pp. 658 in Marek *et. al.* strongly supports our setup and the use of fMRI–which is to isolate anatomical regions of the brain involved in any cognitive process, and subsequently find their correspondence to ML models. See also Caucheteux et al., published at ICML 2021 [1], who adopt a similar paradigm to learn more about the language system specifically. ECoG and EEG are indeed incredibly useful modalities for studies that necessitate high temporal resolution, and in the case of ECoG, broad coverage of a small region of cortex, but cannot address the questions we ask here, namely the information encoded in large-scale spatially-defined brain systems.

• The reviewer mentions "finding above chance accuracies in a single subject". Again, we are not classifying between individuals, but are evaluating each individual separately and seeing an effect across the entire sample of 24 individuals. The fact that every individual scans code "with different visuospatial strategies" makes our results more robust not less–we see a significant effect across our entire sample of 24 different participants, each with allegedly unrelated strategies.

We encourage the reviewer to re-examine our submission in light of our clarifications.