
Supplementary Material of Failure by Interference: Language Models Make Balanced Parentheses Errors When Faulty Mechanisms *Overshadow* Sound Ones

Anonymous Author(s)

Affiliation

Address

email

1 **1. Open access to data and code**

2 Please refer to the *code-and-data* folder of supplementary files for the dataset and code used
3 for RASTEER and interpretability analysis in Section 3. In the future, we plan to release
4 them to the public as well.

5 **2. Broder Impact**

6 This paper aims to understand the inner workings of language models (LMs) by examining
7 how they perform the balanced parentheses task and why they sometimes fail. We view
8 this as foundational interpretability research, with its impact expected to emerge through
9 future applications of interpretability, such as diagnosing unexpected model behaviors and
10 enabling more effective control and steering of LMs to better serve user needs. As automatic
11 code generation has become one of the core applications of LMs, we expect our approach,
12 RASTEER, will benefit practitioners such as software engineers who may use LMs along
13 with our approach for more reliable code generation. We do not perceive any potential
14 negative impact from this work.