EnSToM: 利用放引向量增強持的系

Heejae Suh[1], Yejin Jeon[1], Deokhyung Kang[1], Taehee Park[1], Yejin Min[1], Gary Geunbae Lee[1,2]
[1]POSTECH人工智能究生院,
[2]POSTECH算机科工程系,
{heejaesuh, jeonyj0612, deokhk, taehpark, yeajinmin, gblee}@postech.ac.kr

摘要

小型大言模型 (sLLMs) 具有量高效的,使其适用于源受限的境 然而,sLLMs在任向系中常常以保持一致性,于像服聊天机器人的景至重要 具而言,保模型拒或意入,遵循其期功能,以防止在的用持可性是十分重要的 此,已有的激活工程方法被提出,用于在推理程中操控部激活 。管些方法在某些景下有效,但我的初步揭示了在保一致性方面的局限性 因此,了解一,我提出了一新方法,**En**tropy-scaled **S**teering vectors for **To**pic **M**aintenance (EnSToM) EnSToM根据入的不定性整引強度,而有效理干,同保持的准性 我的表明,微方法相比,EnSToM在相小的据集上了著的性能提升 通在不影效率的情下提高一致性,我的方法增強基于sLLM的系提供了一健的解方案[1] 。

[1]源代可在 https://github.com/linkyouhj/enstom 得