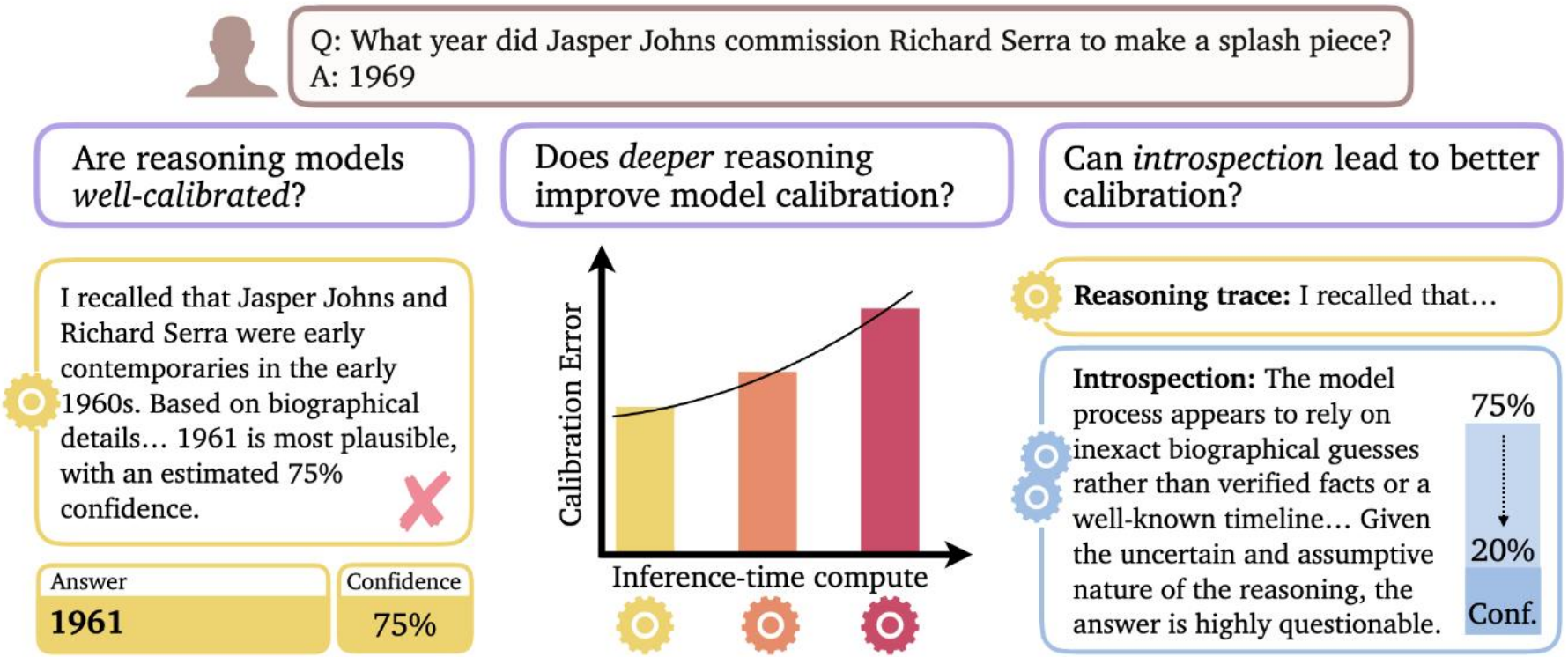
Reasoning about Uncertainty: Do Reasoning Models Know When They Don't Know?

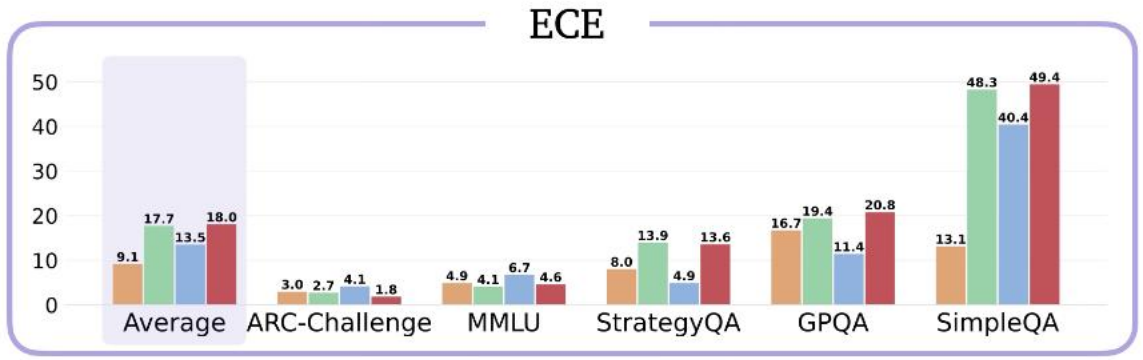
Princeton University

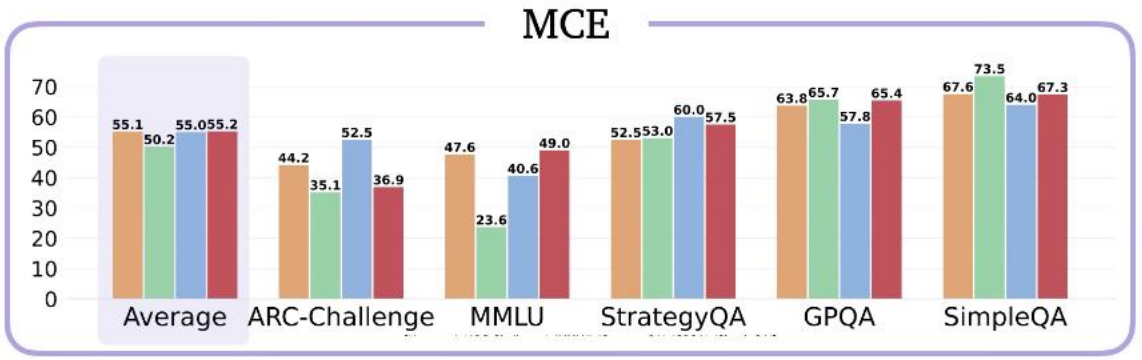
三大问题

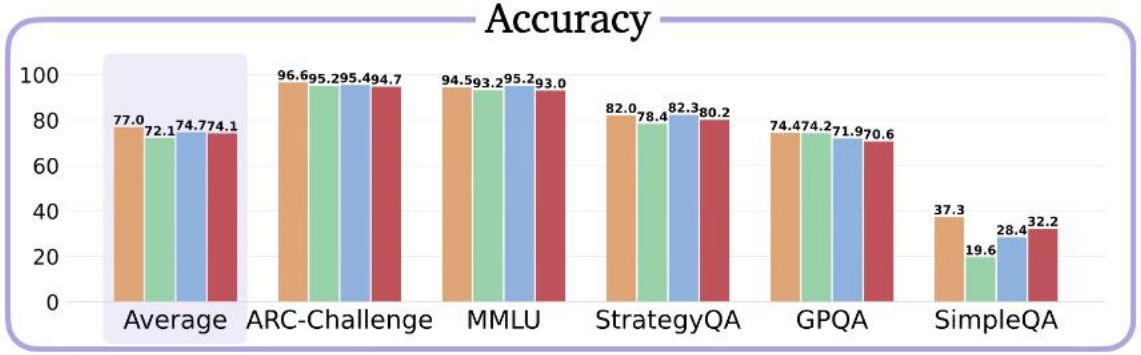


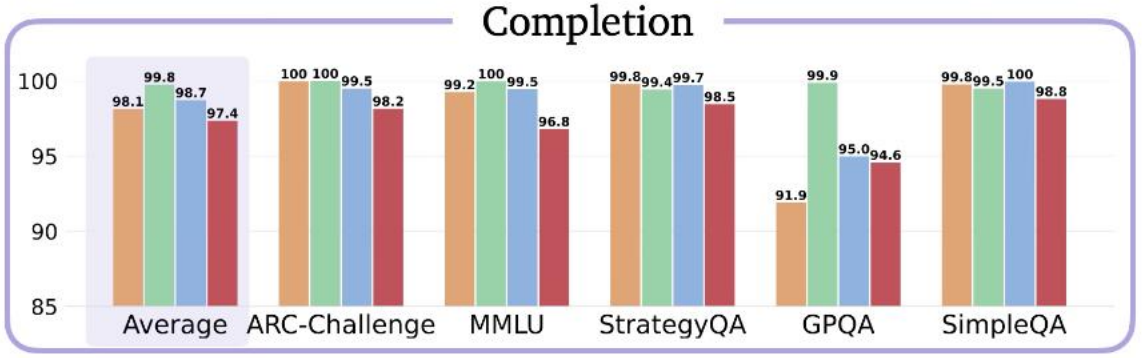
推理模型的校准程度如何　　深度推理可以提升校准率吗　　反思可以提升校准率吗

推理模型的校准程度







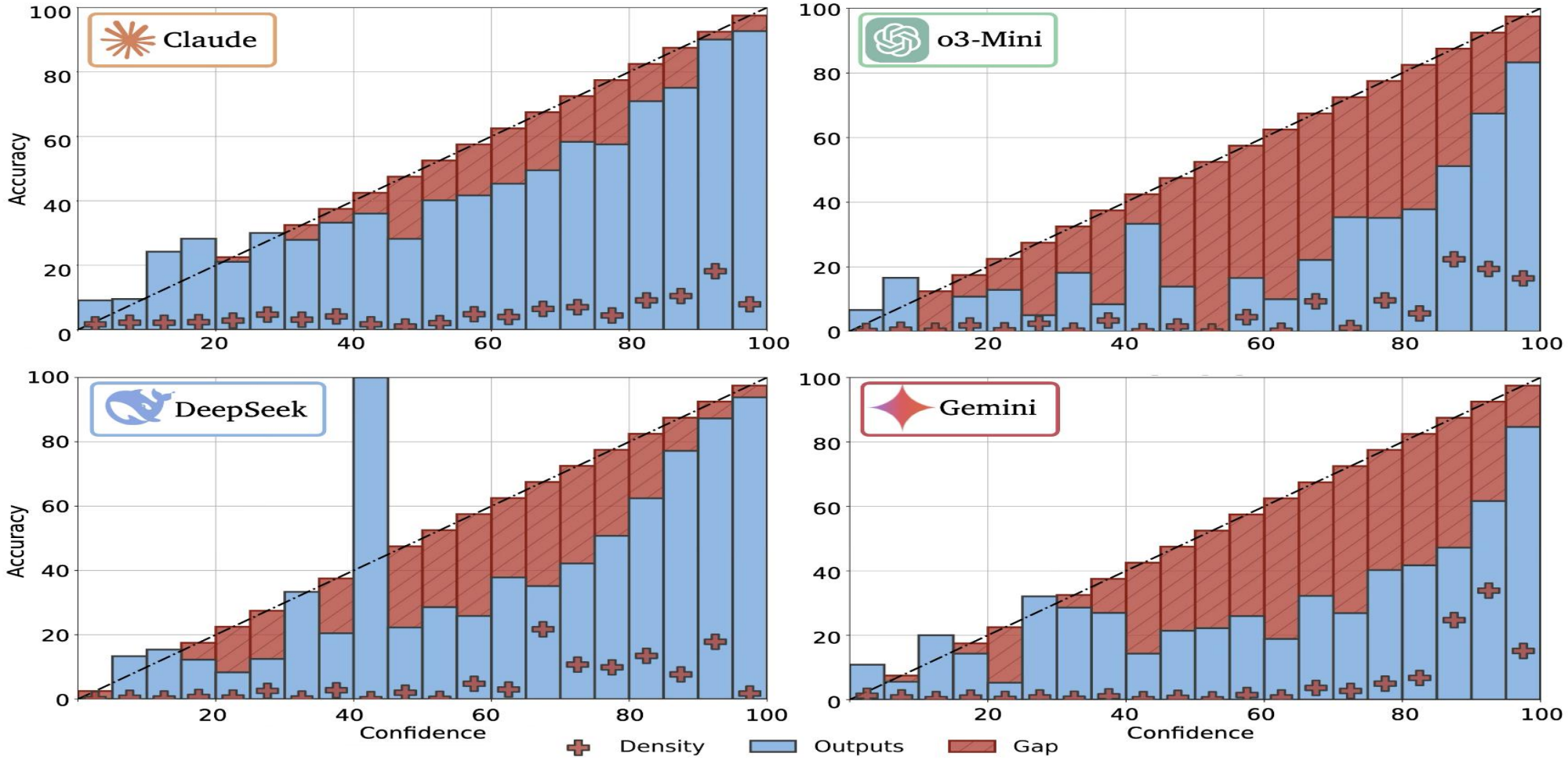


Claude o3-Mini DeepSeek Gemini

1．推理模型在ARC—Challenge和MMLU基准测试中展现出更优的校准性；但在StrategyQA、GPQA和SimpleQA等基准上，这些模型的校准性能出现显著下降

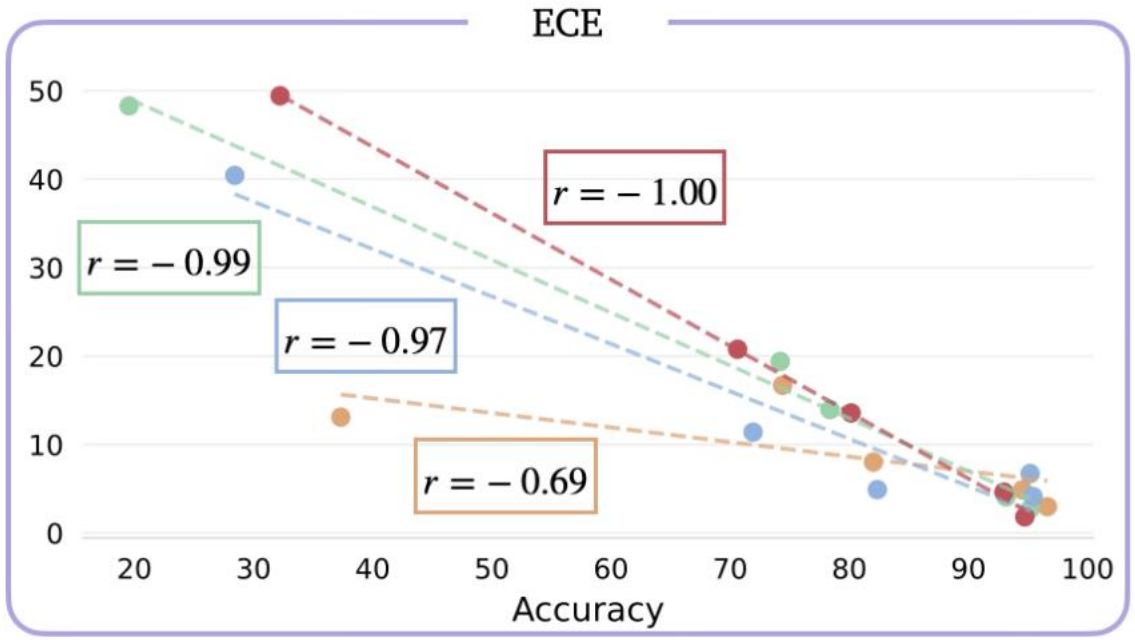
2．模型在特定数据集上的校准表现并不总是能推广至其他数据集，尤其当模型准确率接近完美时，过拟合自信度难以被察觉

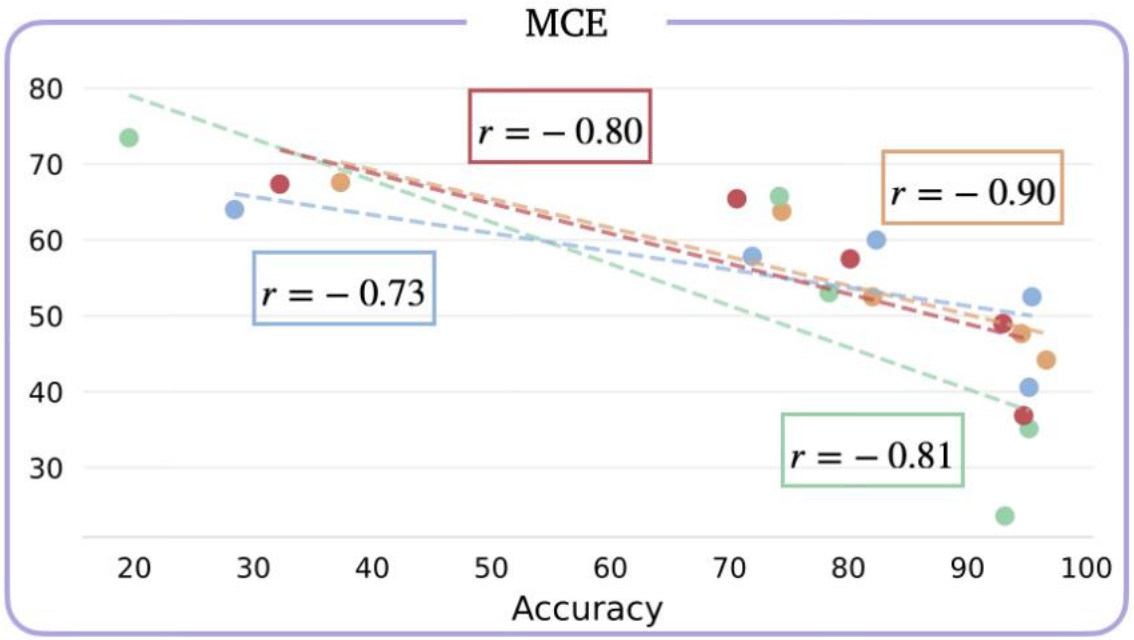
推理模型的校准程度



1. 推理模型普遍存在过度自信现象，其置信度估计通常高于85%
2. Claude是所有模型中校准最优的，其校准误差明显较小

推理模型的校准程度

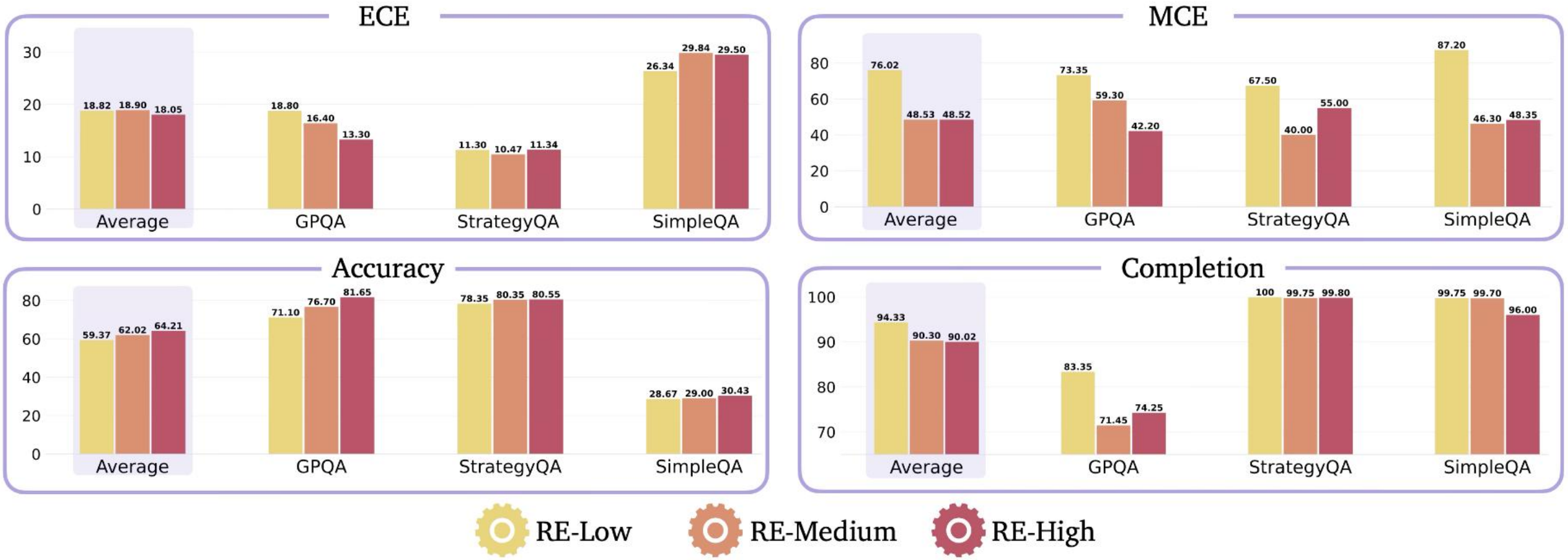




Claude o3-Mini DeepSeek Gemini

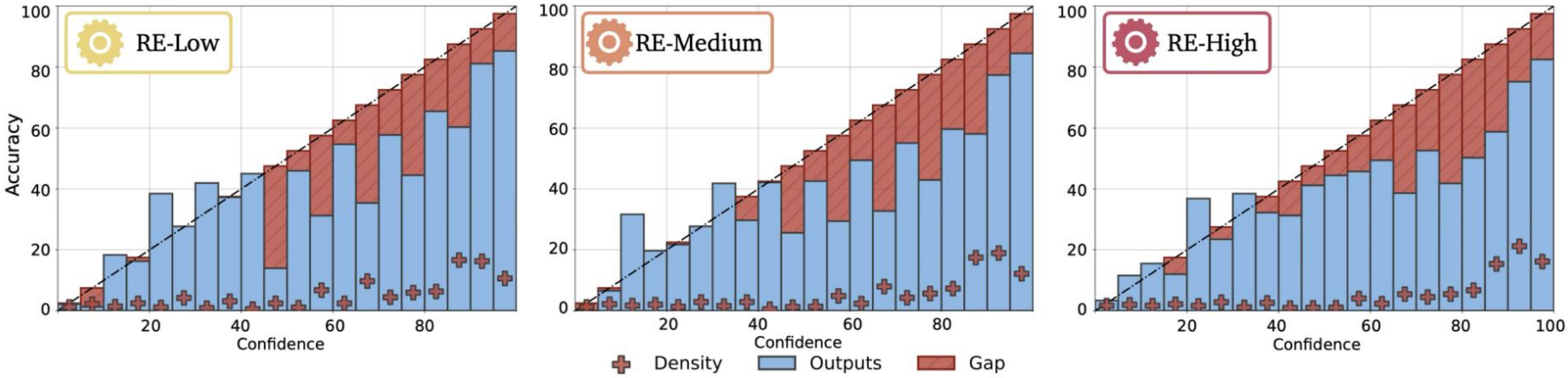
1. 所有模型的准确率与ECE均呈现强负相关性，佐证了推理模型普遍存在校准不足的现象
2. 理想校准模型的置信度应反映其准确率，表现为准确率与ECE间接近零的相关系数

深度推理提升校准率？



1. 更深的推理平均会带来更高的准确率和更好的校准度
2. 低难度数据集（StrategyQA）中，增加推理深度基本不会改变校准度和准确率
3. 高难度数据集（ SimpleQA ）中，当模型准确率达饱和状态后，继续加深推理反而会增大校准误差

深度推理提升校准率？



1. 随着推理深度增加，高置信度响应的比例持续上升，但多数情况下准确率并未同步提升，导致校准度恶化
2. 深度推理会加剧模型过度自信

深度推理提升校准率？

1．更深层推理会通过加强错误思维路径，使模型对其错误回答更加自信

反思可以提升基准率？

Introspective UQ-Low

You are provided with the reasoning trace of a model asked to answer a question and provide the associated confidence between 0 and 100. Your task is to think about the reasoning trace from the first model and provide your confidence in the correctness of the answer provided by the first model.

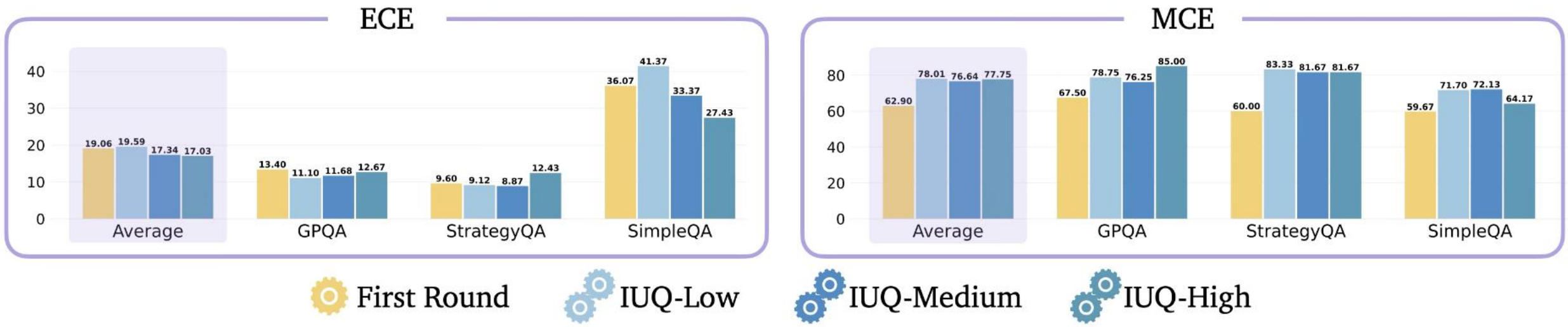
Introspective UQ-Medium

You are provided with the reasoning trace of a model asked to answer a question and provide the associated confidence between 0 and 100. Your task is to identify the flaws in the reasoning trace from the first model and provide your confidence in the correctness of the answer provided by the first model.

Introspective UQ-High

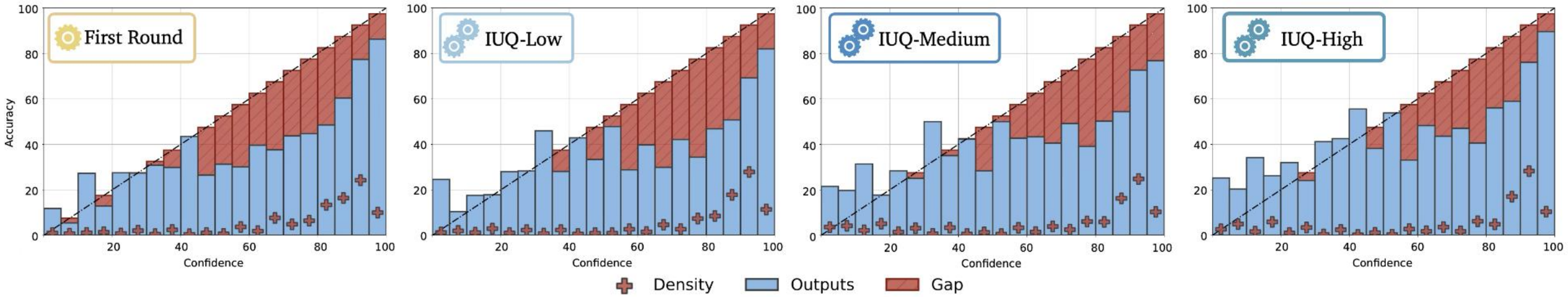
You are provided with the reasoning trace of a model asked to answer a question. Your task is to identify the flaws in the reasoning trace from the first model and provide your confidence in the correctness of the answer provided by the first model.

反思可以提升基准率？



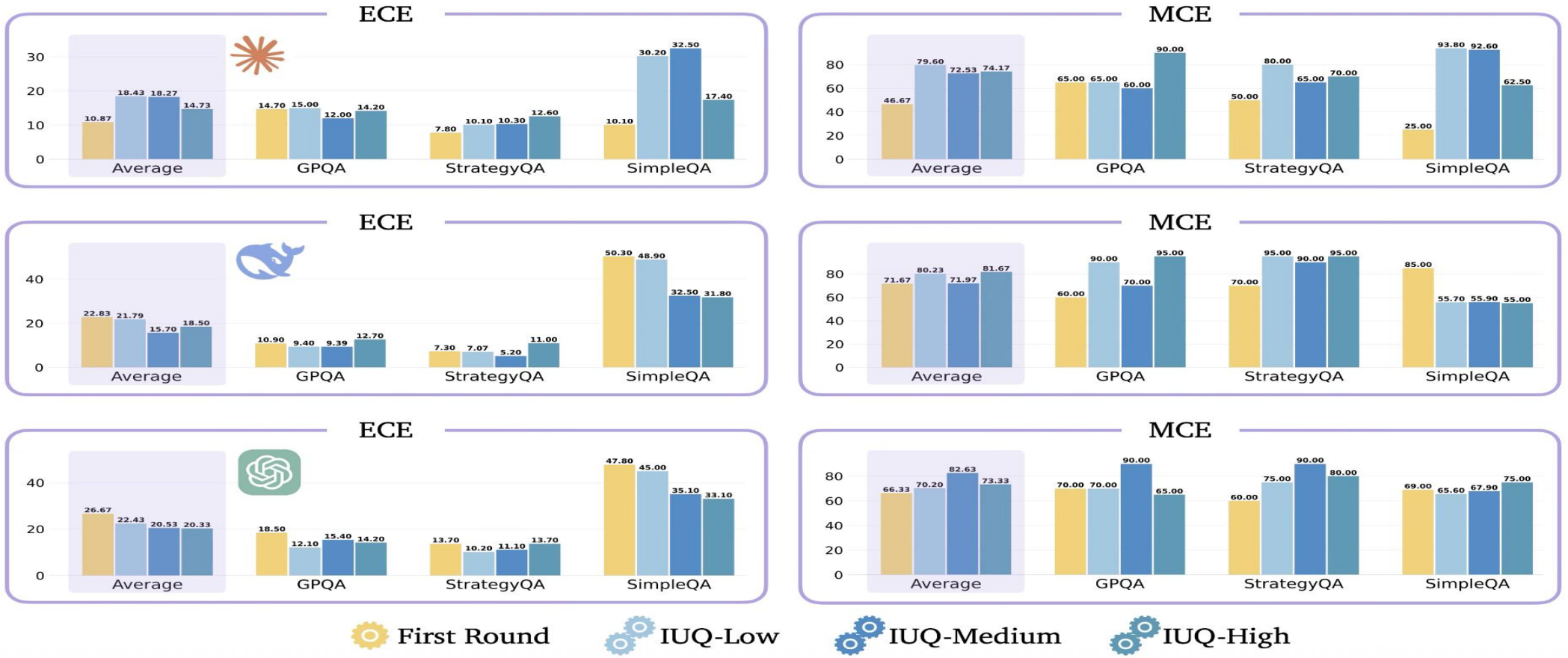
1. IUQ-Low的平均ECE略有上升，IUQ-Medium和IUQ-High中有所下降。
2. 在最具挑战的SimpleQA数据集中，IUQ-High显著减少了校准错误。
3. 在较易数据集StrategyQA中，IUQ-High可能轻微增加ECE

反思可以提升基准率？



1. 当内省批判性不足时（如IUQ-Low），模型的过度自信会加剧
2. 更具批判性的内省（如IUQ-Medium和IUQ-High）则能够改善校准效果。

反思可以提升基准率？



1．我们发现当Claude模型在StrategyQA和SimpleQA数据集中进行不确定性推理时，其校准质量显著恶化

2．相比之下，内省机制有效改善了DeepSeek和o3—Mini的校准效果

Thank you !