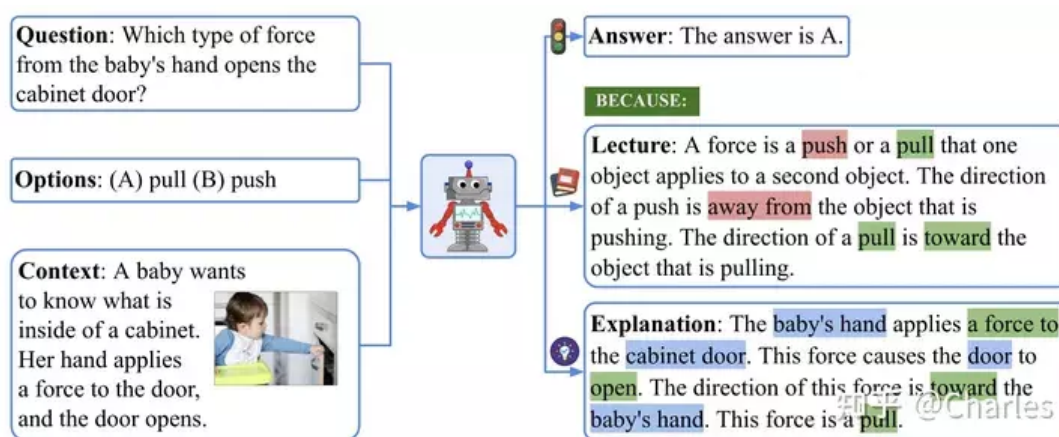


ScienceQA

在回答复杂的问题时，人类可以理解不同模态的信息，并形成一个完整的思维链（Chain of Thought, CoT）。深度学习模型是否可以打开“黑箱”，对其推理过程提供一个思维链？近日，UCLA 和艾伦人工智能研究院（AI2）提出了首个标注详细解释的多模态科学问答数据集 ScienceQA，用于测试模型的多模态推理能力。在 ScienceQA 任务中，作者提出 GPT-3 (CoT) 模型，即在 GPT-3 模型中引入基于思维链的提示学习，从而使得模型能在生成答案的同时，生成相应的推理解释。GPT-3 (CoT) 在 ScienceQA 上实现了 75.17% 的准确率；并且人类评估表明，其可以生成较高质量的解释。

作者收集了全新的科学问答数据集 ScienceQA。ScienceQA 包含 21,208 道来自中小学科学课程的问答多选题。一道典型的问题包含多模态的背景（context）、正确的选项、通用的背景知识（lecture）以及具体的解释（explanation）。

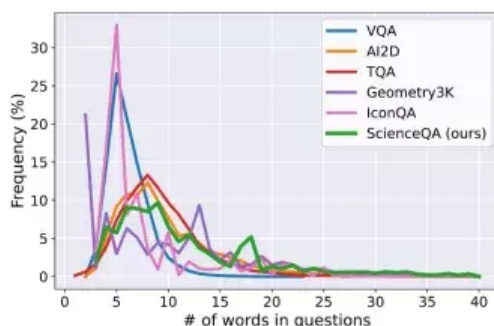


在 ScienceQA 任务中，模型需要在预测答案的同时输出详细地解释。在本文中，作者利用大规模语言模型生成背景知识和解释，作为一种思维链（CoT）来模仿人类具有的多步推理能力。

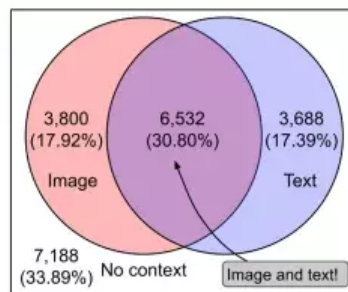
实验表明，目前的多模态问答方法在 ScienceQA 任务不能取得很好的表现。相反，通过基于思维链的提示学习，GPT-3 模型能在 ScienceQA 数据集上取得 75.17% 的准确率，同时可以生成质量较高的解释：根据人类评估，其中 65.2% 的解释相关、正确且完整。思维链也可以帮助 UnifiedQA 模型在 ScienceQA 数据集上取得 3.99% 的提升。

Biology Genes to traits Classification Adaptations Traits and heredity Ecosystems Classification Scientific names Heredity Ecological interactions Cells Plants Animals Plant reproduction	Physics Materials Magnets Velocity and forces Force and motion Particle motion and energy Heat and thermal energy States of matter Kinetic and potential energy Mixture	Geography State capitals Geography Maps Oceania: geography Physical Geography The Americas: geography Oceans and continents Cities States	History Colonial America English colonies in North America The American Revolution World History Greece Ancient Mesopotamia World religions American history Medieval Asia	Civics Social skills Government The Constitution Economics Basic economic principles Supply and demand Banking and finance Global Studies Society and environment
Earth Science Weather and climate Rocks and minerals Astronomy Fossils Earth events Plate tectonics	Chemistry Solutions Physical and chemical change Atoms and molecules Chemical reactions Engineering Designing experiments Engineering practices Units and Measurement Weather and climate	Writing Strategies Supporting arguments Sentences, fragments, and run-ons Word usage and nuance Creative techniques Audience, purpose, and tone Pronouns and antecedents Persuasive strategies Editing and revising Visual elements Opinion writing	Vocabulary Categories Shades of meaning Comprehension strategies Context clues Grammar Sentences and fragments Phrases and clauses Figurative Language Literary devices	Verbs Verb tense Capitalization Formatting Punctuation Fragments Phonology Rhyming Reference Research skills

ScienceQA 包含 21208 个例子，其中有 9122 个不同的问题（question）。10332 道（48.7%）题目有视觉背景信息，10220 道（48.2%）有文本背景信息，6532 道（30.8%）有视觉 + 文本的背景信息。绝大部分问题标注有详细的解释：83.9% 的问题有背景知识标注（lecture），而 91.3% 的问题有详细的解答（explanation）。



(a) Question length distribution of related datasets. SCIENCEQA is distributed more evenly in terms of the number of question words than other datasets.



(b) Question distribution with different context formats. 66.11% of the questions in SCIENCEQA have either an image or text context, while 30.80% have both.

ScienceQA 数据集中问题和背景分布