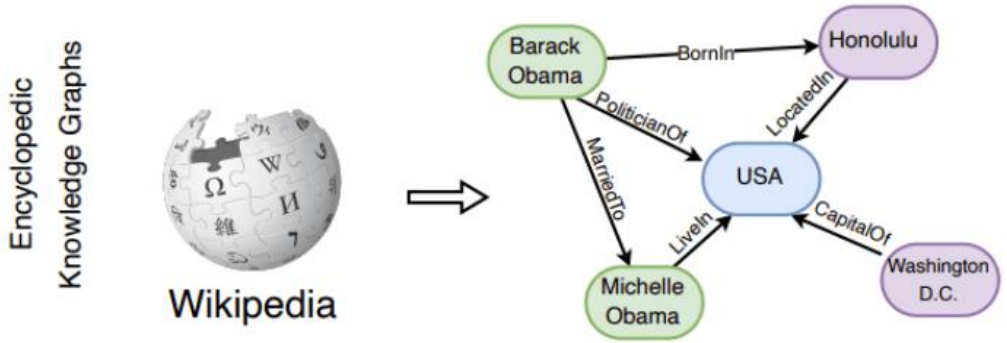
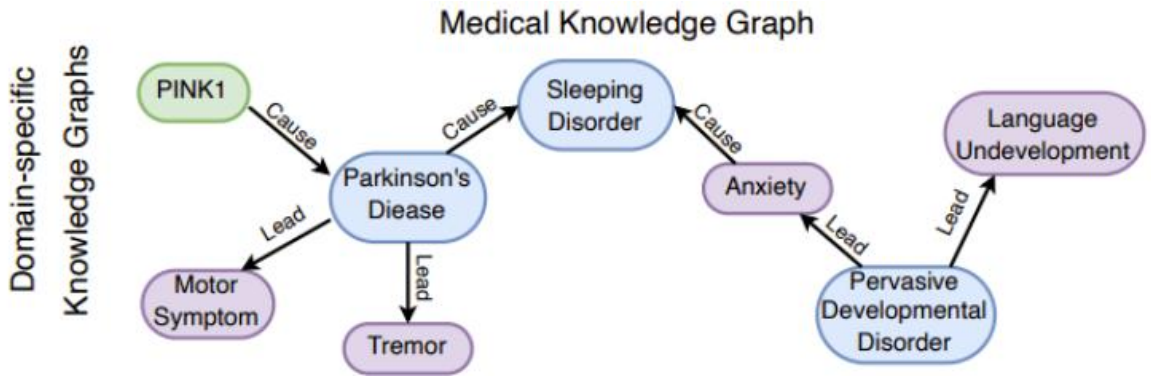
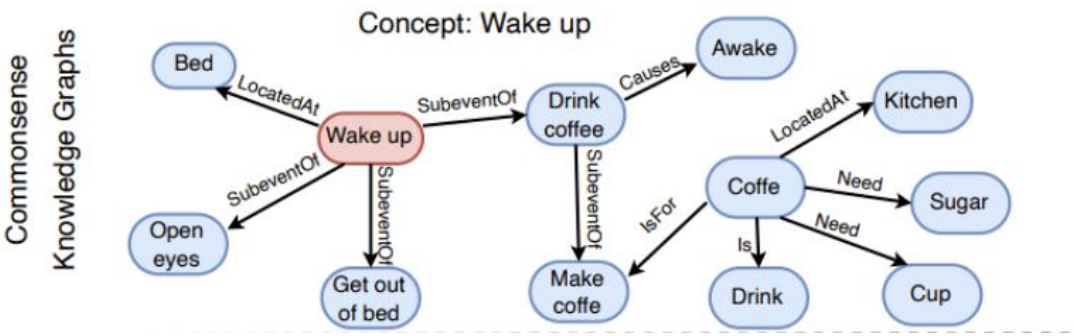
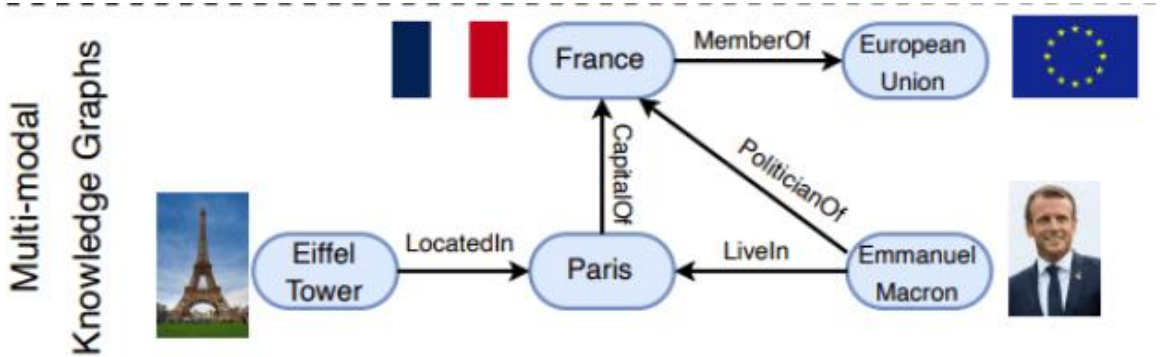
知识图谱

结构化的语义知识库，符号形式，“实体-关系-实体”三元组
通过构建特定领域的知识图谱，就能具备提供精确可靠的特定领域知识的能力。









大语言模型训练方法带来的部分问题分析

- 无法访问最新的信息
- 存在幻觉
- 缺乏一些精准计算的数学技能

数据源带来的问题

GPT-3 training data[1]:9

Dataset	# tokens	Proportion within training
Common Crawl	410 billion	60%
WebText2	19 billion	22%
Books1	12 billion	8%
Books2	55 billion	8%
Wikipedia	3 billion	3%

缺少 “说真话” 的数据源

天真烂漫-> 年轻人
健忘-> 老年人
医护人员 -> 女性
事业有成-> 男性

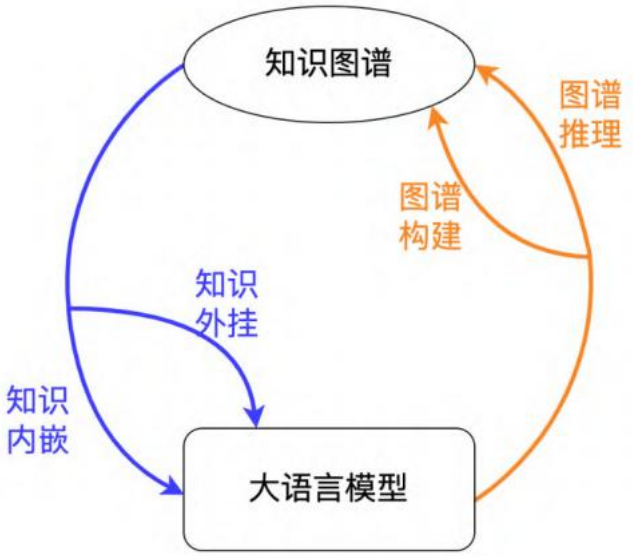
偏见言论

监督学习方法带来的问题

监督信息分布不均匀

行为克隆

*Idea:*利用知识图谱改善大语言模型存在的问题
如何改进创新？



知识图谱与大模型的关系

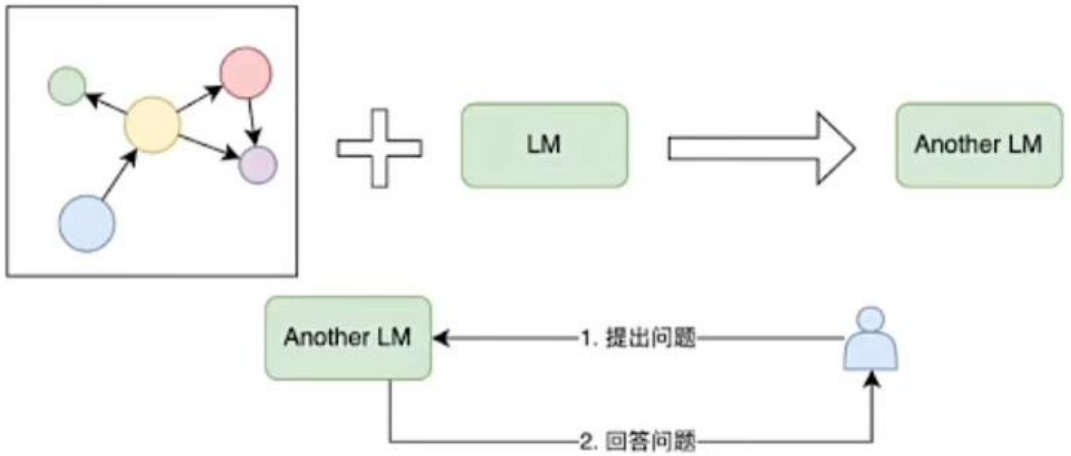
高质量的专业领域知识图谱可以提高LLM的一致性、准确性、可解释性：

- 提升数据源的质量
- 作为额外的信息来源解决监督信息分布不均匀的问题

2种idea：

- 企业合作知识图谱作为模型的训练数据—知识内嵌
- 企业合作知识图谱作为模型的信息来源—知识外挂

Idea1:
企业合作知识图谱作为大语言模型的训练数据



将知识图谱中的信息转换成
自然文本而后作为训练数据

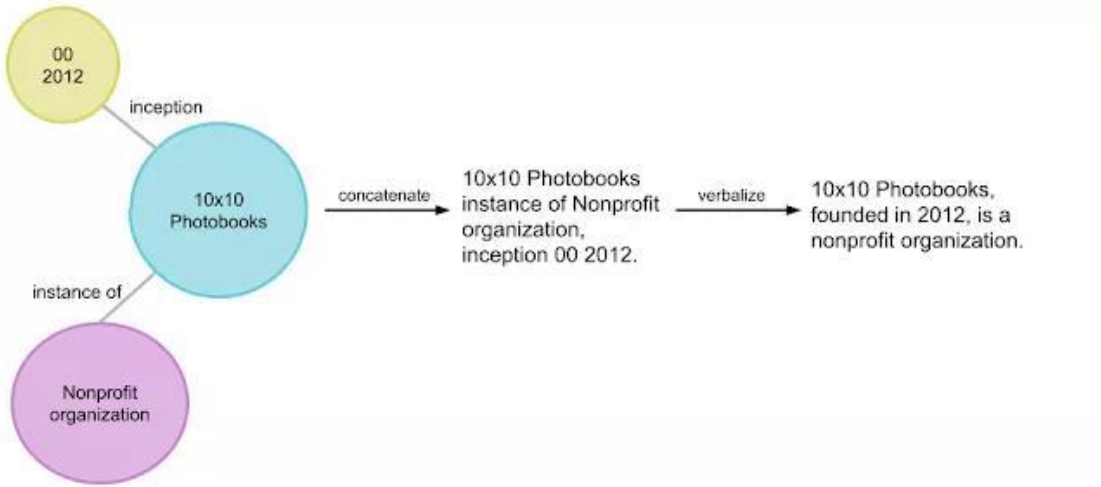
Google提出一个新模型**KELM**(**Knowledge Enhanced Language Model**)，已经被NAACL 2021接受。

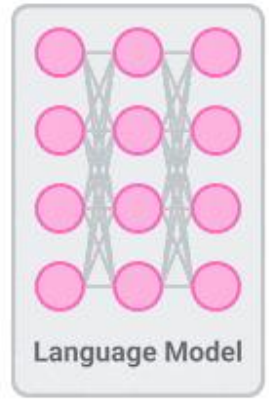
KELM:将知识图谱与语言模型预训练语料库集成

Google主要探索了如何将知识图谱转换为自然语言的句子来增强现有的预训练语料，使其能够在不改变结构的情况下融入语言模型的预训练。

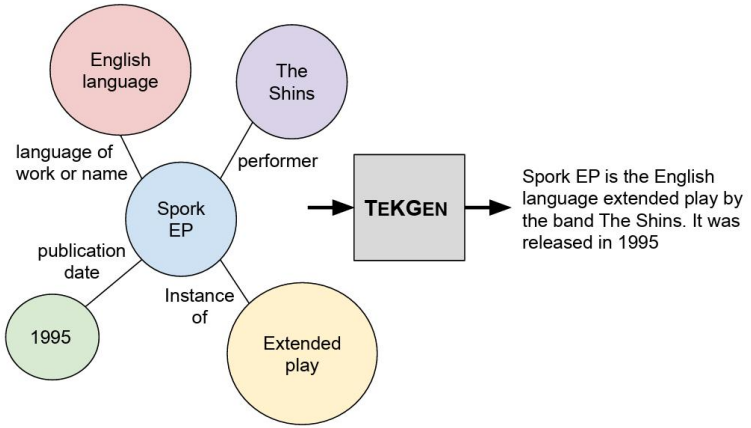
使用的数据集主要是公开的英文知识图谱Wikidata KG，**KELM**模型能够将其转换为自然语言文本，以创建一个合成语料库。

Idea1:
企业合作知识图谱作为大语言模型的训练数据

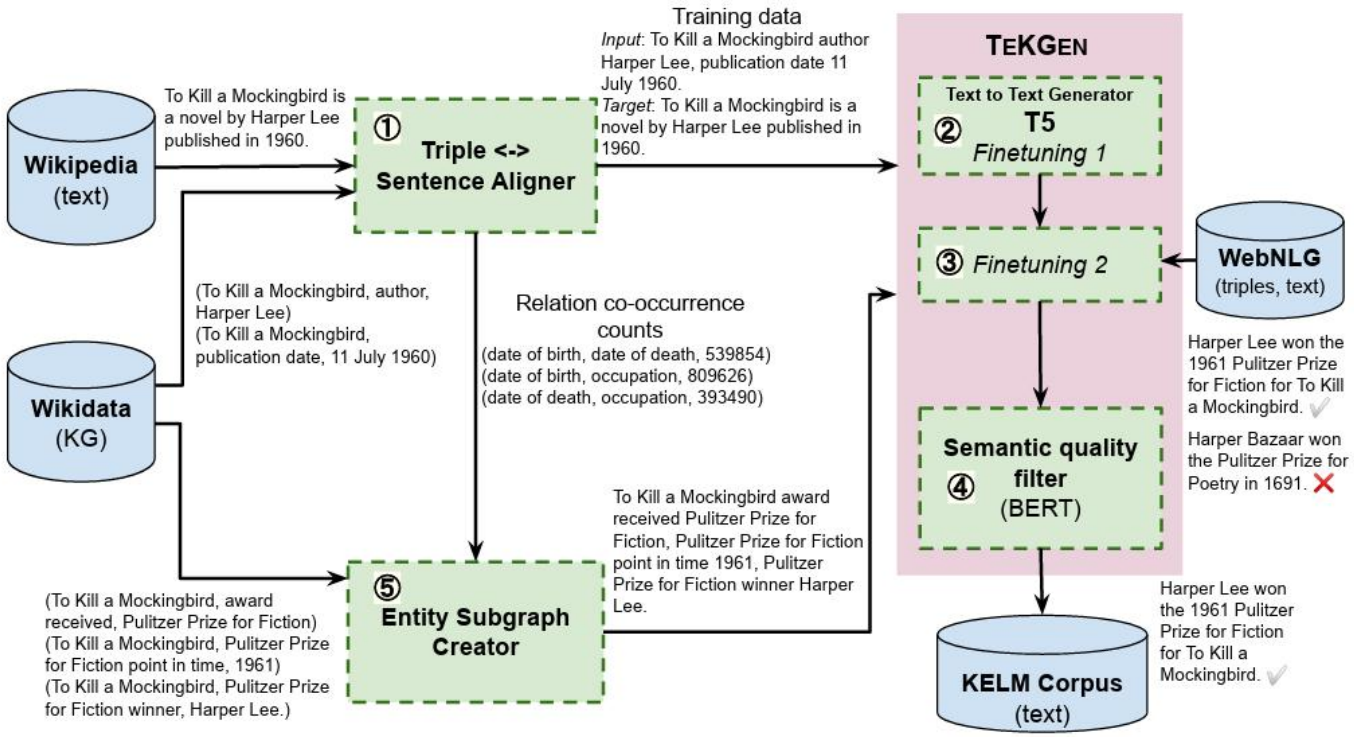




Idea1:
企业合作知识图谱作为大语言模型的训练数据



为了将 Wikidata KG 文本转换为合成的自然的、流畅的语句，Google还开发了一个名为**TEKGEN（Text from KG Generator）**的语言化管道，它由以下几个部分组成: 一个大型启发式构造的、能够自动对齐 Wikipedia启发式对齐器和 Wikidata KG 三元组的训练语料库，一个将 KG 三元组转换为文本的文本到文本生成器(T5) 知识三元组转换为文本的生成器、，一个生成三元组组合语言的实体子图创建器，以及一个消除低质量输出的后处理过滤器。



Idea2:
企业合作知识图谱作为大语言模型的信息来源

把知识图谱独立出来，作为辅助大语言模型的可解释第三方工具

微软提出的解决方案：将Toolformer与ChatGPT结合起来，通过把基础模型与数百万个API连接起来来完成任务。

Toolformer是基于一个预先训练的GPT-J的一个模型，这个模型包含了67亿参数，使用了自监督学习的方法来进行训练，训练过程包括采样和过滤API的调用。

The New England Journal of Medicine is a registered trademark of [QA("Who is the publisher of The New England Journal of Medicine?") → Massachusetts Medical Society] the MMS.

Out of 1400 participants, 400 (or [Calculator(400 / 1400) → 0.29] 29%) passed the test.

The name derives from "la tortuga", the Spanish word for [MT("tortuga") → turtle] turtle.

The Brown Act is California's law [WikiSearch("Brown Act") → The Ralph M. Brown Act is an act of the California State Legislature that guarantees the public's right to attend and participate in meetings of local legislative bodies.] that requires legislative bodies, like city councils, to hold their meetings open to the public.

Idea2:
企业合作知识图谱作为大语言模型的信息来源

Toolformer作为大预言模型，可以通过API来调用和使用不同的工具，每个API调用的输入和输出，都需要被格式化为一个文本对话的序列，才能在会话中自然的流动。

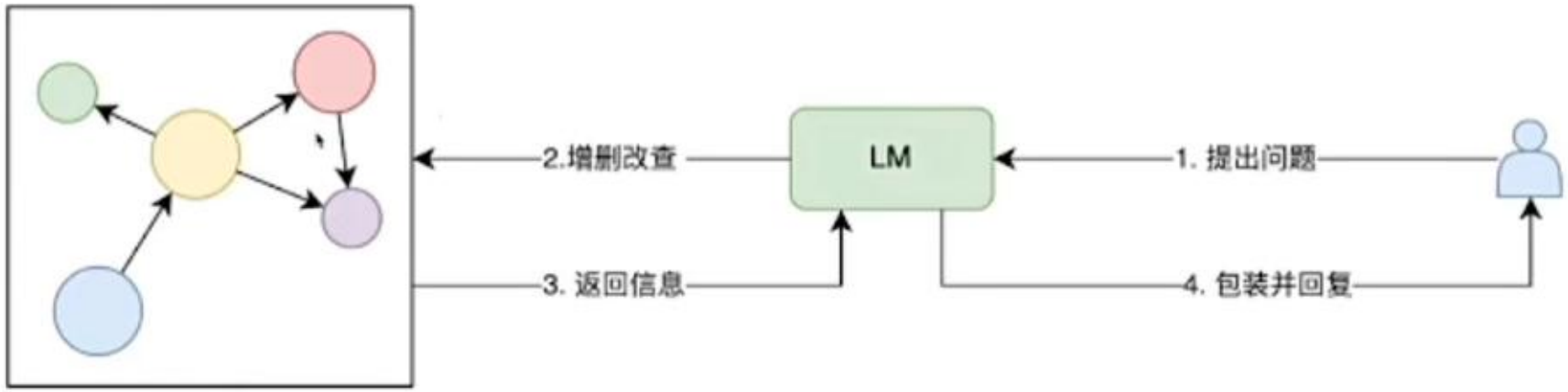
Toolformer是如何利用模型的上下文的学习能力，来对大量潜在的API进行采样和调用的？

Idea2:
企业合作知识图谱作为大语言模型的信息来源

API Name	Example Input	Example Output
Question Answering	Where was the Knights of Columbus founded?	New Haven, Connecticut
Wikipedia Search	Fishing Reel Types	Spin fishing > Spin fishing is distinguished between fly fishing and bait cast fishing by the type of rod and reel used. There are two types of reels used when spin fishing, the open faced reel and the closed faced reel.
Calculator	27 + 4 * 2	35
Calendar	ε	Today is Monday, January 30, 2023.
Machine Translation	sûreté nucléaire	nuclear safety

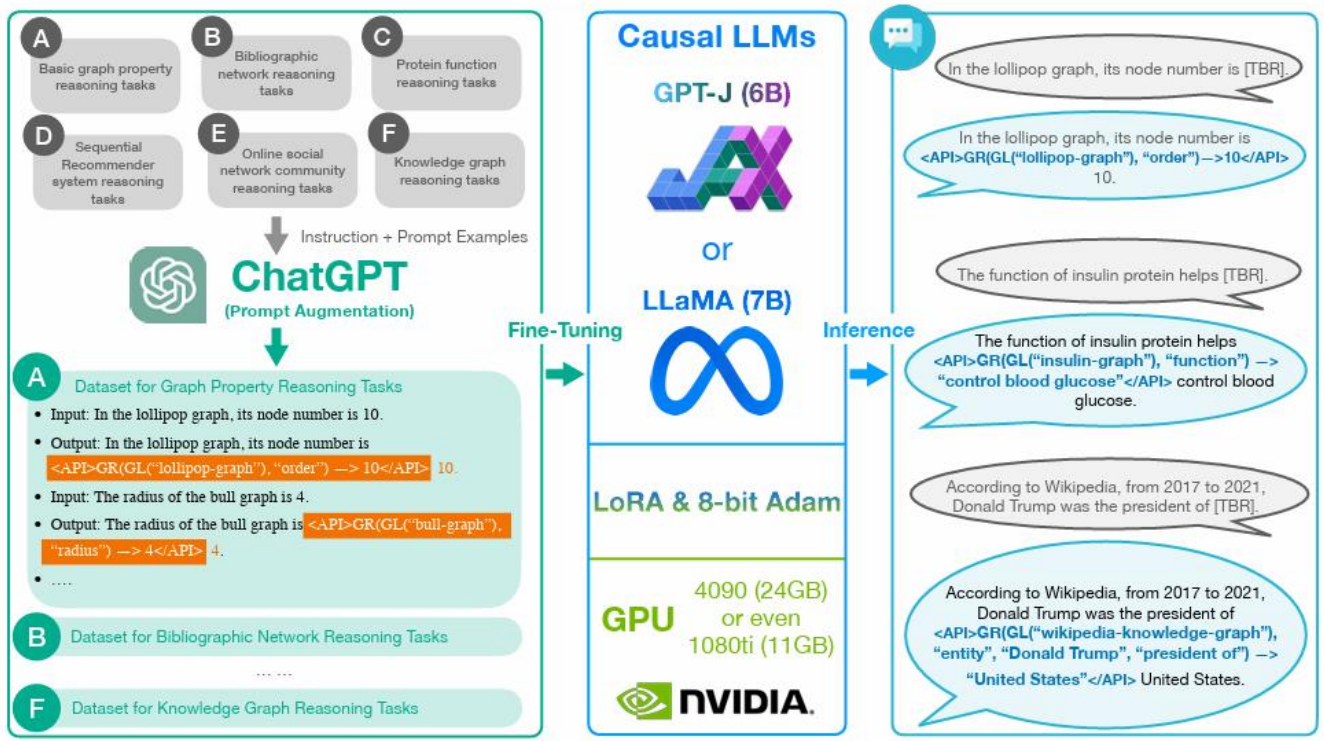
Table 1: Examples of inputs and outputs for all APIs used.

Idea2:
企业合作知识图谱作为大语言模型的信息来源



Idea2:
企业合作知识图谱作为大语言模型的信息来源

Graph-Toolformer



Tasks	API Call Templates	Prompt Examples	
		Inputs	Outputs
Graph Data Loading	$GL(file\text{-}path)$	"The structure of the molecular graph of the benzene ring contains a hexagon."	"The structure of the [GL(file-path="./graphs/benzene-ring")] molecular graph of the benzene ring contains a hexagon."
	$GL(file\text{-}path, node\text{-}subset, link\text{-}subset)$	"There exist a carbon-oxygen double bond in the Acetaldehyde molecular graph."	"There exist a [GL(file-path="./graphs/acetaldehyde", link-subse={C=O})] carbon-oxygen double bond in the Acetaldehyde molecular graph."
	$GL(file\text{-}path) \rightarrow r$	"Lollipop graph look like a spoon."	"[GL(file-path="./graphs/lollipop") -> G_1] Lollipop graph look like a spoon."
Graph Property Reasoning	$GR(graph, "order") \rightarrow r$	"There exist [TBR] nodes in the lollipop graph."	"There exist [GR(G_1, "order") -> 10] nodes in the lollipop graph."
	$GR(graph, "size") \rightarrow r$	"Via [TBR] links, nodes in the lollipop graph are all connected."	"Via [GR(G_1, "size") -> 12] links, nodes in the example lollipop graph are all connected."
	$GR(graph, "density", is\text{-}directed) \rightarrow r$	"The undirected lollipop graph has a density of $\frac{4}{15}$."	"The undirected lollipop graph has a density of [GR(G_1, "density", *is-directed*=False) -> $\frac{4}{15}$]."
	$GR(graph, "eccentricity") \rightarrow r$	"The long 'tail' will lead to large eccentricity [TBR] for many nodes in the lollipop graph."	"The long 'tail' will lead to large eccentricity [GR(G_1, "eccentricity") -> {0 : 7, 1 : 7, 2 : 7, 3 : 6, 4 : 5, 5 : 4, 6 : 4, 7 : 5, 8 : 6, 9 : 7}] for many nodes in the lollipop graph."
	$GR(graph, "eccentricity", \text{node-subset}) \rightarrow r$	"The eccentricity of node #4 in the lollipop graph is [TBR]."	"The eccentricity of node 4 in the lollipop graph is 5 [GR(G_1, "eccentricity", node #4) -> 5]."
	$GR(graph, "radius") \rightarrow r$	"The radius of the lollipop graph is [TBR]."	"The radius of the lollipop graph is [GR(G_1, "radius") -> 4]."
	$GR(graph, "center") \rightarrow r$	"The center of the lollipop graph include node(s) [TBR]."	"The center of the lollipop graph include node(s) [GR(G_1, "center") -> {5, 6}]."
	$GR(graph, "shortest\text{-}path", node_1, node_2) \rightarrow r$	"In the lollipop graph, the length of shortest path between node 1 and node 5 is [TBR]."	"In the lollipop graph, the length of shortest path between node #1 and node #5 is [GR(G_1, "shortest-path", node #1, node #5) -> 3]."
	$GR(graph, "avg\text{-}shortest\text{-}path") \rightarrow r$	"The average length of shortest path for all nodes in the lollipop graph is [TBR]."	"The average length of shortest path for all nodes in the lollipop graph is [GR(G_1, "avg-shortest-path") -> 2.86]."
	$GR(graph, "diameter") \rightarrow r$	"The diameter of the lollipop graph is [TBR] due to the long 'tail'."	"The diameter of the lollipop graph is [GR(G_1, "diameter") -> 7] due to the long 'tail'."
	$GR(graph, "periphery") \rightarrow r$	"The periphery of the lollipop graph includes the nodes [TBR]."	"The periphery of the lollipop graph includes the nodes [GR(G_1, "periphery") -> {0, 1, 2, 9}]."