



Llama 3.1 大模型全解析 与本地部署实战

答辩人：戴炜涛 指导老师：曾煜妮 专业：计算机科学与技术

分工

01

文案
张雅瑞 2022334323029

02

ppt制作

苏泉 2022337621264

王俊杰 2022337621265

03

汇报、PPT修改

戴炜涛 2022337621249

目录

CONTENTS

01

Llama 3.1
简介

02

Llama 3.1
模型概览

03

Llama 3.1
模型架构
解析

04

Llama 3.1
本地部署
方法

05

社区与未
来展望



01

Llama 3.1 简介

○ Llama 3.1模型的发布背景

Llama 3.1模型是由Facebook母公司Meta发布的一款超大型语言模型，旨在推动人工智能技术的创新和普及。这一模型的发布背景体现了Meta在人工智能领域的重要布局与贡献。

Meta，作为科技行业的领军企业，一直致力于推动人工智能技术的研发和应用。Llama 3.1模型是Meta在大型语言模型领域的一次重要突破，展现了其在自然语言处理、对话生成、文本摘要和翻译等任务上的卓越性能。这一模型的发布不仅巩固了Meta在人工智能领域的领先地位，也为全球开发者提供了一个强大的工具，以更低的成本和更高的效率解决复杂的AI问题。



○ Llama 3.1模型开源意义

01



推动全球AI技术交流

开源策略使得全球开发者可以免费使用并改进Llama 3.1模型，这有助于促进不同国家和地区之间的技术交流。

02



促进AI技术创新

开源模型为开发者提供了开放的实验平台，他们可以在此基础上进行各种尝试和探索，发现新的算法和技术。

03



拓展AI应用场景

开源模型使得更多企业和组织能够接触到先进的AI技术，从而将其应用于实际业务场景中。

04



降低AI技术门槛

开源模型降低了AI技术的门槛，使得更多企业和个人能够参与到AI技术的研发和应用中来。



02

Llama 3.1模型概览

Llama 3.1模型特点

多语言处理能力

Llama 3.1支持多种语言的输入与输出，包括但不限于英语、西班牙语、汉语等，使其能够在全球范围内得到广泛应用。



数学推理与问题解决能力

Llama 3.1在数学推理方面表现出色，能够解决复杂的数学问题，如代数、几何等。这使得模型在教育、科研等需要数学推理的领域具有广泛的应用潜力。

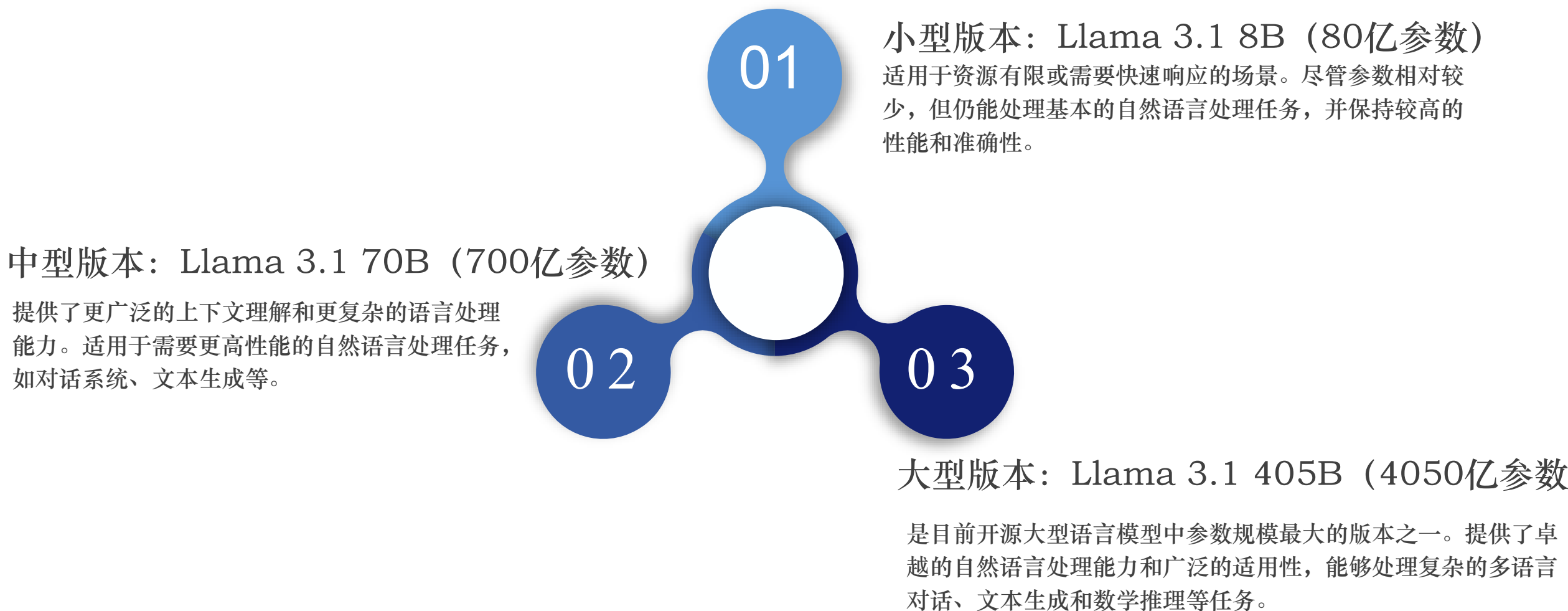


灵活性与可扩展性

Llama 3.1提供不同参数规模的版本，满足不同应用场景的需求。用户可以根据实际需求和计算资源选择适合的模型版本，实现灵活部署和扩展。



○ Llama 3.1 参数规模



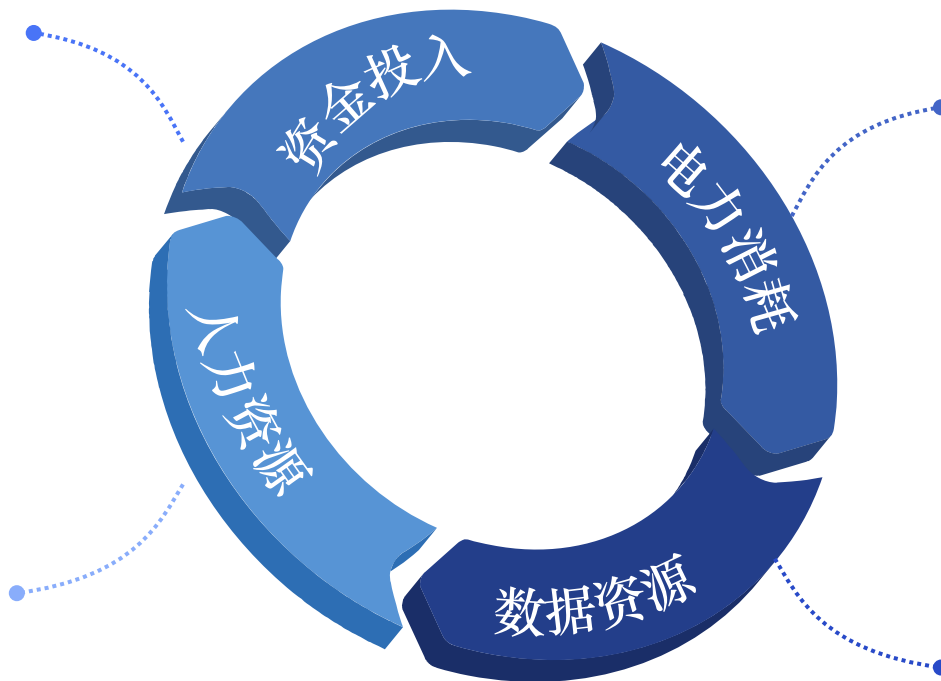
Llama 3.1训练资源投入

资金投入

训练Llama 3.1模型耗费了**数亿美金的巨额成本**。这主要包括购买和维护高性能计算设备（如GPU）、数据收集与处理、模型研发与优化、技术支持与运维等多个方面的费用。

人力资源

训练Llama 3.1模型需要大量的人力资源。这包括AI研究人员、工程师、数据科学家等多个专业的团队。他们需要投入大量的时间和精力来设计和优化模型、处理和分析数据、进行模型的训练和测试等工作。



电力消耗

训练Llama 3.1模型所需的电力消耗也是巨大的。由于模型参数规模庞大，且需要使用大量的训练数据和计算资源，因此训练过程中需要消耗大量的电力。形象地说，其**电力消耗相当于一个小国的用电量**。

数据资源

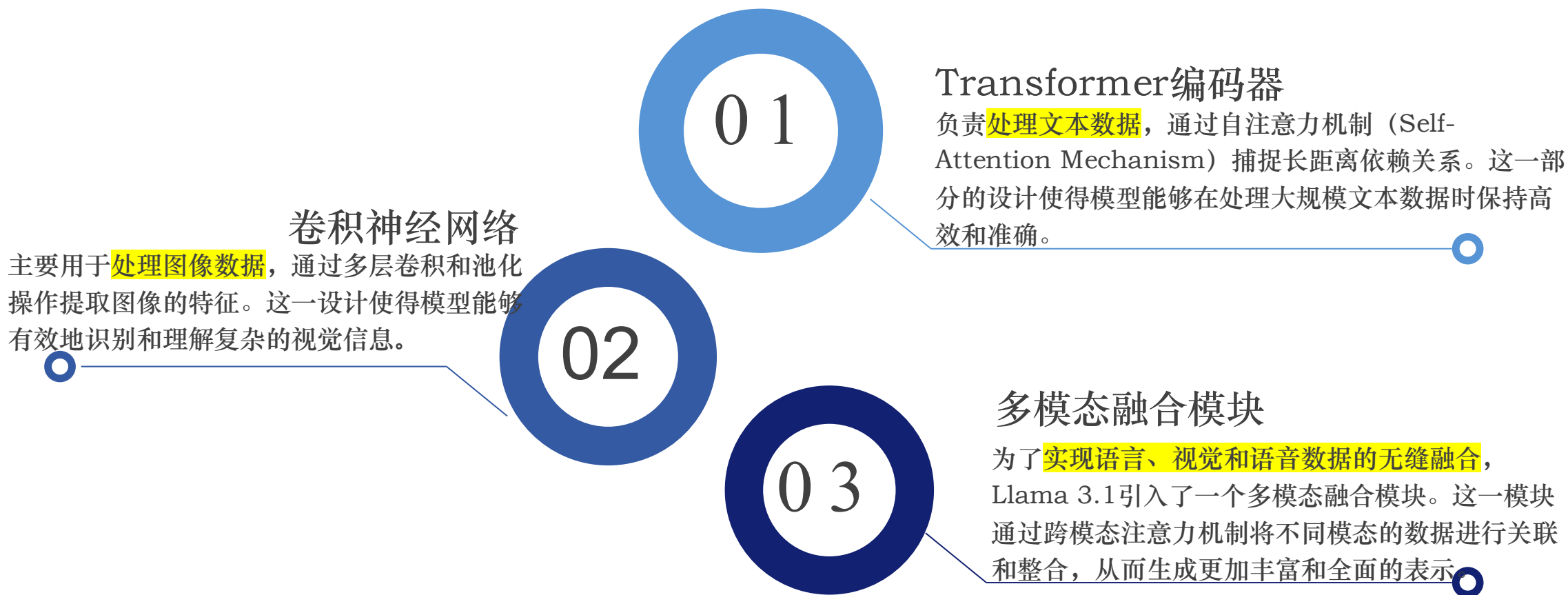
数据资源是训练大型语言模型不可或缺的一部分。Llama 3.1模型在超过15万亿个token的数据集上进行训练，这些**数据涵盖**了多种类型的文本，如新闻文章、书籍、网页内容等。



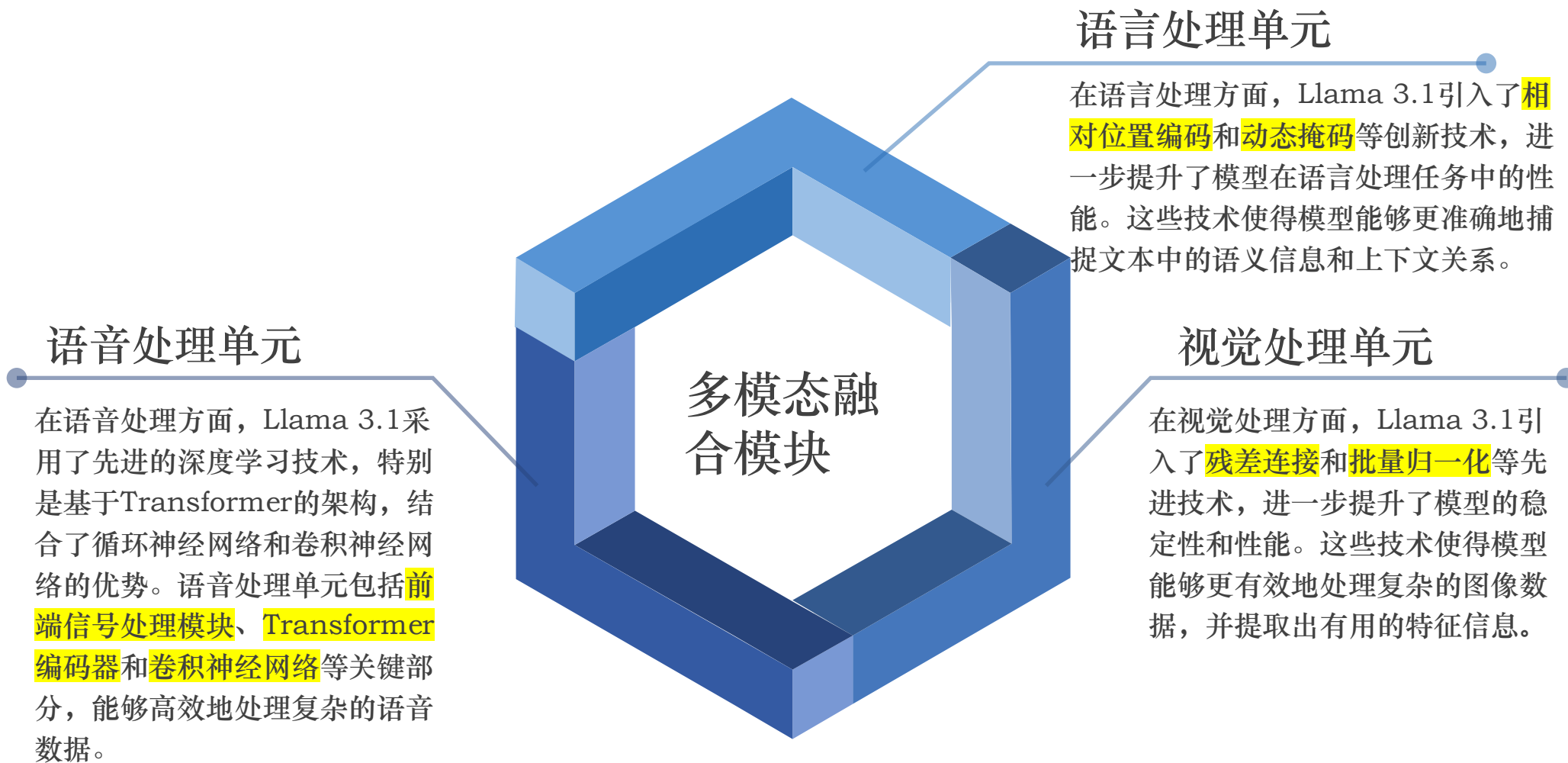
03

Llama 3.1模型架构解析

Llama 3.1模型核心架构



多模态融合模块的深入解析





04

Llama 3.1本地部署方法

Llama 3.1版本选择与环境准备

Llama 3.1版本选择

模型有 8B、70B 及 405B 等不同参数量版本，因8B版本资源需求较低、性能表现相对平衡，这里介绍8B版本的本地部署方法



CPU要求至少需要4个物理核心，时钟频率在2.5GHz以上。



内存要求至少需要16GB的内存，如果使用GPU加速，至少需要8GB的显存。



Python 3.8及以上版本



依赖库：
pytorch、transformers、
SentencePiece、langchain
accelerate、streamlit



Llama 3.1部署

先下载ollama客户端
在cmd里查看是否安装完成:

```
C:\Users\defaultuser0>ollama
Usage:
  ollama [flags]
  ollama [command]

Available Commands:
  serve      Start ollama
  create     Create a model from a Modelfile
  show       Show information for a model
  run        Run a model
  stop       Stop a running model
  pull       Pull a model from a registry
  push       Push a model to a registry
  list       List models
  ps         List running models
  cp         Copy a model
  rm         Remove a model
  help       Help about any command

Flags:
  -h, --help      help for ollama
  -v, --version    Show version information

Use "ollama [command] --help" for more information about a command.
```

安装模型: ollama run llama3.1:8b

```
C:\Users\defaultuser0>ollama run llama3.1:8b
pulling manifest
pulling 8eeb52dfb3bb... 100% 4.7 GB
pulling 948af2743fc7... 100% 1.5 KB
pulling 0ba8f0e314b4... 100% 12 KB
pulling 56bb8bd477a5... 100% 96 B
pulling 1a4c3c319823... 100% 485 B
verifying sha256 digest
writing manifest
success
```

Llama 3.1测试

>>> 8.11和8.9两个数字谁大
很简单！ 8.11 大于 8.9。

>>> 为什么
因为在比较同类数字时，后面的数字是按照从左到右的顺序进行比较的。如果后面的数字相同，那么就比较前面的数字。当我们比较 8.11 和 8.9 时，我们先比较后面的 1 和 9，发现 1 大于 9。因此，8.11 就大于 8.9。

>>> 吃健胃消食片能吃饱吗？
这个问题似乎与前面的话题无关，但我还是会尝试给出答案。

健胃消食片通常是用来帮助消化和减轻肠胃不适的药物。它们可能帮助缓解胀气、恶心和饱腹感等症状，但是并不意味着可以让你吃得更饱。

实际上，健胃消食片的作用是帮助你的身体更好地吸收营养，而不是增加你的食欲或消化能力。所以，即使你服用了健胃消食片，你也需要正常摄入足够的营养和水分来维持健康。

如果你有严重的肠胃不适或饱腹感，建议你寻求专业医疗人员的帮助，而不是自行服用任何药物。

>>> 怀念过去算不算一种刻舟求剑
很好理解了！"刻舟求剑"是一个古老的成语，它描述的是拿着一只舟船（代表过去）去寻找一把剑（代表理想或目标），但是却忘记了自己已经站在了岸上（代表现实）。

你说的怀念过去似乎也是类似的意思。虽然回顾过去可以让我们学习经验和感受快乐，但是如果我们一直执着于过去，不停止在那里，而是让它阻碍我们的前进，那么这也会成为一种障碍。

所以，回答你的问题：是的，它们之间有些相似之处。



05

社区与未来展望

社区开发版本

代码生成微调版本

这种微调版本在大量的代码语料库上进行了训练。它的词汇表和模型结构经过调整，以便更好地理解编程语言的语法、语义和常见的编程模式。



多语言对话微调版本

在多种语言的对话数据上进行了微调。模型能够在不同语言之间进行切换，理解不同文化背景下的语言表达习惯。



情感分析微调版本

基于大量带有情感标签（如正面、负面、中性）的文本数据进行微调。模型能够准确判断输入文本的情感倾向，并且可以分析情感的强度。



未来展望



技术积累与创新

Meta在人工智能领域拥有深厚的技术积累，不断推动技术创新。



市场需求与竞争

随着人工智能技术的快速发展，市场需求日益增长。



开源战略与生态建设

Meta一直秉持开源理念，致力于推动全球AI技术交流与创新。



推动技术融合

它将推动跨学科研究的发展，涉及计算机科学、语言学、心理学、神经科学等多个学科。



应用领域拓展

在物理学、化学、生物学等基础科学领域，Llama模型可以辅助科学家进行理论探索。



社会影响

随着Llama模型性能的提升和应用领域的拓展，会带来一系列伦理挑战。它会推动社会向更加智能化的方向发展。



感谢您的观看