模型推理的方法、系统以及计算设备

本申请要求在 2024 年 3 月 7 日提交中国国家知识产权局、申请号为 202410263183.X 的中国专利申请的优先权，发明名称为“模型推理加速的方法、装置以及计算设备”的中国专利申请的优先权，以及在 2024 年 3 月 29 日提交中国国家知识产权局、申请号为 202410381622.7 的中国专利申请的优先权，发明名称为“模型推理的方法、系统以及计算设备”的中国专利申请的优先权，其全部内容通过引用结合在本申请中。

技术领域

本申请涉及人工智能领域，并且更具体地，涉及一种模型推理的方法、系统以及计算设备。

背景技术

近年来，随着大模型的快速发展，基于大模型构建的智慧应用已经广泛落地，并影响着人们产生信息、使用信息的方式。大模型的一个代表性特点是参数量巨大，为了不断提升大模型的应用效果，当前大模型仍在沿着扩大参数量的方向高速发展。巨大参数量一方面成就了大模型，另一方面也成为大模型普及的掣肘，其中一个重要原因，是大模型推理态的高成本服务化部署。由于参数量巨大，大模型在推理时需要使用大量的计算资源，但受限于自回归解码 Token By Token 的生成方法，大模型推理时对计算资源的利用率较低。如何让模型在推理时充分利用计算资源，提升模型推理吞吐率，降低大模型推理成本，是一个棘手且重要的问题。

相关的生成式模型进行推理的方案中，Draft 生成和 Draft 验证两个关键阶段中，Draft 生成模型和 Draft 验证模型在协同方式和对应方式上具有强耦合的关系，该强耦合的关系在程序启动时已经固化，从而导致 Draft 生成模型和 Draft 验证模型在协同方式和对应方式上的灵活性较差，降低了模型推理的效率。

因此，如何提高模型推理的效率成为亟需要解决的技术问题。

发明内容

本申请提供一种模型推理的方法、系统以及计算设备，能够提高模型推理的效率。

第一方面，提供了一种模型推理的方法，该方法包括：从 Draft 验证模型池中确定至少一个 Draft 验证模型，所述 Draft 验证模型池中包括多个 Draft 验证模型；从 Draft 生成模型池中确定与所述至少一个 Draft 验证模型对应的至少一个 Draft 生成模型，其中，所述 Draft 生成模型池中包括多个 Draft 生成模型；接收第一用户输入的文本提示信息，所述文本提示信息用于生成文本信息；从 Draft 池中获取与所述文本提示信息相关的目标 Draft，所述 Draft 池中存储有多个 Draft，所述多个 Draft 包括所述至少一个 Draft 生成模型生成的 Draft；通过所述至少一个 Draft 验证模型对所述目标 Draft 进行验证，并得到验证结果。根据所述验证结果生成所述文本信息并反馈给所述第一用户。

作为示例，上述文本提示信息例如可以是用于输入的 prompt，该 prompt 通常指的是一段文本或信息，其作为生成模型的一个初始的输入或引导，用于引导生成模型生成 Draft。

上述确定的至少一个 Draft 验证模型可以是 Draft 验证模型池中厂商内置的 Draft 验证模型，或者还可以是用户基于自己的数据对 Draft 验证模型池中厂商内置的 Draft 验证模型进行微调后得到的 Draft 验证模型，或者还可以是该用户自己的 Draft 验证模型（用户通过对接接口将自己的 Draft 验证模型对接至 Draft 验证模型池）。

上述确定的至少一个 Draft 生成模型可以是 Draft 生成模型池中厂商内置的 Draft 生成模型，或者还可以是用户基于自己的数据对 Draft 生成模型池中厂商内置的 Draft 生成模型进行微调后得到的 Draft 生成模型，或者还可以是该用户自己的 Draft 生成模型（用户通过对接接口将自己的 Draft 生成模型对接至 Draft 生成模型池）。

上述技术方案中， Draft 生成模型和 Draft 验证模型在对应方式和协同方式上可以实现完全解耦，Draft 生成模型和 Draft 验证模型之间可以动态组合，任意配比，且 Draft 生成模型和 Draft 验证模型在对

应方式和协同方式上非固定化，且更加灵活，从而可以提高模型推理的效率。

结合第一方面，在第一方面的某些实现方式中，该至少一个 Draft 验证模型和该至少一个 Draft 生成模型之间的对应模式包括：一对一，一对多，多对一，多对多。

上述技术方案中，至少一个 Draft 验证模型和该至少一个 Draft 生成模型之间存在一对一，一对多，多对一，多对多等多种对应关系，且可以实时切换，动态组合。

结合第一方面，在第一方面的某些实现方式中，该方法还包括：该至少一个 Draft 生成模型包括 Draft 同步生成模型，该方法还包括：根据所述文本提示信息获得 Draft 生成请求信息；将所述 Draft 生成请求信息发送给所述 Draft 同步生成模型，所述 Draft 同步生成模型用于根据所述 Draft 生成请求信息生成 Draft；将该 Draft 同步生成模型生成的 Draft 存储在该 Draft 池中。

上述技术方案中，Draft 池可以对系统在运行过程中 Draft 生成模型产生的 Draft 进行记录和管理，使得 Draft 池中记录的 Draft 可以供后续推理使用，让 Draft 得以复用，避免生成的 Draft 仅用于当前请求的推理验证，在一定程度上可以减少 Draft 产生所占用的算力和时间，从而进一步提高了模型推理的效率，降低模型推理的成本。

结合第一方面，在第一方面的某些实现方式中，该 Draft 池中还存储有 Draft 的属性信息。

上述技术方案中，还可以在 Draft 池中记录 Draft 的属性信息，这样为后续推理提供了有效的 Draft 筛选机制，从而提高模型推理的效率，降低模型推理的成本。

结合第一方面，在第一方面的某些实现方式中，根据该 Draft 池中存储的 Draft 的属性信息，从该 Draft 池中获取与该文本提示信息相关的目标 Draft。

上述技术方案中，可以基于 Draft 池中存储的 Draft 的属性信息，为不同的文本提示信息（例如，prompt）筛选、检索更有效的 Draft。

结合第一方面，在第一方面的某些实现方式中，该方法还包括：根据该 Draft 池中存储的 Draft 的属性信息，对该 Draft 池中存储的 Draft 进行检索和/或更新。

上述技术方案中，可以基于 Draft 池中存储的 Draft 的属性信息，对 Draft 池进行自我维护和自动更新。

结合第一方面，在第一方面的某些实现方式中，该方法还包括：根据该 Draft 池中存储的 Draft 的属性信息，对该 Draft 生成模型池中包括的 Draft 生成模型进行操作，该操作包括：激活，休眠，释放。

上述技术方案中，可以基于 Draft 池中存储的 Draft 的属性信息，对该 Draft 生成模型池中包括的 Draft 生成模型进行激活，休眠，释放等操作，从而提高模型推理的效率。

结合第一方面，在第一方面的某些实现方式中，该 Draft 的属性信息包括以下中的至少一种：该 Draft 的领域、内容、对应的 Draft 生成模型的标签、被调用的次数、命中的次数、命中的文本提示信息的列表。

上述技术方案中，可以 Draft 的多维属性信息，使得 Draft 召回更精准，在 Draft 验证阶段带来更高的命中率。

结合第一方面，在第一方面的某些实现方式中，该 Draft 生成请求信息还包括生成 Draft 的数量，该生成 Draft 的数量是根据模型推理资源的使用情况确定的，该模型推理资源用于该 Draft 生成模型池中的 Draft 生成模型产生 Draft 和/或该 Draft 验证模型池中的 Draft 验证模型执行验证过程。

上述技术方案中，可以基于当前推理资源状态、自愿占用情况等实现对 Draft 数量的自适应调整。

结合第一方面，在第一方面的某些实现方式中，该至少一个 Draft 生成模型包括 Draft 异步生成模型，该方法还包括：在推理负载小于预设阈值的情况下，激活该 Draft 异步生成模型生成 Draft；将该 Draft 异步生成模型生成的 Draft 存储在该 Draft 池中。

上述技术方案中，可以在推理负载小于预设阈值的情况下，激活 Draft 异步生成模型生成 Draft，避免 Draft 生成模型和 Draft 验证模型之间同步模式下的相互等待，从而提高模型推理过程对推理资源的利用率，提高模型推理的效率，降低模型推理的成本。

结合第一方面，在第一方面的某些实现方式中，该方法还包括：根据模型推理资源的使用情况，自动申请/释放该模型推理资源。

上述技术方案中，可以结合当前推理系统状态，实时申请/释放推理计算资源。

结合第一方面，在第一方面的某些实现方式中，该 Draft 池中包括共享 Draft 池和第一用户的私有 Draft 池，从该共享 Draft 池和/或该第一用户的私有 Draft 池中获取与该文本提示信息相关的目标 Draft。

上述技术方案中，Draft 池支持多租模式，Draft 池主要由两部分构成，一个为共享 Draft 池，这个池

可以在多个租户之间共享，另一个为客户的私有 Draft 池，每个客户的 Draft 池相互隔离。用户在配置各自的推理服务时，可以选择使用共享 Draft 池进行冷启动。

结合第一方面，在第一方面的某些实现方式中，将该 Draft 同步生成模型生成的 Draft 存储在第一用户的私有 Draft 池中。

上述技术方案中，推理服务启动后，系统会为每个用户划分一个独立的 Draft 池，每个用户的推理服务过程中所产生的 Draft，会写入各自私有的 Draft 池中。不同用户推理服务运行过程中的 Draft 可以独立维护。

结合第一方面，在第一方面的某些实现方式中，将该 Draft 异步生成模型生成的 Draft 存储在第一用户的私有 Draft 池中。

第二方面，提供了一种模型推理的系统，该系统包括：Draft 验证管理单元，Draft 生成管理单元，至少一个 Draft 验证模型，其中，Draft 验证管理单元，用于从 Draft 验证模型池中确定至少一个 Draft 验证模型，所述 Draft 验证模型池中包括多个 Draft 验证模型；Draft 生成管理单元，用于从 Draft 生成模型池中确定与所述至少一个 Draft 验证模型对应的至少一个 Draft 生成模型，其中，所述 Draft 生成模型池中包括多个 Draft 生成模型；所述至少一个 Draft 验证模型，用于接收第一用户输入的文本提示信息，所述文本提示信息用于生成文本信息；所述 Draft 验证管理单元，还用于从 Draft 池中获取与所述文本提示信息相关的目标 Draft，所述 Draft 池中存储有多个 Draft，所述多个 Draft 包括所述至少一个 Draft 生成模型生成的 Draft；所述至少一个 Draft 验证模型，还用于对所述目标 Draft 进行验证，并得到验证结果；所述至少一个 Draft 验证模型，还用于根据所述验证结果生成所述文本信息并反馈给所述第一用户。

结合第二方面，在第二方面的某些实现方式中，所述至少一个 Draft 生成模型包括 Draft 同步生成模型，所述至少一个 Draft 验证模型，还用于根据所述文本提示信息获得 Draft 生成请求信息，并将 Draft 生成请求信息通过所述 Draft 验证管理单发送给所述 Draft 生成管理单元；所述 Draft 生成管理单元，还用于将所述 Draft 生成请求信息发送给所述 Draft 同步生成模型；所述 Draft 同步生成模型，用于根据所述 Draft 生成请求信息生成 Draft 并发送给所述 Draft 生成管理单元；所述 Draft 生成管理单元，还用于将所述 Draft 同步生成模型生成的 Draft 存储在所述 Draft 池中。

结合第二方面，在第二方面的某些实现方式中，所述至少一个 Draft 生成模型包括 Draft 异步生成模型，所述 Draft 验证管理单元，还用于在推理负载小于预设阈值的情况下，激活所述 Draft 异步生成模型生成 Draft；所述 Draft 异步生成模型，用于生成 Draft 并发送给所述 Draft 验证管理单元；所述 Draft 验证管理单元，还用于将所述 Draft 异步生成模型生成的 Draft 存储在所述 Draft 池中。

结合第二方面，在第二方面的某些实现方式中，所述 Draft 池中还存储有 Draft 的属性信息。

结合第二方面，在第二方面的某些实现方式中，所述 Draft 验证管理单元具体用于：根据所述 Draft 池中存储的 Draft 的属性信息，从所述 Draft 池中获取与所述文本提示信息相关的目标 Draft。

结合第二方面，在第二方面的某些实现方式中，所述系统还包括：Draft 分析单元，用于根据所述 Draft 池中存储的 Draft 的属性信息，对所述 Draft 池中存储的 Draft 进行检索和/或更新。

结合第二方面，在第二方面的某些实现方式中，所述 Draft 生成管理单元，还用于根据所述 Draft 池中存储的 Draft 的属性信息，对所述 Draft 生成模型池中包括的 Draft 生成模型进行操作，所述操作包括：激活，休眠，释放。

结合第二方面，在第二方面的某些实现方式中，所述 Draft 的属性信息包括以下中的至少一种：所述 Draft 的领域、内容、对应的 Draft 生成模型的标签、被调用的次数、命中的次数、命中的文本提示信息的列表。

结合第二方面，在第二方面的某些实现方式中，所述 Draft 生成请求信息还包括生成 Draft 的数量，所述生成 Draft 的数量是根据模型推理资源的使用情况确定的，所述模型推理资源用于所述 Draft 生成模型池中的 Draft 生成模型产生 Draft 和/或所述 Draft 验证模型池中的 Draft 验证模型执行验证过程。

结合第二方面，在第二方面的某些实现方式中，所述系统还包括：推理资源分析单元，用于根据模型推理资源的使用情况，自动申请/释放所述模型推理资源。

结合第二方面，在第二方面的某些实现方式中，所述 Draft 池中包括共享 Draft 池和第一用户的私有 Draft 池，所述 Draft 验证管理单元具体用于：从所述共享 Draft 池和/或所述第一用户的私有 Draft 池中获取与所述文本提示信息相关的目标 Draft。

结合第二方面，在第二方面的某些实现方式中，所述 Draft 生成管理单元具体用于：将所述 Draft 同

步生成模型生成的 Draft 存储在第一用户的私有 Draft 池中。

结合第二方面，在第二方面的某些实现方式中，所述 Draft 生成管理单元具体用于：将所述 Draft 异步生成模型生成的 Draft 存储在第一用户的私有 Draft 池中。

结合第二方面，在第二方面的某些实现方式中，所述至少一个 Draft 验证模型和所述至少一个 Draft 生成模型之间的对应模式包括：一对一，一对多，多对一，多对多。

应理解，第二方面以及第二方面的各个实现方式的有益效果与第一方面以及第一方面的各个实现方式的有益效果类似，具体的请参考第一方面以及第一方面的各个实现方式的有益效果，此处不再赘述。

第三方面，提供了一种计算设备，包括处理器和存储器，可选地，还包括输入输出接口。其中该处理器用于控制该输入输出接口收发信息，该存储器用于存储计算机程序，该处理器用于从存储器中调用并运行该计算机程序，使得该执行第一方面或第一方面任意一种可能的实现方式中的方法。

可选地，该处理器可以是通用处理器，可以通过硬件来实现也可以通过软件来实现。当通过硬件实现时，该处理器可以是逻辑电路、集成电路等；当通过软件来实现时，该处理器可以是一个通用处理器，通过读取存储器中存储的软件代码来实现，该存储器可以集成在处理器中，可以位于该处理器之外，独立存在。

第四方面，提供了一种计算设备集群，包括至少一个计算设备，每个计算设备包括处理器和存储器；该至少一个计算设备的处理器用于执行该至少一个计算设备的存储器中存储的指令，以使得该计算设备集群执行第一方面或第一方面任意一种可能的实现方式中的方法。

第五方面，提供了一种芯片，该芯片获取指令并执行该指令来实现上述第一方面以及第一方面的任意一种实现方式中的方法。

可选地，作为一种实现方式，该芯片包括处理器与数据接口，该处理器通过该数据接口读取存储器上存储的指令，执行上述第一方面以及第一方面的任意一种实现方式中的方法。

可选地，作为一种实现方式，该芯片还可以包括存储器，该存储器中存储有指令，该处理器用于执行该存储器上存储的指令，当该指令被执行时，该处理器用于执行第一方面以及第一方面中的任意一种实现方式中的方法。

第六方面，提供了一种包含指令的计算机程序产品，当该指令被计算设备运行时，使得该计算设备执行如上述第一方面以及第一方面的任意一种实现方式中的方法。

第七方面，提供了一种包含指令的计算机程序产品，当该指令被计算设备集群运行时，使得该计算设备集群执行如上述第一方面以及第一方面的任意一种实现方式中的方法。

第八方面，提供了一种计算机可读存储介质，包括计算机程序指令，当该计算机程序指令由计算设备执行时，该计算设备执行如上述第一方面以及第一方面的任意一种实现方式中的方法。

作为示例，这些计算机可读存储包括但不限于如下的一个或者多个：只读存储器（read-only memory，ROM）、可编程 ROM（programmable ROM，PROM）、可擦除的 PROM（erasable PROM，EPROM）、Flash 存储器、电 EPROM（electrically EPROM，EEPROM）以及硬盘驱动器（hard drive）。

可选地，作为一种实现方式，上述存储介质具体可以是非易失性存储介质。

第九方面，提供了一种计算机可读存储介质，包括计算机程序指令，当该计算机程序指令由计算设备集群执行时，该计算设备集群执行如上述第一方面以及第一方面的任意一种实现方式中的方法。

作为示例，这些计算机可读存储包括但不限于如下的一个或者多个：只读存储器（read-only memory，ROM）、可编程 ROM（programmable ROM，PROM）、可擦除的 PROM（erasable PROM，EPROM）、Flash 存储器、电 EPROM（electrically EPROM，EEPROM）以及硬盘驱动器（hard drive）。

可选地，作为一种实现方式，上述存储介质具体可以是非易失性存储介质。

附图说明

图 1 是适用于本申请实施例的一种云场景的示意性框图。

图 2 是本申请实施例提供的一种模型推理系统的示意性框图。

图 3 是本申请实施例提供的另一种模型推理系统的示意性框图。

图 4 是本申请实施例提供的一种模型推理的方法的示意性流程图。

图 5 是本申请实施例提供的一种用户设置 Draft 生成模型和 Draft 验证模型的示意性框图。

图 6 是本申请实施例提供的另一种模型推理系统的示意性框图。

图 7 是本申请实施例提供的另一种模型推理系统的示意性框图。
图 8 是本申请实施例提供的另一种模型推理的方法的示意性流程图。
图 9 是本申请实施例提供的另一种用户设置 Draft 生成模型和 Draft 验证模型的示意性框图。
图 10 是本申请实施例提供的另一种模型推理系统的示意性框图。
图 11 是本申请实施例提供的另一种模型推理系统的示意性框图。
图 12 是本申请实施例提供的另一种模型推理的方法的示意性流程图。
图 13 是本申请实施例提供的另一种用户设置 Draft 生成模型和 Draft 验证模型的示意性框图。
图 14 是本申请实施例提供的另一种模型推理系统的示意性框图。
图 15 是本申请实施例提供的另一种模型推理系统的示意性框图。
图 16 是本申请实施例提供的另一种模型推理的方法的示意性流程图。
图 17 是本申请实施例提供的一种用户 A 设置 Draft 生成模型和 Draft 验证模型的示意性框图。
图 18 是本申请实施例提供的另一种模型推理系统的示意性框图。
图 19 是本申请实施例提供的一种用户 B 设置 Draft 生成模型和 Draft 验证模型的示意性框图。
图 20 是本申请实施例提供的另一种模型推理系统的示意性框图。
图 21 是本申请实施例提供的一种模型推理的装置 2100 的示意性框图。
图 22 是本申请实施例提供的一种计算设备 1500 的架构示意图。
图 23 是本申请实施例提供的一种计算设备集群的架构示意图。
图 24 是本申请实施例提供的计算设备 1500A 和 1500B 之间通过网络进行连接的示意图。

具体实施方式

下面将结合附图，对本申请中的技术方案进行描述。

本申请将围绕包括多个设备、组件、模块等的系统来呈现各个方面、实施例或特征。应当理解和明白的是，各个系统可以包括另外的设备、组件、模块等，并且/或者可以并不包括结合附图讨论的所有设备、组件、模块等。此外，还可以使用这些方案的组合。

另外，在本申请实施例中，“示例的”、“例如”等词用于表示作例子、例证或说明。本申请中被描述为“示例”的任何实施例或设计方案不应被解释为比其它实施例或设计方案更优选或更具优势。确切而言，使用示例的一词旨在以具体方式呈现概念。

本申请实施例中，“相应的（corresponding，relevant）”和“对应的(corresponding)”有时可以混用，应当指出的是，在不强调其区别时，其所要表达的含义是一致的。

本申请实施例描述的业务场景是为了更加清楚地说明本申请实施例的技术方案，并不构成对于本申请实施例提供的技术方案的限定，本领域普通技术人员可知，随着网络架构的演变和新业务场景的出现，本申请实施例提供的技术方案对于类似的技术问题，同样适用。

在本说明书中描述的参考“一个实施例”或“一些实施例”等意味着在本申请的一个或多个实施例中包括结合该实施例描述的特定特征、结构或特点。由此，在本说明书中的不同之处出现的语句“在一个实施例中”、“在一些实施例中”、“在其他一些实施例中”、“在另外一些实施例中”等不是必然都参考相同的实施例，而是意味着“一个或多个但不是所有的实施例”，除非是以其他方式另外特别强调。术语“包括”、“包含”、“具有”及它们的变形都意味着“包括但不限于”，除非是以其他方式另外特别强调。

本申请中，“至少一个”是指一个或者多个，“多个”是指两个或两个以上。“和/或”，描述关联对象的关联关系，表示可以存在三种关系，例如，A 和/或 B，可以表示：包括单独存在 A，同时存在 A 和 B，以及单独存在 B 的情况，其中 A，B 可以是单数或者复数。字符“/”一般表示前后关联对象是一种“或”的关系。“以下至少一项(个)”或其类似表达，是指的这些项中的任意组合，包括单项（个）或复数项（个）的任意组合。例如，a，b，或 c 中的至少一项（个），可以表示：a，b，c，a-b，a-c，b-c，或 a-b-c，其中 a，b，c 可以是单个，也可以是多个。

为了便于描述，下面先对本申请实施例涉及的概念进行解释说明。

1、生成模型

生成模型也可以称为生成式模型，是一种机器学习模型，其目标是生成与训练数据类似的新数据。具体的，生成模型是一种利用概率模型进行数据生成的方法，它可以将数据生成看作是一个从先验分布中抽取样本的过程，这样就可以生成新的数据，而不是仅仅对已有数据进行分类。

2、自回归解码

自回归解码是生成模型中的一种重要的解码策略，具体的，生成模型会逐个生成序列中的 token，自回归解码的方式为 token by token，也即，在自回归解码的过程中，会根据已有的上下文（即之前生成的 token）来计算下一个 token 的概率分布，根据这个概率分布，选择一个 token 作为下一个输出。这个过程是递归的，意味着每个新生成的 token 都会成为后续预测的条件。例如，在文本生成任务中，会先生成序列的第一个 token，然后根据这个 token 来预测第二个 token，如此往复，直到生成完整的序列。

应理解，上述 token 可以是文字、字符、单词、图像块或其他类型的数据单元，具体取决于模型的应用领域和任务。

3、Draft：

一个 Draft 包括带有上下文关系的多个 token，也即，上下文关系的多个 token 被称为一个 Draft。在生成模型的上下文中，Draft 通常意味着生成模型生成的初步结果（例如，初步的版本或未完成的草案），这些结果可能还需要进一步的修正或验证。

4、prompt

prompt 通常指的是一段文本或信息，其作为生成模型的一个初始的输入或引导，用于引导生成模型生成 Draft。也即，prompt 可以看做是生成 Draft 的起点或触发因素，通过提供适当的 prompt，生成模型能够生成一个与该 prompt 内容相关的 Draft。

5、验证模型

验证模型是根据输入的 prompt 对生成的 Draft 中包括的多个 token 进行验证，确定通过验证的至少一个 token。

近年来，随着大模型的快速发展，基于大模型构建的智慧应用已经广泛落地，并影响着人们产生信息、使用信息的方式，比如 ChatGPT、Sora 等。大模型的一个代表性特点是参数量巨大，为了不断提升大模型的应用效果，当前大模型仍在沿着扩大参数量的方向高速发展。2023 年，OpenAI 公布了拥有千亿参数量的 ChatGPT，短短几个月后，便相继公布了有万亿参数量的 GPT4，模型的对话效果提升显著，在业内引起了持续的轰动。巨大参数量一方面成就了大模型，另一方面也成为大模型普及的掣肘，其中一个重要原因，是大模型推理态的高成本服务化部署。由于参数量巨大，大模型在推理时需要使用大量的计算资源，但受限于自回归解码 Token By Token 的生成方法，大模型推理时对计算资源的利用率较低。如何让模型在推理时充分利用计算资源，提升模型推理吞吐率，降低大模型推理成本，是一个棘手且重要的问题。

相关的生成式模型进行推理的方案中，Draft 生成和 Draft 验证两个关键阶段中，Draft 生成模型和 Draft 验证模型在协同方式和对应方式上具有强耦合的关系，该强耦合的关系在程序启动时已经固化，从而导致 Draft 生成模型和 Draft 验证模型在协同方式和对应方式上的灵活性较差，降低了模型推理的效率。

有鉴于此，本申请实施例提供了一种模型推理的方法，该方法可以提高模型推理的效率。

一种可能的实现方式中，本申请实施例提供的方法可以应用到云服务的场景中，由云服务场景中的云管理平台执行所述方法。为了便于描述，下面先结合图 1，对云服务的场景进行详细描述。

图 1 是适用于本申请实施例的一种云场景的示意性框图。如图 1 所示，该云场景中可以包括：云管理平台 110、互联网 120 以及客户端 130。

如图 1 所示，云管理平台 110 用于管理提供多个云服务的基础设施。基础设施包括多个云数据中心，每个云数据中心中包括多个服务器，每个服务器中分别包括云服务资源，为租户提供相应的云服务。

云管理平台 110 可以位于云数据中心，其可以提供访问接口（如界面或应用程序界面（application program interface，API））。租户可操作客户端 130 远程接入访问接口在云管理平台 110 注册云账号和密码，并登录云管理平台 110。云管理平台 110 对云账号和密码鉴权成功后，租户可进一步在云管理平台 110 付费选择并购买特定规格（处理器、内存、磁盘）的虚拟机。付费购买成功后，云管理平台 110 提供所购买的虚拟机的远程登录账号密码，客户端 130 可远程登录该虚拟机，在该虚拟机中安装并运行租户的应用。因此，租户可通过云管理平台 110 在云数据中心中创建、管理、登录和操作虚拟机。其中，虚拟机也可称为云服务器（elastic compute service, ECS）、弹性实例（不同的云服务提供商有不同的叫法）。

应理解，云服务的租户可以是个人、企业、学校、医院、行政机关等。

云管理平台 110 的功能包括但不限于用户控制台、计算管理服务、网络管理服务、存储管理服务、鉴权服务、镜像管理服务。用户控制台提供界面或 API 与租户交互，计算管理服务用于管理运行虚拟机

和容器的服务器以及裸金属服务器，网络管理服务用于管理网络服务（如网关、防火墙等），存储管理服务用于管理存储服务（如数据桶服务），鉴权服务用于管理租户的账号密码，镜像管理服务用于管理虚拟机镜像。租户使用客户端130，通过互联网120可以登录云管理平台110，对租用的云服务进行管理。

图2是本申请实施例提供的一种模型推理系统的示意性框图。如图2所示，在模型推理系统（也可以称为高效解码推理系统）推理启动前，系统管理员需要完成对Draft生成模型、Draft验证模型、计算资源、Draft池以及其他相关的配置，模型推理系统加载配置信息并启动服务。在整个服务期间，模型推理系统会接收外部的请求，通过Draft生成模型、Draft验证模型和Draft池的协同工作，生成文本并反馈给用户。同时，模型推理系统会将推理过程中的诸多运行信息进行统计和分析，并将相关结果发送给系统管理员以及自动调度模块，系统管理员可以手动或者自动完成系统配置更新，更新后的配置可以实时反馈给模型推理系统并实时生效。

图3是本申请实施例提供的另一种模型推理系统的示意性框图。如图3所示，该模型推理系统包括：Draft生成模型池310，Draft生成管理单元320，Draft验证模型池330，Draft验证管理单元340。下面分别对该模型推理系统中包括的各个部分进行详细描述。

1、Draft生成模型池310：

Draft生成模型池310中包括多个Draft生成模型，例如，Draft生成模型1-Draft生成模型M。具体的，该多个Draft生成模型可以通过热插拔的方式接入Draft生成模型池310。

本申请实施例对Draft生成模型池310中Draft生成模型的类型不做具体限定，其类型可以包括但不限于：基于统计的自然语言处理（natural language processing，NLP）模型、传统深度学习模型（例如，循环神经网络（recurrent neural networ，RNN）、长短期记忆网络（long short-Term memory，LSTM））、预训练模型（例如，基于变压器的双向编码器表示（bidirectional encoder representations from transformers，BERT）、生成型预训练转换器（generative pre-trained transformer，GPT））、通用大模型（例如，Llama、Pangu）、领域大模型、用户自定义模型。

每个Draft生成模型会将信息传给Draft生成管理单元320，这些信息包括但不限于：Draft生成模型的标签，Draft生成模型推理（生成Draft）所占用的资源情况，Draft生成模型产生的Draft。

本申请实施例对Draft生成模型产生的Draft的形式不做具体限定，该Draft的形式可以包括但不限于：N-Gram，短语，句子，段落等。

2、Draft生成管理单元320：

Draft生成管理单元320用于管理Draft生成模型池310中的Draft生成模型。具体的，Draft生成管理单元320可以结合当前模型推理系统对Draft的需求判断，以及当前推理资源的使用情况，向Draft生成模型池310发送信息，该信息用于控制Draft生成模型池310中的一个或多个Draft生成模型。

本申请实施例中，Draft生成管理单元320向Draft生成模型池310发送的信息可以包括但不限于：Draft生成模型的标签，对Draft生成模型进行操作的指令，用于Draft生成模型生成Draft的Prompt。其中，对Draft生成模型进行操作包括但不限于：激活、休眠、释放Draft生成模型等。

3、Draft验证模型池330：

Draft验证模型池330中包括多个Draft验证模型，例如，Draft验证模型1-Draft验证模型N。具体的，该多个Draft验证模型可以通过热插拔的方式接入Draft验证模型池330。Draft验证模型用于接收外部请求（包括prompts），形成Response（例如，generated content）并反馈给用户。

本申请实施例对Draft验证模型池330中Draft验证模型的类型不做具体限定，可以是任何规模的LLM生成模型。

每个Draft验证模型会将信息传给Draft验证管理单元340，这些信息包括但不限于：Draft验证模型的标签，Draft验证模型推理（做验证）所占用的资源情况，当前的Prompt，Draft生成模型产生的验证结果等信息。

本申请实施例中，Draft验证模型和Draft生成模型之间的对应方式有多种，对此不做具体限定。

下面列举几种可能的对应方式：

一个Draft生成模型对应一个Draft验证模型；

多个Draft生成模型对应一个Draft验证模型；

一个Draft生成模型对应多个Draft验证模型；

多个Draft生成模型对应多个Draft验证模型。

需要说明的是，上述 Draft 验证模型和 Draft 生成模型之间的对应方式（对应模式）之间可以实时自由切换，完成动态组合。

4、Draft 验证管理单元 340

Draft 验证管理单元 340 用于管理 Draft 验证模型池 330 中的 Draft 验证模型。例如，Draft 验证管理单元 340 主要用于将从 Draft 生成管理单元 320 接收到的 Draft 分发给相应的 Draft 验证模型。

下面以图 3 所示的模型推理系统为例，结合图 4，对本申请实施例提供的一种模型推理的方法进行详细描述。应理解，图 4 的例子仅仅是为了帮助本领域技术人员理解本申请实施例，而非要将申请实施例限制于图 4 所示例的具体数值或具体场景。本领域技术人员根据图 4 所给出的下面的例子，显然可以进行各种等价的修改或变化，这样的修改和变化也落入本申请实施例的范围内。

图 4 是本申请实施例提供的一种模型推理的方法的示意性流程图。如图 4 所示，该方法可以包括步骤 410-480，下面分别对步骤 410-480 进行详细描述。

步骤 410：用户完成 Draft 生成模型和 Draft 验证模型的设置。

本申请实施例中，用户可以设置 Draft 验证模型，以及与该 Draft 验证模型对应的 Draft 生成模型。

应理解，上述与 Draft 验证模型对应的 Draft 生成模型可以是用户自己设置的，或者也可以是系统根据用户输入的 Draft 验证模型，自动为其推荐与该 Draft 验证模型对应的 Draft 生成模型。

举例说明，如图 5 所示，用户选择的 Draft 验证模型为 Llama 13B，选择的 Draft 生成模型为 Draft 生成模型 1 和 Draft 生成模型 2。此时，模型推理系统如图 6 所示。

步骤 420：模型推理系统在运行时，接收到 prompts 后，会发送给 Draft 验证模型池 330 中的 Draft 验证模型 Llama 13B。

本申请实施例中，在运行时，模型推理系统接收到 prompts 后，会发送给 Draft 验证模型池 330 中的 Draft 验证模型，例如，发送给 Draft 验证模型池 330 中的 Llama 13B。

步骤 430：Draft 验证模型 Llama 13B 基于接收到的 prompts 生成 Draft 生成请求并发送给 Draft 验证管理单元 340。

本申请实施例中，Draft 验证模型 Llama 13B 接收到 prompts 后，会基于接收到的 prompts 生成 Draft 生成请求并发送给 Draft 验证管理单元 340。

步骤 440：Draft 验证管理单元 340 基于接收到的 Draft 生成请求生成一个新的 Draft 生成请求，并将该新的 Draft 生成请求发送给 Draft 生成管理单元 320。

本申请实施例中，Draft 验证管理单元 340 接收到 Draft 验证模型 Llama 13B 发送的 Draft 生成请求后，会在该 Draft 生成请求添加 Draft 生成数量等信息后生成一个新的 Draft 生成请求，并将该新的 Draft 生成请求发送给 Draft 生成管理单元 320。

步骤 450：Draft 生成管理单元 320 将新的 Draft 生成请求发送给 Draft 生成模型池 310 中与 Draft 验证模型 Llama 13B 对应的 Draft 生成模型。

本申请实施例中，Draft 生成管理单元 320 从 Draft 验证管理单元 340 接收到新的 Draft 生成请求后，可以将该新的 Draft 生成请求发送给 Draft 生成模型池 310 中与 Draft 验证模型 Llama 13B 对应的 Draft 生成模型，例如，Draft 生成模型池 310 中的 Draft 生成模型 1 和 Draft 生成模型 2。

步骤 460：Draft 生成模型 1 和 Draft 生成模型 2 各自根据接收到的新的 Draft 生成请求生成 Draft，并反馈给 Draft 生成管理单元 320。

本申请实施例中，Draft 生成模型 1 和 Draft 生成模型 2 接收到 Draft 生成管理单元 320 发送的新的 Draft 生成请求后，分别根据该新的 Draft 生成请求各自生成 Draft，并将各自生成的 Draft 反馈给 Draft 生成管理单元 320。

步骤 470：Draft 生成管理单元 320 收集到 Draft 后，会将收集的 Draft 通过 Draft 验证管理单元 340 发送至 Draft 验证模型 Llama 13B。

本申请实施例中，Draft 生成管理单元 320 分别接收 Draft 生成模型 1 和 Draft 生成模型 2 发送的 Draft 后，可以将其接收到的 Draft 发送给 Draft 验证管理单元 340，并由该 Draft 验证管理单元 340 将其接收到的 Draft 发送给 Draft 验证模型 Llama 13B。

步骤 480：Draft 验证模型 Llama 13B 对每条 prompt 对应的一个或多个 Draft 进行验证，形成新的 prompt。

本申请实施例中，Draft 验证模型 Llama 13B 可以对每条 prompt 对应的一个或多个 Draft 进行验证，

具体的，Draft 验证模型 Llama 13B 基于每条 prompt 对应的一个或多个 Draft 可以确认至少 1 个生成的 token，并将该确认的 token 拼接在相应的 prompt（接收到的 prompt）后，形成新的 prompt。

本申请实施例中，Draft 验证模型 Llama 13B 还可以将新的 prompt 再次发送给 Draft 验证模型 Llama 13B，多次迭代重复步骤 430-步骤 480 直至 Draft 验证模型 Llama 13B 基于输入的 prompts 完成内容生成，并反馈给用户。

上述技术方案中，Draft 生成模型和 Draft 验证模型的对应方式解耦，模型之间对应模式可以动态组合，任意配比，并基于模型推理系统的状态进行实时调整。

图 7 是本申请实施例提供的另一种模型推理系统的示意性框图。如图 7 所示，该模型推理系统包括：Draft 生成模型池 310，Draft 生成管理单元 320，Draft 验证模型池 330，Draft 验证管理单元 340，Draft 池 350。

应理解，有关 Draft 生成模型池 310，Draft 生成管理单元 320，Draft 验证模型池 330，以及 Draft 验证管理单元 340 的描述，请参考图 3 中的描述，此处不再赘述。下面重点对 Draft 池 350 的功能进行详细描述。

Draft 池 350 用于存储 Draft 生成模型生成的 Draft，以及每个 Draft 的属性标签信息。

本申请实施例对 Draft 的属性标签信息不做具体限定，该 Draft 的属性标签信息包括但不限于：Draft 对应的内容领域，Draft 对应的生成模型的标签，Draft 被调用的次数，Draft 被命中的次数，Draft 命中的 Prompt 列表等信息。

应理解，图 7 所示的模型推理系统是在图 3 的模型推理系统的基础上增加了 Draft 池 350，基于图 7 的模型推理系统，Draft 生成模型和 Draft 验证模型可以通过以下方式协同：

1)、同步协同：

Draft 验证模型发送 Draft 生成请求，激活对应的 Draft 生成模型产生 Draft，同时产生的 Draft 被记录在 Draft 池 350 中，方便后续模型或者请求复用。

2)、异步协同：

Draft 生成模型自我产生 Draft 并存储在 Draft 池 350 中，Draft 验证模型发送 Draft 生成请求，从 Draft 池 350 中获取相关 Draft。

3)、混合协同：

部分 Draft 生成模型&Draft 验证模型之间同步协同，部分 Draft 生成模型&Draft 验证模型之间异步协同。

下面以图 7 所示的模型推理系统为例，结合图 8，对本申请实施例提供的另一种模型推理的方法进行详细描述。应理解，图 8 的例子仅仅是为了帮助本领域技术人员理解本申请实施例，而非要将申请实施例限制于图 8 所示例的具体数值或具体场景。本领域技术人员根据图 8 所给出的下面的例子，显然可以进行各种等价的修改或变化，这样的修改和变化也落入本申请实施例的范围内。

图 8 是本申请实施例提供的另一种模型推理的方法的示意性流程图。如图 8 所示，该方法可以包括步骤 810-880，下面分别对步骤 810-880 进行详细描述。

步骤 810：用户完成 Draft 生成模型和 Draft 验证模型的设置。

本申请实施例中，用户可以设置 Draft 验证模型，以及与该 Draft 验证模型对应的 Draft 生成模型。

举例说明，如图 9 所示，用户选择的 Draft 验证模型为 Llama 13B，选择的 Draft 生成模型中，同步模型为 Draft 生成模型 1 和 Draft 生成模型 2，异步模型为 Draft 生成模型 3。此时，模型推理系统如图 10 所示。

步骤 820：模型推理系统在运行时，接收到 prompts 后，会发送给 Draft 验证模型池 330 中的 Draft 验证模型 Llama 13B。

与步骤 420 类似，具体的请参考步骤 420 中的描述，此处不再赘述。

步骤 830：Draft 验证模型 Llama 13B 基于接收到的 prompts 生成 Draft 生成请求并发送给 Draft 验证管理单元 340。

步骤 840：Draft 验证管理单元 340 基于接收到的 Draft 生成请求生成一个新的 Draft 生成请求，并将该新的 Draft 生成请求发送给 Draft 池 350。

本申请实施例中，Draft 验证管理单元 340 接收到 Draft 验证模型 Llama 13B 发送的 Draft 生成请求后，会在该 Draft 生成请求添加 Draft 生成数量等信息后生成一个新的 Draft 生成请求，并将该新的 Draft 生成

请求发送给 Draft 池 350。

步骤 850：Draft 池 350 接收到新的 Draft 生成请求后，获取对应的 Draft。

本申请实施例中，Draft 池 350 接收到 Draft 验证管理单元 340 发送的新的 Draft 生成请求后，可以获取到对应的 Draft。

具体的，Draft 池 350 同时执行如下两个步骤：

1、Draft 池 350 将新的 Draft 生成请求继续传递至 Draft 生成管理单元 320，Draft 生成管理单元 320 该新的 Draft 生成请求发送给 Draft 生成模型池 310 中与 Draft 验证模型 Llama 13B 对应的同步模型，例如，Draft 生成模型池 310 中的 Draft 生成模型 1 和 Draft 生成模型 2。Draft 生成模型 1 和 Draft 生成模型 2 各自根据接收到的新的 Draft 生成请求生成 Draft，并将其各自生成的 Draft 通过 Draft 生成管理单元 320 存储在 Draft 池 350 中；

2、Draft 池 350 在池内查找与该新的 Draft 生成请求相关的 Draft。

步骤 860：Draft 池 350 将获得的 Draft 通过 Draft 验证管理单元 340 发送至 Draft 验证模型 Llama 13B。

步骤 870：Draft 验证模型 Llama 13B 对每条 prompt 对应的一个或多个 Draft 进行验证，形成新的 prompt。

本申请实施例中，Draft 验证模型 Llama 13B 可以对每条 prompt 对应的一个或多个 Draft 进行验证，具体的，Draft 验证模型 Llama 13B 基于每条 prompt 对应的一个或多个 Draft 可以确认至少 1 个生成的 token，并将该确认的 token 拼接在相应的 prompt（接收到的 prompt）后，形成新的 prompt。

本申请实施例中，Draft 验证模型 Llama 13B 还可以将 Draft 的命中率等验证信息通过 Draft 验证管理单元 340 发送给 Draft 池 350，并由 Draft 池 350 记录该验证信息。

本申请实施例中，Draft 验证模型 Llama 13B 还可以将新的 prompt 再次发送给 Draft 验证模型 Llama 13B，多次迭代重复步骤 830-步骤 870 直至 Draft 验证模型 Llama 13B 基于输入的 prompts 完成内容生成，并反馈给用户。

步骤 880：当推理服务负载较低时，Draft 池 350 会基于收集的 Draft 的命中率等验证信息，激活 Draft 生成模型池 310 中与 Draft 验证模型 Llama 13B 对应的同步模型。

本申请实施例中，当推理服务负载较低时，Draft 池 350 会基于收集的 Draft 的命中率等验证信息，形成 Draft 生成请求，Draft 生成管理单元 320 会激活 Draft 生成模型池 310 中与 Draft 验证模型 Llama 13B 对应的同步模型，例如，Draft 生成模型 3。该 Draft 生成模型 3 会基于 Draft 生成请求产生对应的 Draft 并保存在 Draft 池 350 中，用于后续模型推理时使用。

上述技术方案中，Draft 生成模型和 Draft 验证模型可以实现完全解耦，模型的对应模式、协同方式可以动态组合，任意配比，可以基于模型推理系统的状态进行实时调整。并且，Draft 池保存了系统运行过程中产生的 Drafts，让 Draft 得以复用，在一定程度上可以减少产生 Draft 所占用的算力和时间。

图 11 是本申请实施例提供的另一种模型推理系统的示意性框图。如图 11 所示，该模型推理系统包括：Draft 生成模型池 310，Draft 生成管理单元 320，Draft 验证模型池 330，Draft 验证管理单元 340，Draft 池 350，Draft 分析单元 360，推理资源分析单元 370。

应理解，图 11 所示的模型推理系统是在图 7 的模型推理系统的基础上增加了 Draft 分析单元 360，推理资源分析单元 370，因此，有关 Draft 生成模型池 310，Draft 生成管理单元 320，Draft 验证模型池 330，以及 Draft 验证管理单元 340 的描述，请参考图 3 中的描述，有关 Draft 池 350 的描述，请参考图 7 中的描述，此处不再赘述。

1、Draft 分析单元 360：

Draft 分析单元 360 会对 Draft 池 350 中 Draft 的使用情况进行统计、分析与管理。

作为示例，Draft 分析单元 360 的主要功能包括但不限于：

a)、对 Draft 池 350 已有 Draft 进行分析，基于分析结果通知 Draft 生成管理单元 320 生成相关的 Draft；

b)、对 Draft 验证管理单元 340 传入的 prompt 进行分析，基于分析结果从 Draft 池 350 中通过多维信息，检索相关的 Draft，若缺少相关 Draft，基于分析结果通知 Draft 生成管理单元 320 生成相关的 Draft；

c)、接收来自 Draft 生成管理单元 320 的 Draft 信息，以及来自 Draft 验证管理单元 340 的验证信息，基于上述信息，完成对 Draft 池 350 中 Draft 的筛选、检索、修改、删除、新增等操作。

2、推理资源分析单元 370：

推理资源分析单元 370 主要用于分析当前推理资源的使用情况。

作为示例，推理资源分析单元 370 的主要功能包括但不限于：

a)、接受 Draft 验证管理单元 340 发送的 Draft 验证模型与请求信息，用于评估当前有多少算力、显存用于模型推理，基于当前推理资源状态、模型资源占用、请求状态，实现 Draft 数量的自适应调整；

b)、接受 Draft 生成管理单元 320 发送的 Draft 生成模型的信息，评估是否可以激活/休眠 Draft 生成模型，以及更新释放 Draft 生成模型后的资源信息；

c)、当前已申请推理计算资源是否需要部分释放，是否需要申请额外的推理计算资源。

下面以图 11 所示的模型推理系统为例，结合图 12，对本申请实施例提供的另一种模型推理的方法进行详细描述。应理解，图 12 的例子仅仅是为了帮助本领域技术人员理解本申请实施例，而非要将申请实施例限制于图 12 所示例的具体数值或具体场景。本领域技术人员根据图 12 所给出的下面的例子，显然可以进行各种等价的修改或变化，这样的修改和变化也落入本申请实施例的范围内。

图 12 是本申请实施例提供的另一种模型推理的方法的示意性流程图。如图 12 所示，该方法可以包括步骤 1210-1280，下面分别对步骤 1210-1280 进行详细描述。

步骤 1210：用户完成 Draft 生成模型和 Draft 验证模型的设置。

本申请实施例中，用户可以设置 Draft 验证模型，以及与该 Draft 验证模型对应的 Draft 生成模型。

举例说明，如图 13 所示，用户选择的 Draft 验证模型为 Llama 13B，选择的 Draft 生成模型中，同步模型为 Draft 生成模型 1 和 Draft 生成模型 2，异步模型为 Draft 生成模型 3。并且，图 13 中，用户还选择开启数据自适应功能以及 Draft 生成模型的自动调度功能。此时，模型推理系统如图 14 所示。

步骤 1220：模型推理系统在运行时，接收到 prompts 后，会发送给 Draft 验证模型池 330 中的 Draft 验证模型 Llama 13B。

与步骤 420 类似，具体的请参考步骤 420 中的描述，此处不再赘述。

步骤 1230：Draft 验证模型 Llama 13B 基于接收到的 prompts 生成 Draft 生成请求并发送给 Draft 验证管理单元 340。

步骤 1240：Draft 验证管理单元 340 基于接收到的 Draft 生成请求生成一个新的 Draft 生成请求，并将该新的 Draft 生成请求发送给 Draft 分析单元 360。

本申请实施例中，Draft 验证管理单元 340 接收到 Draft 验证模型 Llama 13B 发送的 Draft 生成请求后，基于推理资源分析单元 370 给出的资源利用率等信息，自动为每个 prompt 设定 Draft 生成数量等额外信息（数据自适应）后生成一个新的 Draft 生成请求，并将该新的 Draft 生成请求发送给 Draft 分析单元 360。

步骤 1250：Draft 分析单元 360 收到新的 Draft 生成请求后，获取对应的 Draft。

本申请实施例中，Draft 分析单元 360 接收到 Draft 验证管理单元 340 发送的新的 Draft 生成请求后，可以获取到对应的 Draft。

具体的，Draft 分析单元 360 同时执行如下两个步骤：

1、Draft 分析单元 360 将新的 Draft 生成请求继续传递至 Draft 生成管理单元 320，Draft 生成管理单元 320 该新的 Draft 生成请求发送给 Draft 生成模型池 310 中与 Draft 验证模型 Llama 13B 对应的同步模型，例如，Draft 生成模型池 310 中的 Draft 生成模型 1 和 Draft 生成模型 2。Draft 生成模型 1 和 Draft 生成模型 2 各自根据接收到的新的 Draft 生成请求生成 Draft，并将其各自生成的 Draft 通过 Draft 生成管理单元 320 发送给 Draft 分析单元 360；

2、Draft 分析单元 360 从领域、语义等多个角度生成 prompts 的多维分析信息，并基于该 prompts 的多维分析信息在 Draft 池 350 内查找与该新的 Draft 生成请求相关的 Draft。

步骤 1260：Draft 分析单元 360 将获得的 Draft 通过 Draft 验证管理单元 340 发送至 Draft 验证模型 Llama 13B。

步骤 1270：Draft 验证模型 Llama 13B 对每条 prompt 对应的一个或多个 Draft 进行验证，形成新的 prompt。

本申请实施例中，Draft 验证模型 Llama 13B 可以对每条 prompt 对应的一个或多个 Draft 进行验证，具体的，Draft 验证模型 Llama 13B 基于每条 prompt 对应的一个或多个 Draft 可以确认至少 1 个生成的 token，并将该确认的 token 拼接在相应的 prompt（接收到的 prompt）后，形成新的 prompt。

本申请实施例中，Draft 验证模型 Llama 13B 还可以将 Draft 的命中率等验证信息通过 Draft 验证管理单元 340 发送给 Draft 分析单元 360，并由 Draft 分析单元 360 记录该验证信息。

本申请实施例中，Draft 验证模型 Llama 13B 还可以将新的 prompt 再次发送给 Draft 验证模型 Llama

13B，多次迭代重复步骤 1230-步骤 1270 直至 Draft 验证模型 Llama 13B 基于输入的 prompts 完成内容生成，并反馈给用户。

步骤 1280：当推理服务负载较低时，Draft 分析单元 360 会基于收集的 Draft 的命中率等验证信息，激活 Draft 生成模型池 310 中与 Draft 验证模型 Llama 13B 对应的同步模型。

本申请实施例中，当推理服务负载较低时，Draft 分析单元 360 会基于收集的 Draft 的命中率等验证信息，形成 Draft 生成请求，Draft 生成管理单元 320 会结合当前推理资源分析单元 370 给出的推理资源利用等信息，激活 Draft 生成模型池 310 中与 Draft 验证模型 Llama 13B 对应的同步模型，例如，Draft 生成模型 3。该 Draft 生成模型 3 会基于 Draft 生成请求产生对应的 Draft 并保存在 Draft 池 350 中，用于后续模型推理时使用。

本申请实施例中，如果 Draft 生成管理单元 320 要求激活 Draft 生成模型 3，而当前没有充足的推理资源时，推理资源分析单元 370 会新的推理资源供 Draft 生成模型 3 产生 Draft。

本申请实施例中，如果推理资源分析单元 370 给出的推理资源利用信息指示当前有未利用的推理资源，且 Draft 生成模型 3 未被激活，则推理资源分析单元 370 会自动完成推理资源的释放。

上述技术方案中，一方面产生的 Draft 的数量可以基于当前资源的状态实现自适应调整；另一方面，模型推理系统可以基于不同状态完成自我调整，包括调整 Draft 生成模型和 Draft 验证模型之间的对应方式和协同方式、Draft 池的自我维护、推理资源的申请和释放等；另一方面，由于 Draft 分析单元 360 从领域、语义等多个角度生成 prompts 的多维分析信息，并基于该 prompts 的多维分析信息在 Draft 池 350 内查找与 Draft 生成请求相关的 Draft，使得 Draft 召回更精准，在验证阶段带来更高的 Draft 命中率。

图 15 是本申请实施例提供的另一种模型推理系统的示意性框图。如图 15 所示，该模型推理系统包括：Draft 生成模型池 310，Draft 生成管理单元 320，Draft 验证模型池 330，Draft 验证管理单元 340，Draft 池 350，Draft 分析单元 360，推理资源分析单元 370。

应理解，图 15 所示的模型推理系统是在图 11 所示的模型推理系统的基础上，使得 Draft 池 350 支持多租模式，Draft 池 350 主要由两部分构成，一个为预置 Draft 池（也可以称为共享 Draft 池），这个共享 Draft 池可以在多个租户之间共享，另一个为客户的私有 Draft 池，每个客户的 Draft 池相互隔离。不同用户在配置各自的推理服务时，可以选择是否使用共享 Draft 池进行冷启动。服务启动后，系统会为每个用户划分一个独立的 Draft 池，每个用户的推理服务过程中所产生的 Draft，会写入各自的 Draft 池中。

下面以图 15 所示的模型推理系统为例，结合图 16，对本申请实施例提供的另一种模型推理的方法进行详细描述。应理解，图 16 的例子仅仅是为了帮助本领域技术人员理解本申请实施例，而非要将申请实施例限制于图 16 所示例的具体数值或具体场景。本领域技术人员根据图 16 所给出的下面的例子，显然可以进行各种等价的修改或变化，这样的修改和变化也落入本申请实施例的范围内。

图 16 是本申请实施例提供的一种模型推理的方法的示意性流程图。如图 16 所示，该方法可以包括步骤 1610-1680，下面分别对步骤 1610-1680 进行详细描述。

步骤 1610：用户 A 完成 Draft 生成模型和 Draft 验证模型的设置。

举例说明，如图 17 所示，用户 A 选择的 Draft 验证模型为 Llama 13B，选择的 Draft 生成模型中，同步模型为 Draft 生成模型 1 和 Draft 生成模型 2，异步模型为 Draft 生成模型 3。并且，图 17 中，用户 A 还选择开启数据自适应功能，开启 Draft 生成模型的自动调度功能，以及开启使用 Draft 池 350 中的共享 Draft 池。此时，模型推理系统如图 18 所示，图 18 中的 Draft 池 350 中包括共享 Draft 池以及用户 A 的私有 Draft 池。

步骤 1620：模型推理系统在运行时，接收到 prompts 后，会发送给 Draft 验证模型池 330 中的 Draft 验证模型 Llama 13B。

与步骤 420 类似，具体的请参考步骤 420 中的描述，此处不再赘述。

步骤 1630：Draft 验证模型 Llama 13B 基于接收到的 prompts 生成 Draft 生成请求并发送给 Draft 验证管理单元 340。

步骤 1640：Draft 验证管理单元 340 基于接收到的 Draft 生成请求生成一个新的 Draft 生成请求，并将该新的 Draft 生成请求发送给 Draft 分析单元 360。

步骤 1650：Draft 分析单元 360 收到新的 Draft 生成请求后，获取对应的 Draft。

本申请实施例中，Draft 分析单元 360 接收到 Draft 验证管理单元 340 发送的新的 Draft 生成请求后，可以获取到对应的 Draft。

具体的，Draft 分析单元 360 同时执行如下三个步骤：

1、Draft 分析单元 360 将新的 Draft 生成请求继续传递至 Draft 生成管理单元 320，Draft 生成管理单元 320 该新的 Draft 生成请求发送给 Draft 生成模型池 310 中与 Draft 验证模型 Llama 13B 对应的同步模型，例如，Draft 生成模型池 310 中的 Draft 生成模型 1 和 Draft 生成模型 2。Draft 生成模型 1 和 Draft 生成模型 2 各自根据接收到的新的 Draft 生成请求生成 Draft，并将其各自生成的 Draft 通过 Draft 生成管理单元 320 发送给 Draft 分析单元 360；

2、Draft 分析单元 360 从领域、语义等多个角度生成 prompts 的多维分析信息，并基于该 prompts 的多维分析信息在 Draft 池 350 中共享 Draft 池内查找与该新的 Draft 生成请求相关的 Draft；

3、Draft 分析单元 360 从领域、语义等多个角度生成 prompts 的多维分析信息，并基于该 prompts 的多维分析信息在 Draft 池 350 中用户 A 的私有 Draft 池内查找与该新的 Draft 生成请求相关的 Draft。

步骤 1660：Draft 分析单元 360 将获得的 Draft 通过 Draft 验证管理单元 340 发送至 Draft 验证模型 Llama 13B。

步骤 1670：Draft 验证模型 Llama 13B 对每条 prompt 对应的一个或多个 Draft 进行验证，形成新的 prompt。

本申请实施例中，Draft 验证模型 Llama 13B 可以对每条 prompt 对应的一个或多个 Draft 进行验证，具体的，Draft 验证模型 Llama 13B 基于每条 prompt 对应的一个或多个 Draft 可以确认至少 1 个生成的 token，并将该确认的 token 拼接在相应的 prompt（接收到的 prompt）后，形成新的 prompt。

本申请实施例中，Draft 验证模型 Llama 13B 还可以将 Draft 的命中率等验证信息通过 Draft 验证管理单元 340 发送给 Draft 分析单元 360，并由 Draft 分析单元 360 记录该验证信息。该 Draft 分析单元 360 会基于验证信息，对当前获得的 Draft 进行分析，并基于分析结果完成用户 A 的私有 Draft 池中相关 Draft 的操作。

本申请实施例中，Draft 验证模型 Llama 13B 还可以将新的 prompt 再次发送给 Draft 验证模型 Llama 13B，多次迭代重复步骤 1630-步骤 1670 直至 Draft 验证模型 Llama 13B 基于输入的 prompts 完成内容生成，并反馈给用户 A。

步骤 1680：当推理服务负载较低时，Draft 分析单元 360 会基于收集的 Draft 的命中率等验证信息，激活 Draft 生成模型池 310 中与 Draft 验证模型 Llama 13B 对应的同步模型。

本申请实施例中，当推理服务负载较低时，Draft 分析单元 360 会基于收集的 Draft 的命中率等验证信息，形成 Draft 生成请求，Draft 生成管理单元 320 会结合当前推理资源分析单元 370 给出的推理资源利用等信息，激活 Draft 生成模型池 310 中与 Draft 验证模型 Llama 13B 对应的同步模型，例如，Draft 生成模型 3。该 Draft 生成模型 3 会基于 Draft 生成请求产生对应的 Draft 并保存在 Draft 池 350 的用户 A 的私有 Draft 池中，用于后续模型推理时使用。

同样的，如果 Draft 生成管理单元 320 要求激活 Draft 生成模型 3，而当前没有充足的推理资源时，推理资源分析单元 370 会新的推理资源供 Draft 生成模型 3 产生 Draft。如果推理资源分析单元 370 给出的推理资源利用信息指示当前有未利用的推理资源，且 Draft 生成模型 3 未被激活，则推理资源分析单元 370 会自动完成推理资源的释放。

可选的，在一些实施例中，假设用户 B 也完成了系统配置。例如，如图 19 所示，用户 B 选择的 Draft 验证模型为 Llama 7B，选择的 Draft 生成模型中，同步模型为 Draft 生成模型 1 和 Draft 生成模型 4，异步模型为 Draft 生成模型 5。并且，图 19 中，用户 B 也选择了开启数据自适应功能，开启 Draft 生成模型的自动调度功能，以及开启使用 Draft 池 350 中的共享 Draft 池。此时，模型推理系统如图 20 所示，图 20 中的 Draft 池 350 中包括共享 Draft 池，用户 A 的私有 Draft 池以及用户 B 的私有 Draft 池。

需要说明的是，上述用户 B 也可以按照用户 A 执行的步骤 1610-1680 完成模型的推理过程。不同之处在于用户 B 的 Draft 验证模型为 Llama 7B，同步模型为 Draft 生成模型 1 和 Draft 生成模型 4，其中 Draft 生成模型 1 与用户 A 共享，异步模型为 Draft 生成模型 5，用户 B 在推理过程中，Draft 分析单元 360 可以读取 Draft 池 350 中的共享 Draft 池，但仅能对 Draft 池 350 中用户 B 的私有 Draft 池进行读写等操作。

上述技术方案中，Draft 池中的共享 Draft 池可以为多租户共享，帮助用户完成冷启动时的推理加速，Draft 池中不同用户的私有 Draft 池之间相互独立，这样可以独立维护不同用户在推理服务运行过程中产生的 Draft。

上文结合图 1 至图 20，详细描述了本申请实施例提供模型推理的方法，下面将结合图 21-图 24，详细描述本申请装置的实施例。应理解，方法实施例的描述与装置实施例的描述相互对应，因此，未详细描述的部分可以参见前面方法实施例。

图 21 是本申请实施例提供的一种模型推理的装置 2100 的示意性框图。该装置 2100 可以通过软件、硬件或者两者的结合实现。本申请实施例提供的装置 2100 可以实现本申请实施例所示的方法流程，该装置 2100 包括：接收模块 2110，处理模块 2120，反馈模块 2130，其中，接收模块 2110 用于接收第一用户输入的文本提示信息，该文本提示信息用于生成文本信息；处理模块 2120 用于从 Draft 验证模型池中确定至少一个 Draft 验证模型，该 Draft 验证模型池中包括多个 Draft 验证模型；从 Draft 生成模型池中确定与该至少一个 Draft 验证模型对应的至少一个 Draft 生成模型，其中，该 Draft 生成模型池中包括多个 Draft 生成模型，该至少一个 Draft 生成模型包括 Draft 同步生成模型和 Draft 异步生成模型；根据该文本提示信息获得 Draft 生成请求信息，将该 Draft 生成请求信息发送给该 Draft 同步生成模型，该 Draft 同步生成模型用于根据该 Draft 生成请求信息生成 Draft；和/或从 Draft 池中获取与该文本提示信息相关的 Draft，该 Draft 池中包括该 Draft 异步生成模型生成的 Draft；通过该至少一个 Draft 验证模型对该 Draft 同步生成模型生成的 Draft 和/或从 Draft 池中获取的 Draft 进行验证得到验证结果；反馈模块 2130 用于根据该验证结果生成该文本信息并反馈给该第一用户。

可选的，该至少一个 Draft 验证模型和该至少一个 Draft 生成模型之间的对应模式包括：一对一，一对多，多对一，多对多。

可选的，该装置 2100 还包括：存储模块，用于将该 Draft 同步生成模型生成的 Draft 存储在该 Draft 池中。

可选的，该 Draft 池中还存储有 Draft 的属性信息。

可选的，该处理模块 2120 具体用于：根据该 Draft 池中存储的 Draft 的属性信息，从该 Draft 池中获取与该文本提示信息相关的 Draft。

可选的，该处理模块 2120 还用于根据该 Draft 池中存储的 Draft 的属性信息，对该 Draft 池中存储的 Draft 进行检索和/或更新。

可选的，该处理模块 2120 还用于根据该 Draft 池中存储的 Draft 的属性信息，对该 Draft 生成模型池中包括的 Draft 生成模型进行操作，该操作包括：激活，休眠，释放。

可选的，该 Draft 的属性信息包括以下中的至少一种：该 Draft 的领域、内容、对应的 Draft 生成模型的标签、被调用的次数、命中的次数、命中的文本提示信息的列表。

可选的，该 Draft 生成请求信息还包括生成 Draft 的数量，该生成 Draft 的数量是根据模型推理资源的使用情况确定的，该模型推理资源用于该 Draft 生成模型池中的 Draft 生成模型产生 Draft 和/或该 Draft 验证模型池中的 Draft 验证模型执行验证过程。

可选的，该处理模块 2120 还用于在推理负载小于预设阈值的情况下，激活该 Draft 异步生成模型生成 Draft；存储模块，还用于将该 Draft 异步生成模型生成的 Draft 存储在该 Draft 池中。

可选的，该处理模块 2120 还用于根据模型推理资源的使用情况，自动申请/释放该模型推理资源。

可选的，该 Draft 池中包括共享 Draft 池和第一用户的私有 Draft 池，该处理模块 2120 具体用于：从该共享 Draft 池和/或该第一用户的私有 Draft 池中获取与该文本提示信息相关的 Draft。

可选的，该存储模块具体用于：将该 Draft 同步生成模型生成的 Draft 存储在第一用户的私有 Draft 池中。

可选的，该存储模块，还用于将该 Draft 异步生成模型生成的 Draft 存储在第一用户的私有 Draft 池中。

这里的装置 2100 可以以功能模块的形式体现。这里的术语“模块”可以通过软件和/或硬件形式实现，对此不作具体限定。

例如，“模块”可以是实现上述功能的软件程序、硬件电路或二者结合。示例性的，接下来以接收模块 2110 为例，介绍接收模块 2110 的实现方式。类似的，其他模块，例如处理模块 2120，反馈模块 2130，存储模块的实现方式可以参考接收模块 2110 的实现方式。

接收模块 2110 作为软件功能单元的一种举例，接收模块 2110 可以包括运行在计算实例上的代码。其中，计算实例可以包括物理主机（计算设备）、虚拟机、容器中的至少一种。进一步地，上述计算实例可以是一台或者多台。例如，接收模块 2110 可以包括运行在多个主机/虚拟机/容器上的代码。需要说明的是，用于运行该代码的多个主机/虚拟机/容器可以分布在相同的区域（region）中，也可以分布在不

同的 region 中。进一步地，用于运行该代码的多个主机/虚拟机/容器可以分布在相同的可用区（availability zone，AZ）中，也可以分布在不同的 AZ 中，每个 AZ 包括一个数据中心或多个地理位置相近的数据中心。其中，通常一个 region 可以包括多个 AZ。

同样，用于运行该代码的多个主机/虚拟机/容器可以分布在同一个虚拟私有云（virtual private cloud，VPC）中，也可以分布在多个 VPC 中。其中，通常一个 VPC 设置在一个 region 内，同一 region 内两个 VPC 之间，以及不同 region 的 VPC 之间跨区通信需在每个 VPC 内设置通信网关，经通信网关实现 VPC 之间的互连。

接收模块 2110 作为硬件功能单元的一种举例，接收模块 2110 可以包括至少一个计算设备，如服务器等。或者，接收模块 2110 也可以是利用专用集成电路（application-specific integrated circuit，ASIC）实现、或可编程逻辑器件（programmable logic device，PLD）实现的设备等。其中，上述 PLD 可以是复杂程序逻辑器件（complex programmable logical device，CPLD）、现场可编程门阵列（field-programmable gate array，FPGA）、通用阵列逻辑（generic array logic，GAL）或其任意组合实现。

接收模块 2110 包括的多个计算设备可以分布在相同的 region 中，也可以分布在不同的 region 中。接收模块 2110 包括的多个计算设备可以分布在相同的 AZ 中，也可以分布在不同的 AZ 中。同样，接收模块 2110 包括的多个计算设备可以分布在同一个 VPC 中，也可以分布在多个 VPC 中。其中，所述多个计算设备可以是服务器、ASIC、PLD、CPLD、FPGA 和 GAL 等计算设备的任意组合。

因此，在本申请的实施例中描述的各示例的模块，能够以电子硬件、或者计算机软件和电子硬件的结合来实现。这些功能究竟以硬件还是软件方式来执行，取决于技术方案的特定应用和设计约束条件。专业技术人员可以对每个特定的应用来使用不同方法来实现所描述的功能，但是这种实现不应认为超出本申请的范围。

需要说明的是：上述实施例提供的装置在执行上述方法时，仅以上述各功能模块的划分进行举例说明，实际应用中，可以根据需要而将上述功能分配由不同的功能模块完成，即将装置的内部结构划分成不同的功能模块，以完成以上描述的全部或者部分功能。例如，接收模块 2110 可以用于执行上述方法中的任意步骤，处理模块 2120 可以用于执行上述方法中的任意步骤，反馈模块 2130 可以用于执行上述方法中的任意步骤，存储模块可以用于执行上述方法中的任意步骤。接收模块 2110、处理模块 2120、反馈模块 2130、存储模块负责实现的步骤可根据需要指定，通过接收模块 2110、处理模块 2120、反馈模块 2130、存储模块分别实现上述方法中不同的步骤来实现上述装置的全部功能。

另外，上述实施例提供的装置与方法实施例属于同一构思，其具体实现过程详见上文中的方法实施例，这里不再赘述。

本申请实施例提供的方法可以由计算设备执行，该计算设备也可以被称为计算机系统。包括硬件层、运行在硬件层之上的操作系统层，以及运行在操作系统层上的应用层。该硬件层包括处理单元、内存和内存控制单元等硬件，随后对该硬件的功能和结构进行详细说明。该操作系统是任意一种或多种通过进程（process）实现业务处理的计算机操作系统，例如，Linux 操作系统、Unix 操作系统、Android 操作系统、iOS 操作系统或 windows 操作系统等。该应用层包含浏览器、通讯录、文字处理软件、即时通信软件等应用程序。并且，可选地，该计算机系统是智能手机等手持设备，或个人计算机等终端设备，本申请并未特别限定，只要能够通过本申请实施例提供的方法即可。本申请实施例提供的方法的执行主体可以是计算设备，或者，是计算设备中能够调用程序并执行程序的功能模块。

下面结合图 22，对本申请实施例提供的一种计算设备进行详细描述。

图 22 是本申请实施例提供的一种计算设备 1500 的架构示意图。该计算设备 1500 可以是服务器或者计算机或者其他具有计算能力的设备。图 22 所示的计算设备 1500 包括：至少一个处理器 1510 和存储器 1520。

应理解，本申请不限定计算设备 1500 中的处理器、存储器的个数。

处理器 1510 执行存储器 1520 中的指令，使得计算设备 1500 实现本申请提供的方法。或者，处理器 1510 执行存储器 1520 中的指令，使得计算设备 1500 实现本申请提供的各功能模块，从而实现本申请提供的方法。

可选地，计算设备 1500 还包括通信接口 1530。通信接口 1530 使用例如但不限于网络接口卡、收发器一类的收发模块，来实现计算设备 1500 与其他设备或通信网络之间的通信。

可选地，计算设备 1500 还包括系统总线 1540，其中，处理器 1510、存储器 1520 和通信接口 1530 分

别与系统总线 1540 连接。处理器 1510 能够通过系统总线 1540 访问存储器 1520，例如，处理器 1510 能够通过系统总线 1540 在存储器 1520 中进行数据读写或代码执行。该系统总线 1540 是快捷外设部件互连标准（peripheral component interconnect express，PCI）总线或扩展工业标准结构（extended industry standard architecture，EISA）总线等。所述系统总线 1540 分为地址总线、数据总线、控制总线等。为便于表示，图 22 中仅用一条粗线表示，但并不表示仅有一根总线或一种类型的总线。

一种可能的实现方式，处理器 1510 的功能主要是解释计算机程序的指令（或者说，代码）以及处理计算机软件中的数据。其中，该计算机程序的指令以及计算机软件中的数据能够保存在存储器 1520 或者缓存 1516 中。

可选地，处理器 1510 可能是集成电路芯片，具有信号的处理能力。作为示例而非限定，处理器 1510 是通用处理器、数字信号处理器（digital signal processor，DSP）、专用集成电路（application specific integrated circuit，ASIC）、现成可编程门阵列（field programmable gate array，FPGA）或者其他可编程逻辑器件、分立门或者晶体管逻辑器件、分立硬件组件。其中，通用处理器是微处理器等。例如，该处理器 1510 是中央处理单元（central processing unit，CPU）。

可选地，每个处理器 1510 包括至少一个处理单元 1512 和内存控制单元 1514。

可选地，处理单元 1512 也称为核心（core）或内核，是处理器最重要的组成部分。处理单元 1512 是由单晶硅以一定的生产工艺制造出来的，处理器所有的计算、接受命令、存储命令、处理数据都由核心执行。处理单元分别独立地运行程序指令，利用并行计算的能力加快程序的运行速度。各种处理单元都具有固定的逻辑结构，例如，处理单元包括例如，一级缓存、二级缓存、执行单元、指令级单元和总线接口等逻辑单元。

一种实现举例，内存控制单元 1514 用于控制存储器 1520 与处理单元 1512 之间的数据交互。具体地说，内存控制单元 1514 从处理单元 1512 接收内存访问请求，并基于该内存访问请求控制针对内存的访问。作为示例而非限定，内存控制单元是内存管理单元（memory management unit，MMU）等器件。

一种实现举例，各内存控制单元 1514 通过系统总线进行针对存储器 1520 的寻址。并且在系统总线中配置仲裁器（图 22 中未示出），该仲裁器负责处理和协调多个处理单元 1512 的竞争访问。

一种实现举例，处理单元 1512 和内存控制单元 1514 通过芯片内部的连接线，例如地址线，通信连接，从而实现处理单元 1512 和内存控制单元 1514 之间的通信。

可选地，每个处理器 1510 还包括缓存 1516，其中，缓存是数据交换的缓冲区（称作 cache）。当处理单元 1512 要读取数据时，会首先从缓存中查找需要的数据，如果找到了则直接执行，找不到的话则从存储器中找。由于缓存的运行速度比存储器快得多，故缓存的作用就是帮助处理单元 1512 更快地运行。

存储器 1520 能够为计算设备 1500 中的进程提供运行空间，例如，存储器 1520 中保存用于生成进程的计算机程序（具体地说，是程序的代码）。计算机程序被处理器运行而生成进程后，处理器在存储器 1520 中为该进程分配对应的存储空间。进一步的，上述存储空间进一步包括文本段、初始化数据段、位初始化数据段、栈段、堆段等等。存储器 1520 在上述进程对应的存储空间中保存进程运行期间产生的数据，例如，中间数据，或过程数据等等。

可选地，存储器也称为内存，其作用是用于暂时存放处理器 1510 中的运算数据，以及与硬盘等外部存储器交换的数据。只要计算机在运行中，处理器 1510 就会把需要运算的数据调到内存中进行运算，当运算完成后处理单元 1512 再将结果传送出来。

作为示例而非限定，存储器 1520 是易失性存储器或非易失性存储器，或可包括易失性和非易失性存储器两者。其中，非易失性存储器是只读存储器（read-only memory，ROM）、可编程只读存储器（programmable ROM，PROM）、可擦除可编程只读存储器（erasable PROM，EPROM）、电可擦除可编程只读存储器（electrically EPROM，EEPROM）或闪存。易失性存储器是随机存取存储器（random access memory，RAM），其用作外部高速缓存。通过示例性但不是限制性说明，许多形式的 RAM 可用，例如静态随机存取存储器（static RAM，SRAM）、动态随机存取存储器（dynamic RAM，DRAM）、同步动态随机存取存储器（synchronous DRAM，SDRAM）、双倍数据速率同步动态随机存取存储器（double data rate SDRAM，DDR SDRAM）、增强型同步动态随机存取存储器（enhanced SDRAM，ESDRAM）、同步连接动态随机存取存储器（synchlink DRAM，SLDRAM）和直接内存总线随机存取存储器（direct rambus RAM，DR RAM）。应注意，本文描述的系统和方法的存储器 1520 旨在包括但不限于这些和任意其它适合类型的存储器。

以上列举的计算设备 1500 的结构仅为示例性说明，本申请并未限定于此，本申请实施例的计算设备 1500 包括现有技术中计算机系统中的各种硬件，例如，计算设备 1500 还包括除存储器 1520 以外的其他存储器，例如，磁盘存储器等。本领域的技术人员应当理解，计算设备 1500 还可以包括实现正常运行所必须的其他器件。同时，根据具体需要，本领域的技术人员应当理解，上述计算设备 1500 还可包括实现其他附加功能的硬件器件。此外，本领域的技术人员应当理解，上述计算设备 1500 也可仅仅包括实现本申请实施例所必须的器件，而不必包括图 22 中所示的全部器件。

本申请实施例还提供了一种计算设备集群。该计算设备集群包括至少一台计算设备。该计算设备可以是服务器。在一些实施例中，计算设备也可以是台式机、笔记本电脑或者智能手机等终端设备。

如图 23 所示，所述计算设备集群包括至少一个计算设备 1500。计算设备集群中的一个或多个计算设备 1500 中的存储器 1520 中可以存有相同的用于执行上述方法的指令。

在一些可能的实现方式中，该计算设备集群中的一个或多个计算设备 1500 中的存储器 1520 也可以分别存有用于执行上述方法的部分指令。换言之，一个或多个计算设备 1500 的组合可以共同执行上述方法的指令。

需要说明的是，计算设备集群中的不同的计算设备 1500 中的存储器 1520 可以存储不同的指令，分别用于执行上述装置的部分功能。也即，不同的计算设备 1500 中的存储器 1520 存储的指令可以实现将上述装置中的一个或多个模块的功能。

在一些可能的实现方式中，计算设备集群中的一个或多个计算设备可以通过网络连接。其中，所述网络可以是广域网或局域网等等。图 24 示出了一种可能的实现方式。如图 24 所示，两个计算设备 1500A 和 1500B 之间通过网络进行连接。具体地，通过各个计算设备中的通信接口与所述网络进行连接。

应理解，图 24 中示出的计算设备 1500A 的功能也可以由多个计算设备 1500 完成。同样，计算设备 1500B 的功能也可以由多个计算设备 1500 完成。

本实施例中，还提供了一种包含指令的计算机程序产品，所述计算机程序产品可以是包含指令的，能够运行在计算设备上或被储存在任何可用介质中的软件或程序产品。当其在计算设备上运行时，使得计算设备执行上述所提供的方法，或者使得该计算设备实现上述提供的装置的功能。

本实施例中，还提供了一种计算机可读存储介质，计算机可读存储介质可以是计算设备能够存储的任何可用介质或者是包含一个或多个可用介质的数据中心等数据存储设备。所述可用介质可以是磁性介质，（例如，软盘、硬盘、磁带）、光介质（例如，DVD）、或者半导体介质（例如固态硬盘）等。该计算机可读存储介质包括指令，当计算机可读存储介质中的指令在计算设备上被执行时，使得计算设备执行上述所提供的方法。

应理解，在本申请的各种实施例中，上述各过程的序号的大小并不意味着执行顺序的先后，各过程的执行顺序应以其功能和内在逻辑确定，而不应对本申请实施例的实施过程构成任何限定。

本领域普通技术人员可以意识到，结合本文中所公开的实施例描述的各示例的单元及算法步骤，能够以电子硬件、或者计算机软件和电子硬件的结合来实现。这些功能究竟以硬件还是软件方式来执行，取决于技术方案的特定应用和设计约束条件。专业技术人员可以对每个特定的应用来使用不同方法来实现所描述的功能，但是这种实现不应认为超出本申请的范围。

所属领域的技术人员可以清楚地了解到，为描述的方便和简洁，上述描述的系统、装置和单元的具体工作过程，可以参考前述方法实施例中的对应过程，在此不再赘述。

在本申请所提供的几个实施例中，应该理解到，所揭露的系统、装置和方法，可以通过其它的方式实现。例如，以上所描述的装置实施例仅仅是示意性的，例如，所述单元的划分，仅仅为一种逻辑功能划分，实际实现时可以有另外的划分方式，例如多个单元或组件可以结合或者可以集成到另一个系统，或一些特征可以忽略，或不执行。另一点，所显示或讨论的相互之间的耦合或直接耦合或通信连接可以是通过一些接口，装置或单元的间接耦合或通信连接，可以是电性，机械或其它的形式。

所述作为分离部件说明的单元可以是或者也可以不是物理上分开的，作为单元显示的部件可以是或者也可以不是物理单元，即可以位于一个地方，或者也可以分布到多个网络单元上。可以根据实际的需要选择其中的部分或者全部单元来实现本实施例方案的目的。

另外，在本申请各个实施例中的各功能单元可以集成在一个处理单元中，也可以是各个单元单独物理存在，也可以两个或两个以上单元集成在一个单元中。

所述功能如果以软件功能单元的形式实现并作为独立的产品销售或使用时，可以存储在一个计算机可读取存储介质中。基于这样的理解，本申请的技术方案本质上或者说对现有技术做出贡献的部分或者该技术方案的部分可以以软件产品的形式体现出来，该计算机软件产品存储在一个存储介质中，包括若干指令用以使得一台计算机设备（可以是个人计算机，服务器，或者网络设备等）执行本申请各个实施例所述方法的全部或部分步骤。而前述的存储介质包括：U 盘、移动硬盘、只读存储器（read-only memory，ROM）、随机存取存储器（random access memory，RAM）、磁碟或者光盘等各种可以存储程序代码的介质。

以上所述，仅为本申请的具体实施方式，但本申请的保护范围并不局限于此，任何熟悉本技术领域的技术人员在本申请揭露的技术范围内，可轻易想到变化或替换，都应涵盖在本申请的保护范围之内。因此，本申请的保护范围应以所述权利要求的保护范围为准。

权 利 要 求 书

1. 一种模型推理的方法，其特征在于，所述方法包括：

从 Draft 验证模型池中确定至少一个 Draft 验证模型，所述 Draft 验证模型池中包括多个 Draft 验证模型；

从 Draft 生成模型池中确定与所述至少一个 Draft 验证模型对应的至少一个 Draft 生成模型，其中，所述 Draft 生成模型池中包括多个 Draft 生成模型；

接收第一用户输入的文本提示信息，所述文本提示信息用于生成文本信息；

从 Draft 池中获取与所述文本提示信息相关的目标 Draft，所述 Draft 池中存储有多个 Draft，所述多个 Draft 包括所述至少一个 Draft 生成模型生成的 Draft；

通过所述至少一个 Draft 验证模型对所述目标 Draft 进行验证，并得到验证结果；

根据所述验证结果生成所述文本信息并反馈给所述第一用户。

2. 根据权利要求 1 所述的方法，其特征在于，所述至少一个 Draft 生成模型包括 Draft 同步生成模型，所述方法还包括：

根据所述文本提示信息获得 Draft 生成请求信息；

将所述 Draft 生成请求信息发送给所述 Draft 同步生成模型，所述 Draft 同步生成模型用于根据所述 Draft 生成请求信息生成 Draft；

将所述 Draft 同步生成模型生成的 Draft 存储在所述 Draft 池中。

3. 根据权利要求 1 或 2 所述的方法，其特征在于，所述至少一个 Draft 生成模型包括 Draft 异步生成模型，所述方法还包括：

在推理负载小于预设阈值的情况下，激活所述 Draft 异步生成模型生成 Draft；

将所述 Draft 异步生成模型生成的 Draft 存储在所述 Draft 池中。

4. 根据权利要求 1 至 3 中任一项所述的方法，其特征在于，所述 Draft 池中还存储有 Draft 的属性信息。

5. 根据权利要求 4 所述的方法，其特征在于，所述从 Draft 池中获取与所述文本提示信息相关的目标 Draft，包括：

根据所述 Draft 池中存储的 Draft 的属性信息，从所述 Draft 池中获取与所述文本提示信息相关的目标 Draft。

6. 根据权利要求 4 或 5 所述的方法，其特征在于，所述方法还包括：

根据所述 Draft 池中存储的 Draft 的属性信息，对所述 Draft 池中存储的 Draft 进行检索和/或更新。

7. 根据权利要求 4 至 6 中任一项所述的方法，其特征在于，所述方法还包括：

根据所述 Draft 池中存储的 Draft 的属性信息，对所述 Draft 生成模型池中包括的 Draft 生成模型进行操作，所述操作包括：激活，休眠，释放。

8. 根据权利要求 4 至 7 中任一项所述的方法，其特征在于，所述 Draft 的属性信息包括以下中的至少一种：所述 Draft 的领域、内容、对应的 Draft 生成模型的标签、被调用的次数、命中的次数、命中的文本提示信息的列表。

9. 根据权利要求 1 至 8 中任一项所述的方法，其特征在于，Draft 生成请求信息还包括生成 Draft 的数量，所述生成 Draft 的数量是根据模型推理资源的使用情况确定的，所述模型推理资源用于所述 Draft 生成模型池中的 Draft 生成模型产生 Draft 和/或所述 Draft 验证模型池中的 Draft 验证模型执行验证过程。

10. 根据权利要求 1 至 9 中任一项所述的方法，其特征在于，所述方法还包括：

根据模型推理资源的使用情况，自动申请/释放所述模型推理资源。

11. 根据权利要求 1 至 10 中任一项所述的方法，其特征在于，所述 Draft 池中包括共享 Draft 池和第一用户的私有 Draft 池，

所述从 Draft 池中获取与所述文本提示信息相关的目标 Draft，包括：

从所述共享 Draft 池和/或所述第一用户的私有 Draft 池中获取与所述文本提示信息相关的目标 Draft。

12. 根据权利要求 2 至 11 中任一项所述的方法，其特征在于，所述将所述 Draft 同步生成模型生成的 Draft 存储在所述 Draft 池中，包括：

将所述 Draft 同步生成模型生成的 Draft 存储在第一用户的私有 Draft 池中。

13. 根据权利要求 3 至 12 中任一项所述的方法，其特征在于，所述将所述 Draft 异步生成模型生成的 Draft 存储在所述 Draft 池中，包括：

将所述 Draft 异步生成模型生成的 Draft 存储在第一用户的私有 Draft 池中。

14. 根据权利要求 1 至 13 中任一项所述的方法，其特征在于，所述至少一个 Draft 验证模型和所述至少一个 Draft 生成模型之间的对应模式包括：一对一，一对多，多对一，多对多。

15. 一种模型推理的系统，其特征在于，所述系统包括：

Draft 验证管理单元，用于从 Draft 验证模型池中确定至少一个 Draft 验证模型，所述 Draft 验证模型池中包括多个 Draft 验证模型；

Draft 生成管理单元，用于从 Draft 生成模型池中确定与所述至少一个 Draft 验证模型对应的至少一个 Draft 生成模型，其中，所述 Draft 生成模型池中包括多个 Draft 生成模型；

所述至少一个 Draft 验证模型，用于接收第一用户输入的文本提示信息，所述文本提示信息用于生成文本信息；

所述 Draft 验证管理单元，还用于从 Draft 池中获取与所述文本提示信息相关的目标 Draft，所述 Draft 池中存储有多个 Draft，所述多个 Draft 包括所述至少一个 Draft 生成模型生成的 Draft；

所述至少一个 Draft 验证模型，还用于对所述目标 Draft 进行验证，并得到验证结果；

所述至少一个 Draft 验证模型，还用于根据所述验证结果生成所述文本信息并反馈给所述第一用户。

16. 根据权利要求 15 所述的系统，其特征在于，所述至少一个 Draft 生成模型包括 Draft 同步生成模型，

所述至少一个 Draft 验证模型，还用于根据所述文本提示信息获得 Draft 生成请求信息，并将 Draft 生成请求信息通过所述 Draft 验证管理单发送给所述 Draft 生成管理单元；

所述 Draft 生成管理单元，还用于将所述 Draft 生成请求信息发送给所述 Draft 同步生成模型；

所述 Draft 同步生成模型，用于根据所述 Draft 生成请求信息生成 Draft 并发送给所述 Draft 生成管理单元；

所述 Draft 生成管理单元，还用于将所述 Draft 同步生成模型生成的 Draft 存储在所述 Draft 池中。

17. 根据权利要求 15 或 16 所述的系统，其特征在于，所述至少一个 Draft 生成模型包括 Draft 异步生成模型，

所述 Draft 验证管理单元，还用于在推理负载小于预设阈值的情况下，激活所述 Draft 异步生成模型生成 Draft；

所述 Draft 异步生成模型，用于生成 Draft 并发送给所述 Draft 验证管理单元；

所述 Draft 验证管理单元，还用于将所述 Draft 异步生成模型生成的 Draft 存储在所述 Draft 池中。

18. 根据权利要求 15 至 17 中任一项所述的系统，其特征在于，所述 Draft 池中还存储有 Draft 的属性信息。

19. 根据权利要求 18 所述的系统，其特征在于，所述 Draft 验证管理单元具体用于：

根据所述 Draft 池中存储的 Draft 的属性信息，从所述 Draft 池中获取与所述文本提示信息相关的目标 Draft。

20. 根据权利要求 18 或 19 所述的系统，其特征在于，所述系统还包括：

Draft 分析单元，用于根据所述 Draft 池中存储的 Draft 的属性信息，对所述 Draft 池中存储的 Draft 进行检索和/或更新。

21. 根据权利要求 18 至 20 中任一项所述的系统，其特征在于，

所述 Draft 生成管理单元，还用于根据所述 Draft 池中存储的 Draft 的属性信息，对所述 Draft 生成模型池中包括的 Draft 生成模型进行操作，所述操作包括：激活，休眠，释放。

22. 根据权利要求 18 至 21 中任一项所述的系统，其特征在于，所述 Draft 的属性信息包括以下中的至少一种：所述 Draft 的领域、内容、对应的 Draft 生成模型的标签、被调用的次数、命中的次数、命中的文本提示信息的列表。

23. 根据权利要求 15 至 21 中任一项所述的系统，其特征在于，Draft 生成请求信息还包括生成 Draft 的数量，所述生成 Draft 的数量是根据模型推理资源的使用情况确定的，所述模型推理资源用于所述 Draft 生成模型池中的 Draft 生成模型产生 Draft 和/或所述 Draft 验证模型池中的 Draft 验证模型执行验证过

程。

24. 根据权利要求 15 至 23 中任一项所述的系统，其特征在于，所述系统还包括：

推理资源分析单元，用于根据模型推理资源的使用情况，自动申请/释放所述模型推理资源。

25. 根据权利要求 15 至 24 中任一项所述的系统，其特征在于，所述 Draft 池中包括共享 Draft 池和第一用户的私有 Draft 池，所述 Draft 验证管理单元具体用于：

从所述共享 Draft 池和/或所述第一用户的私有 Draft 池中获取与所述文本提示信息相关的目标 Draft。

26. 根据权利要求 16 至 25 中任一项所述的系统，其特征在于，所述 Draft 生成管理单元具体用于：

将所述 Draft 同步生成模型生成的 Draft 存储在第一用户的私有 Draft 池中。

27. 根据权利要求 17 至 26 中任一项所述的系统，其特征在于，所述 Draft 生成管理单元具体用于：

将所述 Draft 异步生成模型生成的 Draft 存储在第一用户的私有 Draft 池中。

28. 根据权利要求 15 至 27 中任一项所述的系统，其特征在于，所述至少一个 Draft 验证模型和所述至少一个 Draft 生成模型之间的对应模式包括：一对一，一对多，多对一，多对多。

29. 一种计算设备集群，其特征在于，包括至少一个计算设备，每个计算设备包括处理器和存储器；

所述至少一个计算设备的处理器用于执行所述至少一个计算设备的存储器中存储的指令，以使得所述计算设备集群执行如权利要求 1 至 14 中任一项所述的方法。

30. 一种包含指令的计算机程序产品，其特征在于，当所述指令被计算设备集群运行时，使得所述计算设备集群执行如权利要求的 1 至 14 中任一项所述的方法。

31. 一种计算机可读存储介质，其特征在于，包括计算机程序指令，当所述计算机程序指令由计算设备集群执行时，所述计算设备集群执行如权利要求 1 至 14 中任一项所述的方法。