

一种基于大语言模型的数据处理方法、装置及设备

技术领域

[01] 本申请涉及人工智能技术领域，尤其是涉及一种基于大语言模型（Large Language Model，简称 LLM）的数据处理方法、装置及设备。

5 背景技术

[02] 随着网络技术的快速发展，网络数据越来越多。为了从大量网络数据中检索到用户所需的数据，需要由用户提供关键词（即原始输入问题）。服务器基于关键词从大量网络数据中检索到用户所需的数据。

10 [03] 然而，由于网络数据的数据量庞大，服务器需要消耗大量计算资源才能够检索到用户所需的数据，且花费大量时间，导致服务器资源消耗大，也使得用户无法及时获取所需数据。

[04] 比如说，实际场景中部署的摄像机在采集大量图片后，将大量图片传输至服务器。服务器为了从这些图片中分析用户所需数据，需要消耗大量计算资源，花费大量时间。

发明内容

15 [05] 有鉴于此，本申请提供一种基于大语言模型的数据处理方法、装置及设备，能够节省服务器的计算资源，并节省时间开销，能够快速确定出目标答案。

[06] 本申请提供一种基于大语言模型的数据处理方法，目标大语言模型包括基础模型和至少两个旁枝模型，所述方法包括：

[07] 获取原始输入问题，并确定所述原始输入问题对应的目标特征；

[08] 从所有旁枝模型中，选取与所述目标特征对应的目标旁枝模型；

20 [09] 向所述基础模型输入所述原始输入问题，得到所述原始输入问题对应的中间处理特征和所述原始输入问题对应的输出处理特征；

25 [10] 向所述目标旁枝模型发送触发指令，以使得所述目标旁枝模型根据所述触发指令从所述基础模型获取所述中间处理特征和所述输出处理特征，对所述中间处理特征进行处理得到所述原始输入问题对应的旁支处理特征，对所述旁支处理特征和所述输出处理特征进行处理得到所述原始输入问题对应的目标答案。

[11] 本申请提供一种基于大语言模型的数据处理装置，目标大语言模型包括基础模型和至少两个旁枝模型，所述装置包括：

[12] 获取模块，用于获取原始输入问题，并确定所述原始输入问题对应的目标特征；从所有旁枝模型中，选取与所述目标特征对应的目标旁枝模型；

30 [13] 处理模块，用于向所述基础模型输入所述原始输入问题，得到所述原始输入问题对应的中间处理特征和所述原始输入问题对应的输出处理特征；

35 [14] 向所述目标旁枝模型发送触发指令，以使得所述目标旁枝模型根据所述触发指令从所述基础模型获取所述中间处理特征和所述输出处理特征，对所述中间处理特征进行处理得到所述原始输入问题对应的旁支处理特征，对所述旁支处理特征和所述输出处理特征进行处理得到所述原始输入问题对应的目标答案。

[15] 本申请提供一种电子设备，包括：处理器和机器可读存储介质，所述机器可读存储介质存储有能够被所述处理器执行的机器可执行指令；所述处理器用于执行机器可执行指令，以实现上述的基于大语言模型的数据处理方法。

[16] 由以上技术方案可见，本申请实施例中，目标大语言模型包括基础模型和至少两个旁枝模型，

如至少两个旁枝模型可以是针对不同行业的旁枝模型。在得到原始输入问题之后，可以从所有旁枝模型中选取目标旁枝模型，基于基础模型和目标旁枝模型确定原始输入问题对应的目标答案。这样，针对不同行业的原始输入问题，可以基于该行业的旁枝模型确定目标答案，从而得到准确的目标答案。比如说，针对行业 A 的原始输入问题，基于行业 A 的旁枝模型确定目标答案，针对行业 B 的原始输入问题，基于行业 B 的旁枝模型确定目标答案，以此类推。这样，针对不同行业的原始输入问题均可以得到准确的目标答案。

[17] 能够借助目标大语言模型得到目标答案，只需要消耗少量计算资源就能够查询到目标答案。获取目标答案的时间更短，从而节省服务器的计算资源，并节省时间开销。能够快速确定出原始输入问题对应的目标答案，提供逻辑分析推理能力，利用目标大语言模型的分析能力给出准确可靠的目标答案。

附图说明

[18] 为了更加清楚地说明本申请实施例或者现有技术中的技术方案，下面将对本申请实施例或者现有技术描述中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图仅仅是本申请中记载的一些实施例，对于本领域普通技术人员来讲，还可以根据本申请实施例的这些附图获得其他的附图。

[19] 图 1 是一种实施方式中的基于大语言模型的数据处理方法的流程示意图；

[20] 图 2A 是一种实施方式中的大语言模型的结构示意图；

[21] 图 2B 是一种实施方式中的行业大语言模型的结构示意图；

[22] 图 3 是一种实施方式中的目标大语言模型的结构示意图；

[23] 图 4 是一种实施方式中的针对目标大语言模型的训练过程的示意图；

[24] 图 5A、图 5B 和图 5C 是一种实施方式中的行业大语言模型的结构示意图；

[25] 图 6 是一种实施方式中的基于目标大语言模型的数据处理过程的示意图；

[26] 图 7 是一种实施方式中的目标大语言模型的结构示意图；

[27] 图 8 是一种实施方式中的基于大语言模型的数据处理装置的结构示意图；

[28] 图 9 是一种实施方式中的电子设备的硬件结构图。

具体实施方式

[29] 在本申请实施例使用的术语仅仅是出于描述特定实施例的目的，而非限制本申请。本申请和权利要求书中所使用的单数形式的“一种”、“所述”和“该”也旨在包括多数形式，除非上下文清楚地表示其它含义。还应当理解，本文中使用的术语“和/或”是指包含一个或多个相关联的列出项目的任何或所有可能组合。

[30] 应当理解，尽管在本申请实施例可能采用术语第一、第二、第三等来描述各种信息，但这些信息不应限于这些术语。这些术语仅用来将同一类型的信息彼此区分开。例如，在不脱离本申请范围的情况下，第一信息也可以被称为第二信息，类似地，第二信息也可以被称为第一信息。取决于语境，此外，所使用的词语“如果”可以被解释成为“在……时”或“当……时”或“响应于确定”。

[31] 本申请实施例中提出一种基于大语言模型的数据处理方法，该方法可以应用于任意电子设备，如服务器等。一个例子中，目标大语言模型可以包括基础模型和至少两个旁枝模型。参见图 1 所示，该方法可以包括：

[32] 步骤 101、获取原始输入问题，并确定原始输入问题对应的目标特征。

[33] 步骤 102、从所有旁枝模型中，选取与所述目标特征对应的目标旁枝模型。

[34] 步骤 103、向基础模型输入原始输入问题，得到该原始输入问题对应的中间处理特征和该原始输入问题对应的输出处理特征。

[35] 步骤 104、向目标旁枝模型发送触发指令，以使得目标旁枝模型根据该触发指令从基础模型获

取该中间处理特征和该输出处理特征，对该中间处理特征进行处理得到该原始输入问题对应的旁支处理特征，对该旁支处理特征和该输出处理特征进行处理得到该原始输入问题对应的目标答案。

[36] 一个例子中，针对每个旁枝模型，该旁枝模型可以是针对一个行业的模型，且不同旁枝模型是针对不同行业的模型。

5 [37] 从所有旁枝模型中，选取与目标特征对应的目标旁枝模型，可以包括但不限于：若目标特征包括原始输入问题所属的目标行业，则从所有旁枝模型中，选取与目标行业对应的旁枝模型作为目标旁枝模型。

[38] 一个例子中，针对每个旁枝模型，该旁枝模型可以是针对一个行业的模型，且不同旁枝模型是针对不同行业的模型。

10 [39] 从所有旁枝模型中，选取与目标特征对应的目标旁枝模型，可以包括但不限于：获取与原始输入问题对应的会话标识；其中，在用户设备与目标大语言模型建立连接时，可以为用户设备分配会话标识，在用户设备每次向目标大语言模型提供原始输入问题时，同步提供该会话标识；

[40] 若存在该会话标识对应的旁枝模型，则可以将该会话标识对应的旁枝模型确定为原始输入问题对应的目标旁枝模型；

15 [41] 若不存在该会话标识对应的旁枝模型，且目标特征包括原始输入问题所属的目标行业，则从所有旁枝模型中，选取与目标行业对应的旁枝模型作为目标旁枝模型，并记录该会话标识与目标旁枝模型的对应关系。

[42] 一个例子中，确定与原始输入问题对应的目标行业的过程，可以包括但不限于：获取原始输入问题对应的目标关键词，通过查询行业词典得到目标关键词对应的目标行业；其中，行业词典包括关键词与行业的映射关系。

20 [43] 或者，将该原始输入问题输入至行业识别模型得到该原始输入问题对应的目标行业；其中，该行业识别模型可以用于识别行业。

[44] 一个例子中，针对每个旁枝模型，该旁枝模型可以是针对一个语种的模型，且不同旁枝模型是针对不同语种的模型。

25 [45] 确定原始输入问题对应的目标特征，可以包括但不限于：若目标特征包括原始输入问题对应的目标语种，则对原始输入问题进行分词处理得到多个 token；基于多个 token 确定与原始输入问题对应的目标语种；

[46] 从所有旁枝模型中，选取与所标特征对应的目标旁枝模型，可以包括但不限于：从所有旁枝模型中，选取与目标语种对应的旁枝模型作为目标旁枝模型。

30 [47] 一个例子中，针对目标大语言模型的训练过程，可以包括但不限于：

[48] 基于基础模型对应的样本数据训练得到基础模型。

35 [49] 针对每个待训练旁枝模型，对基础模型与该待训练旁枝模型进行旁路连接得到行业模型。基于该待训练旁枝模型对应的样本数据，对该待训练旁枝模型进行训练得到已训练的旁枝模型。在训练过程中，对该待训练旁枝模型的网络参数进行调整得到已训练的旁枝模型，不对基础模型的网络参数进行调整。

[50] 在得到每个待训练旁枝模型对应的已训练的旁枝模型后，将基础模型与所有已训练的旁枝模型进行旁路连接得到目标大语言模型。

40 [51] 一个例子中，上述执行顺序只是为了方便描述给出的示例，在实际应用中，还可以改变步骤之间的执行顺序，对此执行顺序不做限制。而且，在其它实施例中，并不一定按照本说明书示出和描述的顺序来执行相应方法的步骤，其方法所包括的步骤可以比本说明书所描述的更多或更少。此外，本说明书中所描述的单个步骤，在其它实施例中可能被分解为多个步骤进行描述。本说明书中所描述的多个步骤，在其它实施例中也可能被合并为单个步骤进行描述。

45 [52] 由以上技术方案可见，本申请实施例中，目标大语言模型包括基础模型和至少两个旁枝模型，如至少两个旁枝模型可以是针对不同行业的旁枝模型。在得到原始输入问题之后，可以从所有旁枝模型中选取目标旁枝模型，基于基础模型和目标旁枝模型确定原始输入问题对应的目标答案。这样，针对不同行业的原始输入问题，可以基于该行业的旁枝模型确定目标答案，从而得到准确的目标答

案。比如说，针对行业 A 的原始输入问题，基于行业 A 的旁枝模型确定目标答案，针对行业 B 的原始输入问题，基于行业 B 的旁枝模型确定目标答案，以此类推。这样，针对不同行业的原始输入问题均可以得到准确的目标答案。

5 [53] 能够借助目标大语言模型得到目标答案，只需要消耗少量计算资源就能够查询到目标答案。获取目标答案的时间更短，从而节省服务器的计算资源，并节省时间开销。能够快速确定出原始输入问题对应的目标答案，提供逻辑分析推理能力，利用目标大语言模型的分析能力给出准确可靠的目标答案。

[54] 以下结合具体应用场景，对本申请实施例的上述技术方案进行说明。

10 [55] 由于网络数据（如文本数据、视频数据、语音数据、图片数据等）的数据量很大，服务器需要消耗大量计算资源才能够检索到用户所需数据，且需要花费大量时间，从而导致服务器的大量资源消耗，且用户很长时间才能够获知所需数据。比如说，实际场景部署大量摄像机，这些摄像机采集大量图片，为了从这些图片中分析用户所需数据，需要消耗大量计算资源，花费大量时间。

15 [56] 针对上述发现，可以借助大语言模型（Large Language Model, LLM）对数据进行分析，比如说，将问题数据输入至大语言模型，由大语言模型输出目标答案。这样，只需要消耗少量计算资源就能够从大量数据中查询到问题数据对应的目标答案，获取目标答案的时间更短。从而节省服务器的计算资源，节省时间开销，利用大语言模型的分析能力给出准确可靠的目标答案。

20 [57] 比如说，ChatGPT 是一种基于自然语言处理和机器学习的大语言模型。可以模拟人类语言交流，实现与用户的智能对话，已经广泛应用于智能客服、聊天机器人、语音助手等领域，为人们提供了更方便、快捷和智能化的服务。基于此，可以借助 ChatGPT 对数据进行分析，得到问题数据对应的目标答案。

25 [58] 随着人工智能的快速发展，基于大语言模型，各行业会使用行业数据对大语言模型进行微调训练，得到新大语言模型，新大语言模型也可以称为行业大语言模型或者专家大语言模型（后续以行业大语言模型为例）。比如说，针对通信行业，可以采用通信行业的数据对大语言模型进行微调训练，得到通信行业的行业大语言模型。针对医疗行业，可以采用医疗行业的数据对大语言模型进行微调训练，得到医疗行业的行业大语言模型。针对教育行业，可以采用教育行业的数据对大语言模型进行微调训练，得到教育行业的行业大语言模型。

30 [59] 然而，由于大语言模型的网络参数很多，即使为大语言模型添加少量网络参数（如针对通信行业、医疗行业、教育行业等的网络参数），那么，在对大语言模型进行微调训练时，也需要对大量网络参数进行调整，需要消耗大量时间和大量计算资源才能够完成大语言模型的微调训练，资源消耗很大。

[60] 比如说，参见图 2A 所示，为大语言模型的结构示意图，大语言模型可以包括网络 A、网络 B 和网络 C，网络 A、网络 B 和网络 C 均可以包括大量网络参数，如万亿级别的网络参数等。在大语言模型训练完成之后，各行业还可以使用行业数据对大语言模型进行微调训练，从而得到行业大语言模型。

35 [61] 比如说，参见图 2B 所示，为行业大语言模型的结构示意图，在大语言模型的基础上，行业大语言模型额外增加了网络 A1、网络 B1 和网络 C1，网络 A1、网络 B1 和网络 C1，均包括少量网络参数，如数十个、数百个网络参数等。

40 [62] 在使用行业数据对大语言模型进行微调训练时，在梯度的回传过程中，需要对网络 A、网络 B、网络 C、网络 A1、网络 B1 和网络 C1 的网络参数进行调整，虽然网络 A1、网络 B1 和网络 C1 的网络参数比较少，但是，网络 A、网络 B 和网络 C 的网络参数很多，如万亿级别的网络参数，因此，微调训练过程需要消耗大量时间和大量计算资源才能够完成大语言模型的微调训练。

[63] 针对上述发现，本申请提供一种基于大语言模型的数据处理方法，可以采用旁枝模型的方式对大语言模型进行调整，多个行业可以采用同一大语言模型。

45 [64] 比如说，参见图 3 所示，为目标大语言模型的结构示意图，在大语言模型的基础上，目标大语言模型额外增加了旁枝模型 1、旁枝模型 2 和旁枝模型 3，当然，这里是以 3 个旁枝模型为例，旁枝模型的数量可以更多或者更少。

[65] 在图 3 中，大语言模型可以包括网络 A、网络 B 和网络 C，可以将该大语言模型称为基础模型，

即基础模型包括网络 A、网络 B 和网络 C。

5 [66] 旁枝模型 1 可以包括网络 A1、网络 B1 和网络 C1，如旁枝模型 1 可以是针对通信行业的旁枝模型，基础模型和旁枝模型 1 可以构成通信行业的行业大语言模型。在使用通信行业的行业数据对大语言模型进行微调训练时，在梯度的回传过程中，只需要对网络 A1、网络 B1 和网络 C1 的网络参数进行调整，而不需要对网络 A、网络 B 和网络 C 的网络参数进行调整。显然，由于网络 A1、网络 B1 和网络 C1 的网络参数比较少，因此，微调训练过程只需要消耗少量时间和少量计算资源，从而能够快速完成大语言模型

10 [67] 旁枝模型 2 可以包括网络 A2、网络 B2 和网络 C2，如旁枝模型 2 可以是针对医疗行业的旁枝模型，基础模型和旁枝模型 2 可以构成医疗行业的行业大语言模型。在使用医疗行业的行业数据对大语言模型进行微调训练时，在梯度的回传过程中，只需要对网络 A2、网络 B2 和网络 C2 的网络参数进行调整，而不需要对网络 A、网络 B 和网络 C 的网络参数进行调整。微调训练过程只需要消耗少量时间和少量计算资源，从而快速完成微调训练，节省计算资源。

15 [68] 旁枝模型 3 可以包括网络 A3、网络 B3 和网络 C3，如旁枝模型 3 可以是针对教育行业的旁枝模型，基础模型和旁枝模型 3 可以构成教育行业的行业大语言模型。在使用教育行业的行业数据对大语言模型进行微调训练时，在梯度的回传过程中，只需要对网络 A3、网络 B3 和网络 C3 的网络参数进行调整，而不需要对网络 A、网络 B 和网络 C 的网络参数进行调整。微调训练过程只需要消耗少量时间和少量计算资源，从而快速完成微调训练，节省计算资源。

20 [69] 一个例子中，可以涉及目标大语言模型的训练过程和基于目标大语言模型的数据处理过程。在目标大语言模型的训练过程中，可以训练得到目标大语言模型，目标大语言模型可以包括基础模型和至少两个旁枝模型。在基于目标大语言模型的数据处理过程中，可以将原始输入问题输入至目标大语言模型，由目标大语言模型输出原始输入问题对应的目标答案。以下对目标大语言模型的训练过程和基于目标大语言模型的数据处理过程进行说明。

[70] 第一，针对目标大语言模型的训练过程。比如说，参见图 4 所示，为针对目标大语言模型的训练过程的示意图，该过程可以包括：

25 [71] 步骤 401、基于基础模型对应的样本数据训练得到基础模型。

[72] 比如说，基础模型（即大语言模型）的结构可以参见图 2A 所示，基础模型可以包括网络 A、网络 B 和网络 C，对此基础模型的结构不做限制。网络 A 可以是 Transformer 结构，也可以是其它结构，网络 B 可以是 Transformer 结构，也可以是其它结构，网络 C 可以是 Transformer 结构，也可以是其它结构。

30 [73] 比如说，可以获取基础模型对应的样本数据（如大量样本数据），可以采用样本数据训练得到基础模型，对此基础模型的训练过程不做限制。

[74] 步骤 402、针对每个待训练旁枝模型，将基础模型与该待训练旁枝模型进行旁路连接得到行业模型（即行业大语言模型），基于该待训练旁枝模型对应的样本数据对该待训练旁枝模型进行训练，得到该待训练旁枝模型对应的已训练的旁枝模型。

35 [75] 一个例子中，在待训练旁枝模型的训练过程中，对该待训练旁枝模型的网络参数进行调整得到已训练的旁枝模型，不对基础模型的网络参数进行调整。

[76] 比如说，针对通信行业的待训练旁枝模型，待训练旁枝模型包括网络 A1、网络 B1 和网络 C1，将基础模型与该待训练旁枝模型进行旁路连接得到通信行业的行业大语言模型，参见图 5A 所示，为行业大语言模型的结构示意图。

40 [77] 在此基础上，可以基于该待训练旁枝模型对应的样本数据（即通信行业的行业数据）对该行业大语言模型进行训练，得到该待训练旁枝模型对应的已训练的旁枝模型（记为旁枝模型 1）。在训练过程中，只对该待训练旁枝模型的网络参数（即网络 A1、网络 B1 和网络 C1 的网络参数）进行调整得到旁枝模型 1，而不对基础模型的网络参数进行调整，对此训练过程不做限制。

45 [78] 比如说，针对医疗行业的待训练旁枝模型，待训练旁枝模型包括网络 A2、网络 B2 和网络 C2，将基础模型与该待训练旁枝模型进行旁路连接得到医疗行业的行业大语言模型，参见图 5B 所示，为行业大语言模型的结构示意图。

[79] 在此基础上，可以基于医疗行业的行业数据对该行业大语言模型进行训练，得到已训练的旁枝

模型 2。在训练过程中，只对网络 A2、网络 B2 和网络 C2 的网络参数）进行调整得到旁枝模型 2，而不对基础模型的网络参数进行调整。

[80] 比如说，针对教育行业的待训练旁枝模型，待训练旁枝模型包括网络 A3、网络 B3 和网络 C3，将基础模型与该待训练旁枝模型进行旁路连接得到教育行业的行业大语言模型，参见图 5C 所示，为行业大语言模型的结构示意图。

[81] 在此基础上，可以基于教育行业的行业数据对该行业大语言模型进行训练，得到已训练的旁枝模型 3。在训练过程中，只对网络 A3、网络 B3 和网络 C3 的网络参数）进行调整得到旁枝模型 3，而不对基础模型的网络参数进行调整。

[82] 步骤 403、在得到每个待训练旁枝模型对应的已训练的旁枝模型后，将基础模型与所有已训练的旁枝模型进行旁路连接得到目标大语言模型。

[83] 比如说，可以将基础模型与旁枝模型 1、旁枝模型 2 和旁枝模型 3 进行旁路连接，得到目标大语言模型，即目标大语言模型包括基础模型、旁枝模型 1、旁枝模型 2 和旁枝模型 3。参见图 3 所示，为目标大语言模型的结构示意图。

[84] 一个例子中，目标大语言模型是在基础模型上搭建旁枝模型，将基础模型的部分输出作为旁枝模型的输入，所有待训练的网络参数均位于旁枝模型。由于基础模型仅提供输入，因此，反向传播的复杂度取决于旁枝模型的规模，并不需要直接在基础模型上执行反向传播，从而可以明显提升训练效率。

[85] 比如说，由于每个旁枝模型的网络参数个数相对较小，相对于基础模型来说不会增加很多网络参数。因此，在训练旁枝模型（即调整旁枝模型的网络参数）时，训练效率较高，对内存的占用比较小，占用的计算资源比较少。能够快速完成训练，可以很好地实现快速、低成本地扩展模型能力，增强模型性能。

[86] 一个例子中，在得到基础模型之后，针对特定行业，对该行业的旁枝模型进行训练后，将旁枝模型与基础模型进行旁路连接得到该行业的行业大语言模型。行业大语言模型可以表示为基础模型加一个旁挂的旁枝模型，且行业大语言模型对应与该行业有关的任务。其中，行业大语言模型可以分为旁枝模型与基础模型等两部分，可以将行业大语言模型记为：Model = Base + Ladder。

[87] 一个例子中，可以将基础模型与多个行业的旁枝模型进行旁路连接，得到目标大语言模型，即目标大语言模型包括多个行业大语言模型。目标大语言模型包括基础模型和多个旁枝模型，每个旁枝模型与一个行业大语言模型对应。

[88] 比如说，行业大语言模型 1 记为 Base+Ladder¹，行业大语言模型 2 记为 Base+Ladder²，以此类推。显然，各行业大语言模型的 Base 是一致的。在将这些行业大语言模型进行旁路连接起来后，可以得到一个支持多个行业的目标大语言模型。

[89] 一个例子中，针对旁枝模型中的每个网络（如网络 A1、网络 B1、网络 C1、网络 A2、网络 B2 和网络 C2、网络 A3、网络 B3 和网络 C3 等），该网络可以是 Transformer 结构（如轻量化（lightweight）的 Transformer 结构，轻量化是指 Transformer 结构的网络参数比较少），也可以是其它结构，对此不做限制。

[90] 第二，基于目标大语言模型的数据处理过程。比如说，参见图 6 所示，为基于目标大语言模型的数据处理过程的示意图，该过程可以包括：

[91] 步骤 601、获取原始输入问题。比如说，可以接收用户输入的原始输入问题，也可以采用其它方式获取原始输入问题，对此不做限制。

[92] 一个例子中，原始输入问题可以是图片数据格式的输入问题，也可以是文本数据格式的输入问题，也可以是语音数据格式的输入问题，也可以是视频数据格式的输入问题，对此原始输入问题的数据格式不做限制。

[93] 步骤 602、从所有旁枝模型中，选取与原始输入问题对应的目标旁枝模型。比如说，可以确定原始输入问题对应的目标特征，从所有旁枝模型中，选取与目标特征对应的目标旁枝模型作为原始输入问题对应的目标旁枝模型。

[94] 一个例子中，目标大语言模型可以包括 Gating（门控）模块（也称为 Gating 网络），即目标大

语言模型的入口存在 Gating 模块。Gating 模块用于获取原始输入问题，并从所有旁枝模型中选取与原始输入问题对应的目标旁枝模型。比如说，参见图 7 所示，目标大语言模型的入口存在 Gating 模块。

[95] 一个例子中，为了从所有旁枝模型选取目标旁枝模型，可以采用如下方式：

- 5 [96] 方式 1：目标大语言模型包括至少两个旁枝模型，针对每个旁枝模型，该旁枝模型可以是针对一个行业的模型，且不同旁枝模型是针对不同行业的模型。比如说，参见图 3 所示，旁枝模型 1 是针对通信行业的旁枝模型，旁枝模型 2 是针对医疗行业的旁枝模型，旁枝模型 3 是针对教育行业的旁枝模型。

- 10 [97] 在此基础上，可以确定与原始输入问题所属的目标行业（即原始输入问题对应的目标特征包括原始输入问题所属的目标行业），从所有旁枝模型中，选取与目标行业对应的旁枝模型作为目标旁枝模型。比如说，若原始输入问题对应的目标行业是通信行业，则将旁枝模型 1 作为目标旁枝模型。若原始输入问题对应的目标行业是医疗行业，则将旁枝模型 2 作为目标旁枝模型。若原始输入问题对应的目标行业是教育行业，则将旁枝模型 3 作为目标旁枝模型。

- 15 [98] 一个例子中，为了获取目标行业对应的旁枝模型，还可以预先配置行业与旁枝模型之间的映射关系。基于此，在得到与原始输入问题对应的目标行业之后，可以通过查询该映射关系，得到目标行业对应的旁枝模型。

[99] 一个例子中，为了确定与原始输入问题对应的目标行业，可以预先配置行业词典，该行业词典包括关键词与行业的映射关系，对此行业词典不做限制。

[100] 可以获取原始输入问题对应的目标关键词。

- 20 [101] 比如说，若原始输入问题是文本格式的原始输入问题，则对原始输入问题进行分词处理，得到多个词语。可以将某个词语作为目标关键词，也可以将部分词语作为目标关键词，对此目标关键词的选取方式不做限制。

- 25 [102] 又例如，若原始输入问题是图片格式的原始输入问题，则可以从图片中提取出目标关键词，比如说，将图片输入给关键词提取模型得到目标关键词，也可以采用图像算法对图像进行分析得到目标关键词，对此不做限制。

[103] 又例如，若原始输入问题是音频格式的原始输入问题，则可以从音频中提取出目标关键词，对此提取方式不做限制。若原始输入问题是视频格式的原始输入问题，则可以从视频中提取出目标关键词，对此提取方式不做限制。

- 30 [104] 在得到目标关键词之后，可以通过目标关键词查询行业词典，得到与目标关键词对应的目标行业，继而将目标行业对应的旁枝模型作为目标旁枝模型。

[105] 一个例子中，为了确定与原始输入问题对应的目标行业，可以预先训练一个行业识别模型，该行业识别模型是用于识别行业的网络模型，对此行业识别模型的网络结构和训练过程不做限制，能够识别出行业即可。

- 35 [106] 在得到原始输入问题之后，可以将原始输入问题输入至行业识别模型，由行业识别模型基于原始输入问题进行处理，得到原始输入问题对应的目标行业，并输出原始输入问题对应的目标行业。这样，可以得到原始输入问题对应的目标行业，继而将目标行业对应的旁枝模型作为目标旁枝模型。

[107] 方式 2：目标大语言模型包括至少两个旁枝模型，针对每个旁枝模型，该旁枝模型可以是针对一个行业的模型，且不同旁枝模型是针对不同行业的模型。

- 40 [108] 在此基础上，可以获取与原始输入问题对应的会话标识（Session）。比如说，在用户设备（即用户）与目标大语言模型建立连接时，目标大语言模型可以为用户设备分配会话标识。在用户设备每次向目标大语言模型提供原始输入问题时，同步提供该会话标识，即原始输入问题均会携带该会话标识，且每次原始输入问题的会话标识相同。这样，在获取到原始输入问题时，还可以得到原始输入问题对应的会话标识，继而基于该会话标识进行分析。

- 45 [109] 若不存在该会话标识对应的旁枝模型（如针对用户设备向目标大语言模型输入的首个原始输入问题，则不存在该会话标识对应的旁枝模型），则确定与原始输入问题所属的目标行业（即原

始输入问题对应的目标特征包括原始输入问题所属的目标行业)，从所有旁枝模型中，选取与目标行业对应的旁枝模型作为目标旁枝模型，并记录会话标识与目标旁枝模型的对应关系。

5 [110] 比如说，若不存在该会话标识对应的旁枝模型，则可以采用方式 1 确定目标旁枝模型，对此过程不再重复赘述。在得到目标旁枝模型之后，可以记录该会话标识与目标旁枝模型的对应关系，以便后续过程查询该对应关系。

[111] 若存在该会话标识对应的旁枝模型（如针对用户设备向目标大语言模型输入的非首个原始输入问题，则存在该会话标识对应的旁枝模型，即，首个原始输入问题时，记录的该会话标识与旁枝模型的对应关系），则可以直接将该会话标识对应的旁枝模型确定为原始输入问题对应的目标旁枝模型。

10 [112] 方式 3：目标大语言模型包括至少两个旁枝模型，针对每个旁枝模型，该旁枝模型可以是针对一个语种的模型，且不同旁枝模型是针对不同语种的模型。

15 [113] 比如说，在训练得到目标大语言模型时，旁枝模型 1 是针对语种 A 的旁枝模型（即采用语种 A 的样本数据训练旁枝模型 1），旁枝模型 2 是针对语种 B 的旁枝模型（即采用语种 B 的样本数据训练旁枝模型 2），旁枝模型 3 是针对语种 C 的旁枝模型（即采用语种 C 的样本数据训练旁枝模型 3）。

[114] 在此基础上，可以对原始输入问题进行分词处理得到多个 token。比如说，以原始输入问题是文本数据格式的输入问题为例，可以将原始输入问题切割成小块，这些小块称为 token，这个过程称为分词，即可以得到多个 token。

20 [115] 基于多个 token 确定与原始输入问题对应的目标语种（即原始输入问题对应的目标特征包括原始输入问题对应的目标语种）。比如说，对每个 token 进行分析得到 token 的语种，将 token 的语种作为原始输入问题对应的目标语种。

[116] 从所有旁枝模型中，选取与目标语种对应的旁枝模型作为目标旁枝模型。比如说，若目标语种是语种 A，则将旁枝模型 1 作为目标旁枝模型，以此类推。

25 [117] 一个例子中，为了获取目标语种对应的旁枝模型，还可以预先配置语种与旁枝模型之间的映射关系。基于此，在得到与原始输入问题对应的目标语种之后，可以通过查询该映射关系，得到目标语种对应的旁枝模型。

[118] 方式 4：目标大语言模型包括至少两个旁枝模型，针对每个旁枝模型，该旁枝模型可以是针对一个语种的模型，且不同旁枝模型是针对不同语种的模型。

30 [119] 在此基础上，可以获取与原始输入问题对应的会话标识。若不存在该会话标识对应的旁枝模型，则可以对原始输入问题进行分词处理得到多个 token，基于多个 token 确定与原始输入问题对应的目标语种，可以从所有旁枝模型中选取与目标语种对应的旁枝模型作为目标旁枝模型，并记录会话标识与目标旁枝模型的对应关系。若存在该会话标识对应的旁枝模型，则可以直接将该会话标识对应的旁枝模型确定为原始输入问题对应的目标旁枝模型。

[120] 上述方式只是从所有旁枝模型选取目标旁枝模型的示例，对此不做限制。

35 [121] 步骤 603、向基础模型输入原始输入问题，得到原始输入问题对应的中间处理特征和原始输入问题对应的输出处理特征。

[122] 步骤 604、向目标旁枝模型发送触发指令，以使得目标旁枝模型根据触发指令从基础模型获取中间处理特征和输出处理特征，对中间处理特征进行处理得到原始输入问题对应的旁支处理特征，对旁支处理特征和输出处理特征进行处理得到原始输入问题对应的目标答案。

40 [123] 比如说，参见图 7 所示，Gating 模块可以将原始输入问题输入至基础模型，并向目标旁枝模型（如旁枝模型 2）发送触发指令，该触发指令用于触发目标旁枝模型执行推理过程，而其它旁枝模型不会执行推理过程。

45 [124] 比如说，触发指令可以是原始输入问题，即通过将原始输入问题输入至目标旁枝模型以触发目标旁枝模型执行推理过程。或者，触发指令可以是一个消息，即通过将消息输入至目标旁枝模型以触发目标旁枝模型执行推理过程。

[125] 示例性的，针对步骤 603，可以通过基础模型对原始输入问题进行处理，得到该原始输入

问题对应的中间处理特征和该原始输入问题对应的输出处理特征。

[126] 比如说,参见图7所示,基础模型的网络A在得到原始输入问题之后,基于原始输入问题进行处理得到中间处理特征X1,将中间处理特征X1输入至网络B。网络B在得到中间处理特征X1之后,基于中间处理特征X1进行处理得到中间处理特征X2,将中间处理特征X2输入至网络C。

5 网络C在得到中间处理特征X2之后,基于中间处理特征X2进行处理得到输出处理特征X3。

[127] 针对步骤604,目标旁枝模型接收到该触发指令之后,从基础模型获取中间处理特征和输出处理特征,对该中间处理特征进行处理得到该原始输入问题对应的旁支处理特征,对该旁支处理特征和该输出处理特征进行处理得到该原始输入问题对应的目标答案。

10 [128] 比如说,目标旁枝模型(如旁枝模型2)在接收到该触发指令之后,触发旁枝模型2执行推理过程。在推理过程中,旁枝模型2的网络A2获取网络A的输出特征(即中间处理特征X1)。比如说,网络A会将中间处理特征X1存储到某个缓存区,网络A2可以从该缓存区中读取中间处理特征X1。

15 [129] 网络A2对中间处理特征X1进行处理得到旁支处理特征Y1,并将旁支处理特征Y1输入至网络B2。网络B2获取网络B的输出特征(即中间处理特征X2),并对中间处理特征X2和旁支处理特征Y1进行融合,基于融合后特征进行处理得到旁支处理特征Y2,并将旁支处理特征Y2输入至网络C2。

[130] 网络C2获取网络C的输出特征(即输出处理特征X3),并对输出处理特征X3和旁支处理特征Y2进行融合,基于融合后特征进行处理得到目标答案。

[131] 显然,目标答案是原始输入问题对应的目标答案,可以输出该目标答案。

20 [132] 由以上技术方案可见,本申请实施例中,针对不同行业的原始输入问题,可以基于该行业的旁枝模型确定目标答案,从而得到准确的目标答案。能够借助目标大语言模型得到目标答案,只需要消耗少量计算资源就能够查询到目标答案。获取目标答案的时间更短,从而节省服务器的计算资源,并节省时间开销。能够快速确定出原始输入问题对应的目标答案,提供逻辑分析推理能力,利用目标大语言模型的分析能力给出准确可靠的目标答案。

25 [133] 基于同一发明构思,还提供与上述基于大语言模型的数据处理方法对应的基于大语言模型的数据处理装置及电子设备,由于数据处理装置及电子设备解决问题的原理与基于大语言模型的数据处理方法相似,因此,数据处理装置及电子设备的实施可以参见基于大语言模型的数据处理方法,重复之处不再赘述。

30 [134] 基于与上述方法同样的申请构思,本申请实施例中提出一种基于大语言模型的数据处理装置,目标大语言模型可以包括基础模型和至少两个旁枝模型,参见图8所示,为该装置的结构示意图,该装置可以包括:

[135] 获取模块81,用于获取原始输入问题,并确定所述原始输入问题对应的目标特征;从所有旁枝模型中,选取与所述目标特征对应的目标旁枝模型;

35 [136] 处理模块82,用于向所述基础模型输入所述原始输入问题,得到所述原始输入问题对应的中间处理特征和所述原始输入问题对应的输出处理特征;

[137] 向所述目标旁枝模型发送触发指令,以使得所述目标旁枝模型根据所述触发指令从所述基础模型获取所述中间处理特征和所述输出处理特征,对所述中间处理特征进行处理得到所述原始输入问题对应的旁支处理特征,对所述旁支处理特征和所述输出处理特征进行处理得到所述原始输入问题对应的目标答案。

40 [138] 所述获取模块81从所有旁枝模型中,选取与所述目标特征对应的目标旁枝模型时具体用于:若所述目标特征包括所述原始输入问题所属的目标行业,从所有旁枝模型中,选取与 said 目标行业对应的旁枝模型作为所述目标旁枝模型。

[139] 一个例子中,所述获取模块81从所有旁枝模型中,选取与 said 目标特征对应的目标旁枝模型时具体用于:

45 [140] 获取与 said 原始输入问题对应的会话标识;其中,在用户设备与 said 目标大语言模型建立连接时,为 said 用户设备分配会话标识,在 said 用户设备每次向 said 目标大语言模型提供原始输入

问题时，同步提供所述会话标识；

[141] 若存在所述会话标识对应的旁枝模型，则将所述会话标识对应的旁枝模型确定为所述原始输入问题对应的目标旁枝模型；

5 [142] 若不存在所述会话标识对应的旁枝模型，且所述目标特征包括所述原始输入问题所属的目标行业，则从所有旁枝模型中，选取与所述目标行业对应的旁枝模型作为所述目标旁枝模型，记录所述会话标识与所述目标旁枝模型的对应关系。

10 [143] 一个例子中，所述获取模块 81 确定与所述原始输入问题对应的目标行业时具体用于：获取所述原始输入问题对应的目标关键词，通过查询行业词典得到所述目标关键词对应的目标行业；其中，所述行业词典包括关键词与行业的映射关系；或，将所述原始输入问题输入至行业识别模型得到所述原始输入问题对应的目标行业；其中，所述行业识别模型用于识别行业。

[144] 一个例子中，所述获取模块 81 确定所述原始输入问题对应的目标特征时具体用于：若所述目标特征包括所述原始输入问题对应的目标语种，则对所述原始输入问题进行分词处理得到多个 token；基于所述多个 token 确定与原始输入问题对应的目标语种；

15 [145] 所述获取模块 81 从所有旁枝模型中，选取与所述目标特征对应的目标旁枝模型时具体用于：从所有旁枝模型中，选取与所述目标语种对应的旁枝模型作为所述目标旁枝模型。

[146] 一个例子中，所述装置还包括：训练模块，用于训练得到所述目标大语言模型；所述训练模块训练得到所述目标大语言模型时具体用于：

[147] 基于基础模型对应的样本数据训练得到基础模型；针对每个待训练旁枝模型，对所述基础模型与该待训练旁枝模型进行旁路连接得到行业模型；

20 [148] 基于该待训练旁枝模型对应的样本数据，对该待训练旁枝模型进行训练得到已训练的旁枝模型；其中，在训练过程中，对该待训练旁枝模型的网络参数进行调整得到已训练的旁枝模型，不对所述基础模型的网络参数进行调整；

[149] 在得到每个待训练旁枝模型对应的已训练的旁枝模型后，将所述基础模型与所有已训练的旁枝模型进行旁路连接得到所述目标大语言模型。

25 [150] 基于与上述方法同样的构思，本申请一个例子中提出一种电子设备，参见图 9 所示，电子设备包括处理器 911 和机器可读存储介质 912，机器可读存储介质 912 存储有能够被处理器 911 执行的机器可执行指令；处理器 911 用于执行机器可执行指令，以实现上述示例公开的基于大语言模型的数据处理方法。

30 [151] 一个例子中，处理器 911 可以包括一个或多个处理核心，如 4 核心处理器、8 核心处理器等。处理器 911 可以采用 DSP(Digital Signal Processing, 数字信号处理)、FPGA(Field - Programmable Gate Array, 现场可编程门阵列)、PLA (Programmable Logic Array, 可编程逻辑阵列) 中的至少一种硬件形式来实现。

35 [152] 处理器 911 也可以包括主处理器和协处理器，主处理器是用于对在唤醒状态下的数据进行处理的处理器，也称 CPU (Central Processing Unit, 中央处理器)。协处理器是用于对在待机状态下的数据进行处理的低功耗处理器。

[153] 在一些实施例中，处理器 911 可以集成有 GPU (Graphics Processing Unit, 图像处理器)，GPU 用于负责显示屏所需要显示的内容的渲染和绘制。

40 [154] 一个例子中，电子设备还可选包括有：外围设备接口 913 和至少一个外围设备。处理器 911 和外围设备接口 913 之间可以通过总线或信号线相连。各个外围设备可以通过总线、信号线或电路板与外围设备接口 913 相连。外围设备可以包括：射频电路 914 和电源 915 中的至少一种。

[155] 射频电路 914 用于接收和发射 RF (Radio Frequency, 射频) 信号，也称电磁信号。射频电路 914 通过电磁信号与通信网络以及其它通信设备进行通信。射频电路 914 将电信号转换为电磁信号进行发送，或者，将接收到的电磁信号转换为电信号。可选地，射频电路 914 包括：天线系统、RF 收发器、一个或多个放大器、调谐器、振荡器、数字信号处理器、用户身份模块卡等等。

45 [156] 射频电路 914 可以通过至少一种无线通信协议来与用户设备进行通信。该无线通信协议包括但不限于：万维网、城域网、内联网、各代移动通信网络 (2G、3G、4G 及 5G)、无线局域网和

/或 WiFi(Wireless Fidelity, 无线保真) 网络。

[157] 电源 915 用于为电子设备中的各个组件进行供电, 电源 915 可以是交流电、直流电、一次性电池或可充电电池。

5 [158] 基于与上述方法同样的申请构思, 本申请实施例还提供一种机器可读存储介质, 所述机器可读存储介质上存储有若干计算机指令, 所述计算机指令被处理器执行时, 能够实现本申请上述示例公开的基于大语言模型的数据处理方法。

10 [159] 其中, 上述机器可读存储介质可以是任何电子、磁性、光学或其它物理存储装置, 可以包含或存储信息, 如可执行指令、数据, 等等。例如, 机器可读存储介质可以是: RAM (Radom Access Memory, 随机存取存储器)、易失存储器、非易失性存储器、闪存、存储驱动器 (如硬盘驱动器)、固态硬盘、任何类型的存储盘 (如光盘、dvd 等), 或者类似的存储介质, 或者它们的组合。

[160] 上述实施例阐明的系统、装置、模块或单元, 具体可以由计算机实体实现, 或者由具有某种功能的产品来实现。一种典型的实现设备为计算机, 计算机的具体形式可以是个人计算机、膝上型计算机、蜂窝电话、相机电话、智能电话、个人数字助理、媒体播放器、导航设备、电子邮件收发设备、游戏控制台、平板计算机、可穿戴设备或者这些设备中的任意几种设备的组合。

15 [161] 为了描述的方便, 描述以上装置时以功能分为各种单元分别描述。当然, 在实施本申请时可以把各单元的功能在同一个或多个软件和/或硬件中实现。

20 [162] 本领域内的技术人员应明白, 本申请的实施例可提供为方法、系统、或计算机程序产品。因此, 本申请可采用完全硬件实施例、完全软件实施例、或结合软件和硬件方面的实施例的形式。而且, 本申请实施例可采用在一个或多个其中包含有计算机可用程序代码的计算机可用存储介质 (包括但不限于磁盘存储器、CD-ROM、光学存储器等) 上实施的计算机程序产品的形式。

25 [163] 本申请是参照根据本申请实施例的方法、设备 (系统)、和计算机程序产品的流程图和/或方框图来描述的。应理解可以由计算机程序指令实现流程图和/或方框图中的每一流程和/或方框、以及流程图和/或方框图中的流程和/或方框的结合。可提供这些计算机程序指令到通用计算机、专用计算机、嵌入式处理机或其它可编程数据处理设备的处理器以产生一个机器, 使得通过计算机或其它可编程数据处理设备的处理器执行的指令产生用于实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能的装置。

30 [164] 而且, 这些计算机程序指令也可以存储在能引导计算机或其它可编程数据处理设备以特定方式工作的计算机可读存储器中, 使得存储在该计算机可读存储器中的指令产生包括指令装置的制造品, 该指令装置实现在流程图一个流程或者多个流程和/或方框图一个方框或者多个方框中指定的功能。

[165] 这些计算机程序指令也可装载到计算机或其它可编程数据处理设备上, 使得在计算机或者其它可编程设备上执行一系列操作步骤以产生计算机实现的处理, 从而在计算机或其它可编程设备上执行的指令提供用于实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能的步骤。

35 [166] 以上所述仅为本申请的实施例而已, 并不用于限制本申请。对于本领域技术人员来说, 本申请可以有各种更改和变化。凡在本申请的精神和原理之内所作的任何修改、等同替换、改进等, 均应包含在本申请的权利要求范围之内。

权利要求书

- 1、一种基于大语言模型的数据处理方法，其特征在于，目标大语言模型包括基础模型和至少两个旁枝模型，所述方法包括：
- 获取原始输入问题，并确定所述原始输入问题对应的目标特征；
- 5 从所有旁枝模型中，选取与所述目标特征对应的目标旁枝模型；
- 向所述基础模型输入所述原始输入问题，得到所述原始输入问题对应的中间处理特征和所述原始输入问题对应的输出处理特征；
- 向所述目标旁枝模型发送触发指令，以使得所述目标旁枝模型根据所述触发指令从所述基础模型获取所述中间处理特征和所述输出处理特征，对所述中间处理特征进行处理得到所述原始输入问题对应的旁支处理特征，对所述旁支处理特征和所述输出处理特征进行处理得到所述原始输入问题对应的目标答案。
- 10 2、根据权利要求1所述的方法，其特征在于，
- 所述从所有旁枝模型中，选取与所述目标特征对应的目标旁枝模型，包括：
- 若所述目标特征包括所述原始输入问题所属的目标行业，则从所有旁枝模型中，选取与所述目标行业对应的旁枝模型作为所述目标旁枝模型。
- 15 3、根据权利要求1所述的方法，其特征在于，
- 所述从所有旁枝模型中，选取与所述目标特征对应的目标旁枝模型，包括：
- 获取与所述原始输入问题对应的会话标识；其中，在用户设备与所述目标大语言模型建立连接时，为所述用户设备分配会话标识，在所述用户设备每次向所述目标大语言模型提供原始输入问题时，同步提供所述会话标识；
- 20 若存在所述会话标识对应的旁枝模型，则将所述会话标识对应的旁枝模型确定为所述原始输入问题对应的目标旁枝模型；
- 若不存在所述会话标识对应的旁枝模型，且所述目标特征包括所述原始输入问题所属的目标行业，则从所有旁枝模型中，选取与所述目标行业对应的旁枝模型作为所述目标旁枝模型，记录所述会话标识与所述目标旁枝模型的对应关系。
- 25 4、根据权利要求2或3所述的方法，其特征在于，
- 确定与所述原始输入问题对应的目标行业的过程，包括：
- 获取所述原始输入问题对应的目标关键词，通过查询行业词典得到所述目标关键词对应的目标行业；其中，所述行业词典包括关键词与行业的映射关系；
- 30 或，将所述原始输入问题输入至行业识别模型得到所述原始输入问题对应的目标行业；其中，所述行业识别模型用于识别行业。
- 5、根据权利要求1所述的方法，其特征在于，
- 所述确定所述原始输入问题对应的目标特征，包括：若所述目标特征包括所述原始输入问题对应的目标语种，则对所述原始输入问题进行分词处理得到多个 token；基于所述多个 token 确定与所述原始输入问题对应的目标语种；
- 35

所述从所有旁枝模型中，选取与所述目标特征对应的目标旁枝模型，包括：
从所有旁枝模型中，选取与所述目标语种对应的旁枝模型作为所述目标旁枝模型。

6、根据权利要求1所述的方法，其特征在于，
针对所述目标大语言模型的训练过程，包括：

5 基于基础模型对应的样本数据训练得到基础模型；
针对每个待训练旁枝模型，对所述基础模型与该待训练旁枝模型进行旁路连接得到行业模型；
基于该待训练旁枝模型对应的样本数据，对该待训练旁枝模型进行训练得到已训练的旁枝模型；
其中，在训练过程中，对该待训练旁枝模型的网络参数进行调整得到已训练的旁枝模型，不对所述基础模型的网络参数进行调整；

10 在得到每个待训练旁枝模型对应的已训练的旁枝模型后，将所述基础模型与所有已训练的旁枝模型进行旁路连接得到所述目标大语言模型。

7、一种基于大语言模型的数据处理装置，其特征在于，目标大语言模型包括基础模型和至少两个旁枝模型，所述装置包括：

15 获取模块，用于获取原始输入问题，并确定所述原始输入问题对应的目标特征；从所有旁枝模型中，选取与所述目标特征对应的目标旁枝模型；

处理模块，用于向所述基础模型输入所述原始输入问题，得到所述原始输入问题对应的中间处理特征和所述原始输入问题对应的输出处理特征；

20 向所述目标旁枝模型发送触发指令，以使得所述目标旁枝模型根据所述触发指令从所述基础模型获取所述中间处理特征和所述输出处理特征，对所述中间处理特征进行处理得到所述原始输入问题对应的旁支处理特征，对所述旁支处理特征和所述输出处理特征进行处理得到所述原始输入问题对应的目标答案。

8、根据权利要求7所述的装置，其特征在于，所述获取模块从所有旁枝模型中，选取与所述目标特征对应的目标旁枝模型时具体用于：

25 若所述目标特征包括所述原始输入问题所属的目标行业，则从所有旁枝模型中，选取与所述目标行业对应的旁枝模型作为所述目标旁枝模型。

9、根据权利要求7所述的装置，其特征在于，所述获取模块从所有旁枝模型中，选取与所述目标特征对应的目标旁枝模型时具体用于：

30 获取与所述原始输入问题对应的会话标识；其中，在用户设备与所述目标大语言模型建立连接时，为所述用户设备分配会话标识，在所述用户设备每次向所述目标大语言模型提供原始输入问题时，同步提供所述会话标识；

若存在所述会话标识对应的旁枝模型，则将所述会话标识对应的旁枝模型确定为所述原始输入问题对应的目标旁枝模型；

35 若不存在所述会话标识对应的旁枝模型，且所述目标特征包括所述原始输入问题所属的目标行业，则从所有旁枝模型中，选取与所述目标行业对应的旁枝模型作为所述目标旁枝模型，记录所述会话标识与所述目标旁枝模型的对应关系。

- 10、根据权利要求 8 或 9 所述的装置，其特征在于，
所述获取模块确定与所述原始输入问题对应的目标行业时具体用于：
获取所述原始输入问题对应的目标关键词，通过查询行业词典得到所述目标关键词对应的目标行业；其中，所述行业词典包括关键词与行业的映射关系；
- 5 或，将所述原始输入问题输入至行业识别模型得到所述原始输入问题对应的目标行业；其中，所述行业识别模型用于识别行业。
- 11、根据权利要求 7 所述的装置，其特征在于，
所述获取模块确定所述原始输入问题对应的目标特征时具体用于：若所述目标特征包括所述原始输入问题对应的目标语种，则对所述原始输入问题进行分词处理得到多个 token；基于所述多个
10 token 确定与原始输入问题对应的目标语种；
所述获取模块从所有旁枝模型中，选取与所述目标特征对应的目标旁枝模型时具体用于：从所有旁枝模型中，选取与所述目标语种对应的旁枝模型作为所述目标旁枝模型。
- 12、根据权利要求 7 所述的装置，其特征在于，所述装置还包括：
训练模块，用于训练得到所述目标大语言模型；
- 15 其中，所述训练模块训练得到所述目标大语言模型时具体用于：
基于基础模型对应的样本数据训练得到基础模型；
针对每个待训练旁枝模型，对所述基础模型与该待训练旁枝模型进行旁路连接得到行业模型；
基于该待训练旁枝模型对应的样本数据，对该待训练旁枝模型进行训练得到已训练的旁枝模型；
其中，在训练过程中，对该待训练旁枝模型的网络参数进行调整得到已训练的旁枝模型，不对所述
20 基础模型的网络参数进行调整；
在得到每个待训练旁枝模型对应的已训练的旁枝模型后，将所述基础模型与所有已训练的旁枝模型进行旁路连接得到所述目标大语言模型。
- 13、一种电子设备，其特征在于，包括：处理器和机器可读存储介质，所述机器可读存储介质存储有能够被所述处理器执行的机器可执行指令；所述处理器用于执行机器可执行指令，以实现权
25 利要求 1-6 任一所述的方法。