A11_常见问题

文档版本 01

发布日期 2024-12-02

华为云计算技术有限公司

商标声明


和其他华为商标均为华为技术有限公司的商标。
本文档提及的其他所有商标或注册商标，由各自的所有人拥有。

注意

您购买的产品、服务或特性等应受华为云计算技术有限公司商业合同和条款的约束，本文档中描述的全部或部分产品、服务或特性可能不在您的购买或使用范围之内。除非合同另有约定，华为云计算技术有限公司对本文档内容不做任何明示或暗示的声明或保证。

由于产品版本升级或其他原因，本文档内容会不定期进行更新。除非另有约定，本文档仅作为使用指导，本文档中的所有陈述、信息和建议不构成任何明示或暗示的担保。

华为云计算技术有限公司

地址： 贵州省贵安新区黔中大道交兴功路华为云数据中心 邮编：550029

网址： https://www.huaweicloud.com/

目 录

1 大模型概念类问题

1.1 如何对盘古大模型的安全性展开评估和防护

盘古大模型的安全性主要从以下方面考虑：

- 数据安全和隐私保护：大模型涉及大量训练数据，这些数据是重要资产。为确保数据安全，需在数据和模型训练的全生命周期内，包括数据提取、加工、传输、训练、推理和删除的各个环节，提供防篡改、数据隐私保护、加密、审计和数据主权保护等机制。在训练和推理过程中，通过数据脱敏、隐私计算等技术手段识别并保护敏感数据，有效防止隐私泄露，保障个人隐私数据安全。
- 内容安全：通过预训练和强化学习价值观提示（prompt），构建正向的意识形态。通过内容审核模块过滤违法及违背社会道德的有害信息。
- 模型安全：通过模型动态混淆技术，使模型在运行过程中保持混淆状态，有效防止结构信息和权重信息在被窃取后暴露。
- 系统安全：通过网络隔离、身份认证和鉴权、Web安全等技术保护大模型系统安全，增强自身防护能力，以抵御外部安全攻击。

1.2 训练智能客服系统大模型需考虑哪些方面

根据智能客服场景，建议从以下方面考虑：

- 根据企业实际服务的场景和积累的数据量，评估是否需要构建行业模型，如电商、金融等。
- 根据每个客户的金牌客服话术，可以对对话模型进行有监督微调，进一步优化其性能。
- 根据每个客户的实际对话知识，如帮助文档、案例库和FAQ库等，可以使用“先搜后推”的解决方案。客户的文档库可以实时更新，大模型的应答可以无缝实时更新。（搜索+大模型解决方案）

2 大模型微调训练类问题

2.1 无监督领域知识数据量无法支持增量预训练，如何进行模型学习

一般来说，建议采用增量预训练的方式让模型学习领域知识，但预训练对数据量的要求较大，如果您的无监督文档量级过小，达不到预训练要求，您可以通过一些手段将其转换为有监督数据，再将转换后的领域知识与目标任务数据混合，使用微调的方式让模型学习。

这里提供了一些将无监督数据转换为有监督数据的方案，供您参考：

- **基于规则构建：**您可以通过采用一些简单的规则来构建有监督数据。比如：

  表 2-1 采用规则将无监督数据构建为有监督数据的常用方法

规则场景	说明
文本生成：根据标题、关键词、简介生成段落。	若您的无监督文档中含标题、关键词、简介等结构化信息，可以将有监督的问题设置为“请根据标题xxx/关键性xxx/简介xxx，生成一段不少于xx个字的文本。”，将回答设置为符合要求的段落。
续写：根据段落的首句、首段续写成完整的段落。	若您的无监督文档没有任何结构化信息，可以将有监督的问题设置为“以下是一篇文章的第一个句子：xxx/第一段落：xxx。请根据以上的句子/段落，续写为一段不少于xx个字的文本。”，再将回答设置为符合要求的段落。
扩写：根据段落的其中一句或者一段续写成完整的段落。	若您的无监督文档没有任何结构化信息，可以将有监督的问题设置为“以下是一篇文章的某个句子：xxx/某个段落：xxx。请根据以上的句子/段落，扩写成一段不少于xx个字的文本。”，再将回答设置为符合要求的段落。

规则场景	说明
填空：从段落随机掩盖一个或多个词语、句子、段落，再将段落完形填空。	若您的无监督文档没有任何结构化信息，可以将有监督的问题设置为“以下的文章中有一些词语/句子/段落缺失，文章如下：xxx。请结合文章内容，将缺失的信息补充完整。”，再将回答设置为符合要求的信息。

使用规则构建的优点是快速且成本低，缺点是数据多样性较低。

- **基于大模型的数据泛化：**您可以通过调用大模型（比如盘古提供的任意一个规格的基础功能模型）来获取有监督场景。一个比较常见的方法是，将无监督的文本按照章节、段落、字符数进行切片，让模型基于这个片段生成问答对，再将段落、问题和答案三者组装为有监督数据。使用模型构建的优点是数据丰富度更高，缺点是成本较高。

  说明

  当您将无监督数据构建为有监督数据时，请尽可能保证数据的多样性。建议将不同文本构建为不同的场景，甚至将同一段文本构建为多个不同的场景。

  不同规格的模型支持的长度不同，当您将无监督数据构建为有监督数据时，请确保数据长度符合模型长度限制。

2.2 如何调整训练参数，使盘古大模型效果最优

模型微调参数的选择没有标准答案，不同的场景，有不同的调整策略。一般微调参数的影响会受到以下几个因素的影响：

- **目标任务的难度：**如果目标任务的难度较低，模型能较容易的学习知识，那么少量的训练轮数就能达到较好的效果。反之，若任务较复杂，那么可能就需要更多的训练轮数。
- **数据量级：**如果微调数据很多，从客观上来说越多的数据越能接近真实分布，那么可以使用较大的学习率和较大的批量大小，以提高训练效率。如果微调数据量相对较少，则可以使用较小的学习率和较小的数据批量大小，避免过拟合。
- **通用模型的规格：**如果模型参数规模较小，那么可能需要较大的学习率和较大的批量大小，以提高训练效率。如果规模较大，那么可能需要较小的学习率和较小的批量大小，防止内存溢出。

这里提供了一些微调参数的建议值和说明，供您参考：

表 2-2 微调参数的建议和说明

训练参数	范围	建议值	说明
训练轮数（epoch）	1~50	2/4/8/10	训练轮数是指需要完成全量训练数据集训练的次数。训练轮数越大，模型学习数据的迭代步数就越多，可以学得更深入，但过高会导致过拟合；训练轮数越小，模型学习数据的迭代步数就越少，过低则会导致欠拟合。 您可根据任务难度和数据规模进行调整。一般来说，如果目标任务的难度较大或数据量级很小，可以使用较大的训练轮数，反之可以使用较小的训练轮数。 如果您没有专业的调优经验，可以优先使用平台提供的默认值，再结合训练过程中模型的收敛情况动态调整。
数据批量大小（batch_size）	>=1	4/8	数据批量大小是指对数据集进行分批读取训练时，所设定的每个批次数据大小。批量大小越大，训练速度越快，但是也会占用更多的内存资源，并且可能导致收敛困难或者过拟合；批量大小越小，内存消耗越小，但是收敛速度会变慢，同时模型更容易受到数据噪声的影响，从而导致模型收敛困难。 您可根据数据和模型的规模进行调整。一般来说，如果数据量级很小或模型参数规模很大，可以使用较小的批量大小，反之可以使用较大的批量大小。 如果您没有专业的调优经验，可以优先使用平台提供的默认值，再结合训练过程中的实际情况动态调整。
学习率（learning_rate）	0~1	1e-6~5e-4	学习率是在梯度下降的过程中更新权重时的超参数，过高会导致模型在最优解附近震荡，甚至跳过最优解，无法收敛，过低则会导致模型收敛速度过慢。 您可根据数据和模型的规模进行调整。一般来说，如果数据量级很小或模型参数规模很大，可以使用较小的学习率，反之可以使用较大的学习率。 如果您没有专业的调优经验，可以优先使用平台提供的默认值，再结合训练过程中模型的收敛情况动态调整。
学习率衰减比率（learning_rate_decay_ratio）	0~1	0.01~0.1	学习率衰减比率用于设置训练过程中的学习率衰减的最小值。计算公式为：最小学习率=学习率*学习率衰减比率。

说明

参数的选择没有标准答案，您需要根据任务的实际情况进行调整，以上建议值仅供参考。

2.3 如何判断盘古大模型训练状态是否正常

判断训练状态是否正常，通常可以通过观察训练过程中Loss（损失函数值）的变化趋势。损失函数是一种衡量模型预测结果和真实结果之间的差距的指标，正常情况下越小越好。

您可以从平台的训练日志中获取到每一步的Loss，并绘制成Loss曲线，来观察其变化趋势。一般来说，一个正常的Loss曲线应该是单调递减的，即随着训练的进行，Loss值不断减小，直到收敛到一个较小的值。

以下给出了几种正常的Loss曲线形式：

图 2-1 正常的 Loss 曲线：平滑下降

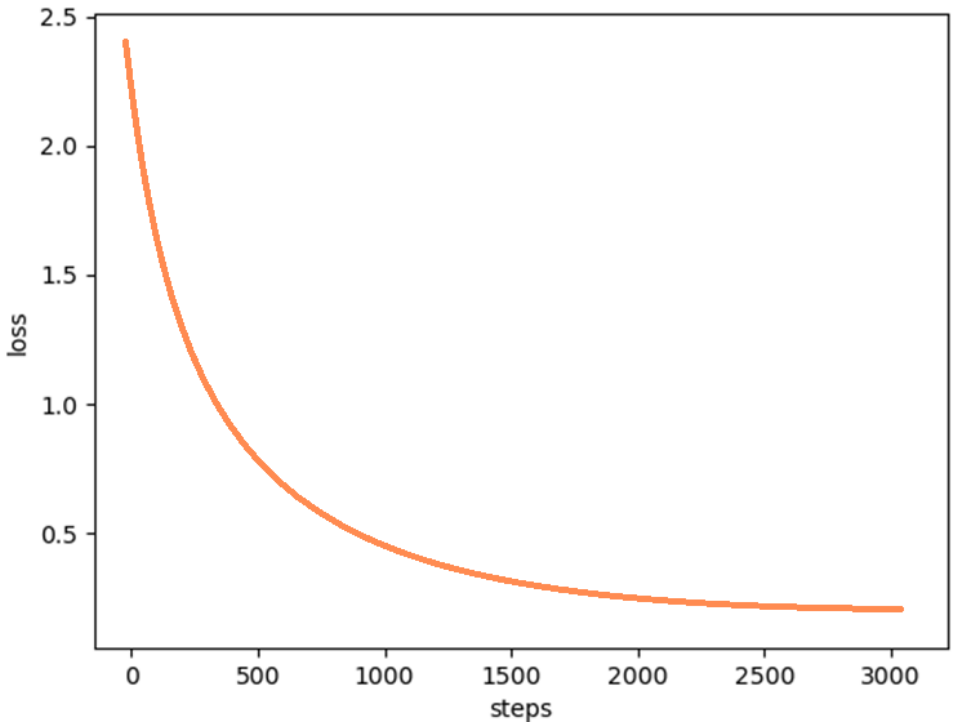
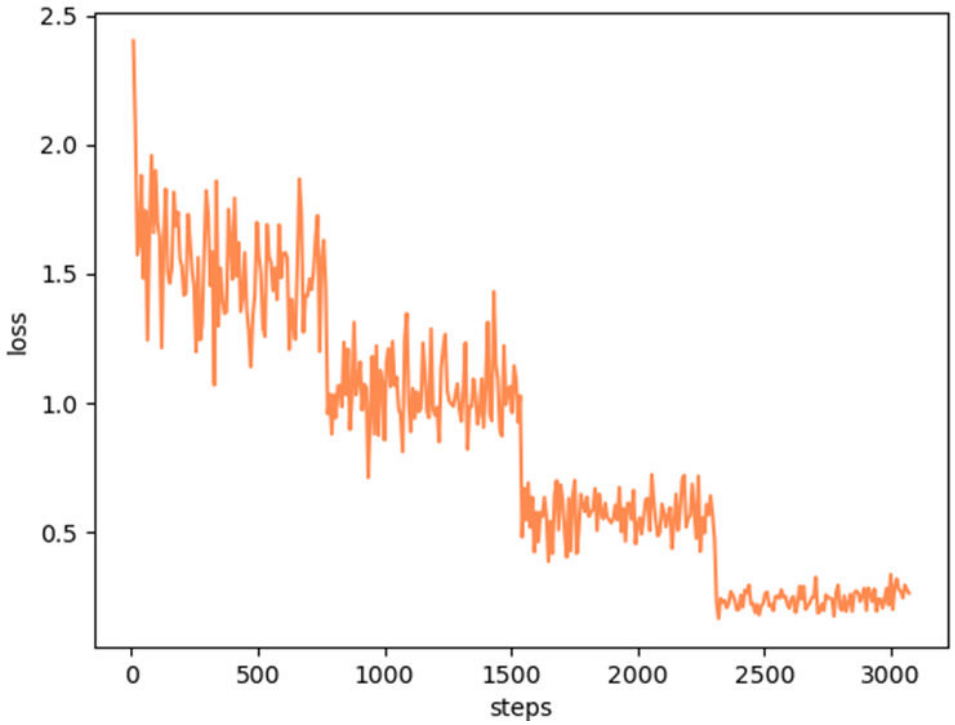


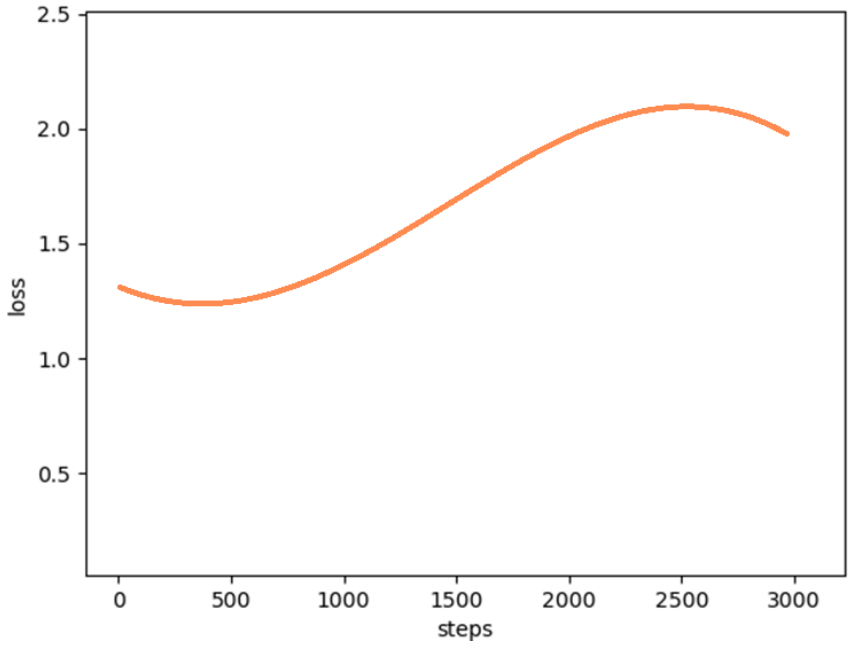
图 2-2 正常的 Loss 曲线：阶梯下降



如果您发现Loss曲线出现了以下几种情况，可能意味着模型训练状态不正常：

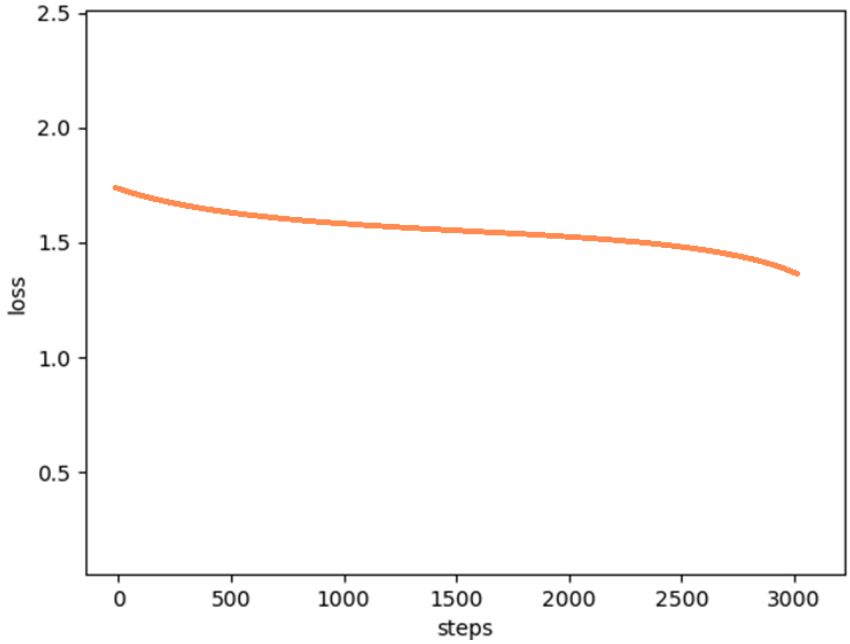
- **Loss曲线上升：**Loss上升的原因可能是由于数据质量差，或者学习率设置得过大，使得模型在最优解附近震荡，甚至跳过最优解，导致无法收敛。您可以尝试提升数据质量或者减小学习率的方式来解决。

图 2-3 异常的 Loss 曲线：上升



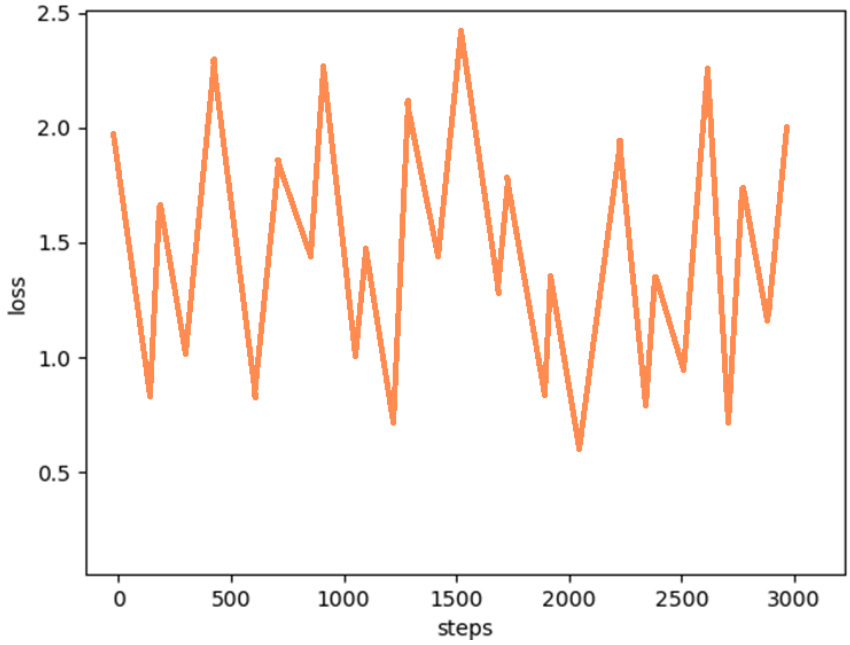
- **Loss曲线平缓，保持高位：**Loss保持平缓且保持高位不下降的原因可能是由于目标任务的难度较大，或者模型的学习率设置得过小，导致模型的收敛速度太慢，无法达到最优解。您可以尝试增大训练轮数或者增大学习率的方式来解决。

图 2-4 异常的 Loss 曲线：平缓且保持高位



- **Loss曲线异常抖动：**Loss曲线异常抖动的原因可能是由于训练数据质量差，比如数据存在噪声或者分布不均衡，导致训练过程不稳定。你可以尝试提升数据质量的方式来解决。

图 2-5 异常的 Loss 曲线：异常抖动



2.4 如何评估微调后的盘古大模型是否正常

评估模型效果的方法有很多，通常可以从以下几个方面来评估模型训练效果：

- **Loss曲线：**通过Loss曲线的变化趋势来评估训练效果，确认训练过程是否出现了过拟合或欠拟合等异常情况。
- **模型评估：**使用平台的“模型评估”功能，“模型评估”将对您之前上传的测试集进行评估。通过查看测试集样本的PPL、BLEU和ROUGE等指标，进行横向（相同训练数据+不同规格的通用模型）或纵向（不同训练数据训练的多个模型版本）对比来判断训练过程是否出现了问题。
- **人工评测：**您可以采用人工评测的方式，参照目标任务构造评测集，通过横向或纵向评估评测集的方式来验证模型效果。

2.5 如何调整推理参数，使盘古大模型效果最优

推理参数（解码参数）是一组用于控制模型生成预测结果的参数，其可以用于控制模型生成结果的样式，如长度、随机性、创造性、多样性、准确性和丰富度等等。

当前，平台支持的推理参数包括：温度、核采样以及话题重复度控制，如下提供了这些推理参数的建议值和说明，供您参考：

表 2-3 推理参数的建议和说明

推理参数	范围	建议值	说明
温度（temperature）	0~1	0.3	温度主要用于控制模型输出的随机性和创造性。温度越高，输出的随机性和创造性越高；温度越低，输出结果越可以被预测，确定性相对也就越高。 您可根据真实的任务类型进行调整。一般来说，如果目标任务的需要生成更具创造性的内容，可以使用较高的温度，反之如果目标任务的需要生成更为确定的内容，可以使用较低的温度。 请注意，温度和核采样的作用相近，在实际使用中，为了更好观察是哪个参数对结果造成的影响，因此不建议同时调整这两个参数。 如果您没有专业的调优经验，可以优先使用建议，再结合推理的效果动态调整。
核采样（top_p）	0~1	1	核采样主要用于控制模型输出的多样性。核采样值越大，输出的多样性越高；核采样值越小，输出结果越可以被预测，确定性相对也就越高。 您可根据真实的任务类型进行调整。一般来说，如果目标任务的需要生成更具多样性的内容，可以使用较大的核采样，反之如果目标任务的需要生成更为确定的内容，可以使用较小的核采样。 请注意，温度和核采样的作用相近，在实际使用中，为了更好观察是哪个参数对结果造成的影响，因此不建议同时调整这两个参数。 如果您没有专业的调优经验，可以优先使用建议，再结合推理的效果动态调整。
话题重复度控制（presence_penalty）	-2~2	0	话题重复度控制主要用于控制模型输出的话题重复程度。 参数设置正值，模型倾向于生成新的、未出现过的内容；参数设置负值，倾向于生成更加固定和统一的内容。 如果您没有专业的调优经验，可以优先使用建议，再结合推理的效果动态调整。

为了让您更好的理解这几个参数的作用，如下列举了一些常见场景，以及对应的调参指导，供您参考：

- **文本生成：**对于文本生成场景（宣传文案生成、信稿文本生成、文学创作等），通常希望生成的文本有一点的多样性，建议在保证不过于随机的基础上，增大“温度”或“核采样”的值（二者选其一调整）。若发现生成的文本过于发散，可以降低“话题重复度控制”的值，保证内容统一；反之若发现内容过于单一，甚至出现了复读机式的重复内容生成，则需要增加“话题重复度控制”的值。

- **知识问答：**对于文本生成场景（开放问答、基于搜索内容回答等），从客观上来说，回答需要是确定且唯一的，建议降低“温度”或“核采样”的值（二者选其一调整）。若需要每次生成完全相同的回答，可以将“温度”置为0。

说明

参数的选择没有标准答案，您需要根据任务的实际情况进行调整，以上建议值仅供参考。

2.6 为什么微调后的盘古大模型总是重复相同的回答

当您将微调的模型部署以后，输入一个与目标任务同属的问题，模型生成了复读机式的结果，即回答中反复出现某一句话或某几句话。这种情况可能是由于以下几个原因导致的，建议您依次排查：

- **推理参数设置：**请检查推理参数中的“话题重复度控制”或“温度”或“核采样”等参数的设置，适当增大其中一个参数的值，可以提升模型回答的多样性。
- **数据质量：**请检查训练数据中是否存在文本重复的异常数据，可以通过规则进行清洗。
- **训练参数设置：**若数据质量存在问题，且因训练参数设置的不合理而导致过拟合，该现象会更加明显。请检查训练参数中的 “训练轮次”或“学习率”等参数的设置，适当降低这些参数的值，降低过拟合的风险。

2.7 为什么微调后的盘古大模型的回答中会出现乱码

当您将微调的模型部署以后，输入一个与目标任务同属的问题，模型生成的结果中出现了其他语言、异常符号、乱码等字符。这种情况可能是由于以下几个原因导致的，建议您依次排查：

- **数据质量：**请检查训练数据中是否存在包含异常字符的数据，可以通过规则进行清洗。
- **训练参数设置：**若数据质量存在问题，且因训练参数设置的不合理而导致过拟合，该现象会更加明显。请检查训练参数中的 “训练轮次”或“学习率”等参数的设置，适当降低这些参数的值，降低过拟合的风险。
- **推理参数设置：**请检查推理参数中的“温度”或“核采样”等参数的设置，适当减小其中一个参数的值，可以提升模型回答的确定性，避免生成异常内容。

2.8 为什么微调后的盘古大模型的回答会异常中断

当您将微调的模型部署以后，输入一个与目标任务同属的问题，模型生成的结果不完整，出现了异常截断。这种情况可能是由于以下几个原因导致的，建议您依次排查：

- **推理参数设置：**请检查推理参数中的“最大Token限制”参数的设置，适当增加该参数的值，可以增大模型回答生成的长度，避免生成异常截断。请注意，该参数值存在上限，请结合目标任务的实际需要以及模型支持的长度限制来调整。
- **模型规格：**不同规格的模型支持的长度不同，若目标任务本身需要生成的长度已经超过模型上限，建议您替换可支持更长长度的模型。
- **数据质量：**请检查训练数据中是否存在包含异常截断的数据，可以通过规则进行清洗。

2.9 为什么微调后的盘古大模型只能回答训练样本中的问题

当您将微调的模型部署以后，输入一个已经出现在训练样本中的问题，模型生成的结果很好，一旦输入了一个从未出现过的数据（目标任务相同），回答却完全错误。这种情况可能是由于以下几个原因导致的，建议您依次排查：

- **训练参数设置：**您可以通过绘制Loss曲线查询来确认模型的训练过程是否出现了问题，这种情况大概率是由于训练参数设置的不合理而导致了过拟合。请检查训练参数中的 “训练轮次”或“学习率”等参数的设置，适当降低这些参数的值，降低过拟合的风险。
- **数据质量：**请检查训练数据的质量，若训练样本出现了大量重复数据，或者数据多样性很差，则会加剧该现象。

2.10 为什么在微调后的盘古大模型中输入训练样本问题，回答完全不同

当您将微调的模型部署以后，输入一个已经出现在训练样本中，或虽未出现但和训练样本差异很小的问题，回答完全错误。这种情况可能是由于以下几个原因导致的，建议您依次排查：

- **训练参数设置：**您可以通过绘制Loss曲线查询来确认模型的训练过程是否出现了问题，这种情况大概率是由于训练参数设置的不合理而导致了欠拟合，模型没有学到任何知识。请检查训练参数中的 “训练轮次”或“学习率”等参数的设置，适当增大“训练轮次”的值，或根据实际情况调整“学习率”的值，帮助模型更好收敛。
- **数据质量：**请检查训练数据的质量，若训练样本和目标任务不一致或者分布差异较大，则会加剧该现象。

2.11 为什么微调后的盘古大模型评估结果很好，但实际场景表现很差

当您在微调过程中，发现模型评估的结果很好，一旦将微调的模型部署以后，输入一个与目标任务同属的问题，回答的结果却不理想。这种情况可能是由于以下几个原因导致的，建议您依次排查：

- **测试集质量：**请检查测试集的目标任务和分布与实际场景是否一致，质量较差的测试集无法反映模型的真实结果。
- **数据质量：**请检查训练数据的质量，若训练样本和目标任务不一致或者分布差异较大，则会加剧该现象。此外，若可预见实际场景会不断发生变化，建议您定期更新训练数据，对模型进行微调更新。

2.12 为什么多轮问答场景的盘古大模型微调效果不好

当您的目标任务是多轮问答，并且使用了多轮问答数据进行微调，微调后却发现多轮回答的效果不理想。这种情况可能是由于以下几个原因导致的，建议您依次排查：

- **数据格式：**多轮问答场景需要按照指定的数据格式来构造，问题需要拼接上历史所有轮对话的问题和回答。比如，当前是第三轮对话，数据中的问题字段需要包

含第一轮的问题、第一轮的回答、第二轮的问题、第二轮的回答以及第三轮的问题，答案字段则为第三轮的回答。以下给出了几条多轮问答的数据样例供您参考：

原始对话示例：

```
A：你是谁?
B：您好，我是盘古大模型。
A：你可以做什么?
B：我可以做很多事情，比如xxxx
A：你可以讲个笑话吗?
B：当然可以啦，以下是xxxx
A：可以把这个笑话改成xxxx
B：好的，以下是修改后的xxxx
```

拼接后的微调数据格式示例：

```
{"context": ["你是谁？ ", "您好，我是盘古大模型。", "你可以做什么？ ", "我可以做很多事情，比如xxxx", "你可以讲个笑话吗？ ", "当然可以啦，以下是xxxx", "可以把这个笑话改成xxxx"], "target": "好的，以下是修改后的xxxx"}
```

说明

多轮问答场景的输入（“context”字段）请务必使用“[问题, 回答, 问题, 回答, 问题, ……]”的方式来构造，若您的数据是同一个角色连续多次对话的“多轮问题”，可以将同一个角色的对话采用某个分隔符拼接到一个字符串中。例如：

原始对话示例：

```
A：xxx号话务员为您服务！
A：先生您好，有什么可以帮助您的?
B：你好，是这样的
B：我家里上不了网了
B：网连不上
A：先生，您家的网络无法连接是吗
A：请问您尝试重新插拔网线吗?
B：是的，我试了
B：还是不行
```

拼接后的微调数据格式示例：

```
{"context": ["xxx号话务员为您服务！ 先生您好，有什么可以帮助您的？ ", "你好，是这样的 我家里上不了网了 网连不上", "先生，您家的网络无法连接是吗 请问您尝试重新插拔网线吗？ "], "target": "是的，我试了 还是不行"}
```

- **数据质量：**若数据格式没有问题，仍然发现模型效果不好，您可以根据具体问题针对性的提升您的数据质量。比如，随着对话轮数的增加，模型出现了遗忘，可以检查构造的训练数据中轮数是否普遍较少，建议根据实际情况增加数据中的对话轮数。

2.13 数据量足够，为什么盘古大模型微调效果仍然不好

这种情况可能是由于以下原因导致的，建议您排查：

- **数据质量：**请检查训练数据的质量，若训练样本和目标任务不一致或者分布差异较大、样本中存在异常数据、样本的多样性较差，都将影响模型训练的效果，建议提升您的数据质量。

2.14 数据量和质量均满足要求，为什么盘古大模型微调效果不好

这种情况可能是由于以下原因导致的，建议您排查：

- **训练参数设置：**您可以通过绘制Loss曲线查询来确认模型的训练过程是否出现了问题，这种情况大概率是由于训练参数设置的不合理而导致了欠拟合或过拟合。

请检查训练参数中的 “训练轮次”或“学习率”等参数的设置，根据实际情况调整训练参数，帮助模型更好学习。

- **Prompt设置：**请检查您使用的Prompt，对于同一个目标任务，建议在推理阶段使用和训练数据相同或相似的PROMPT，才能发挥出模型的最佳效果。
- **模型规格：**理论上模型的参数规模越大，模型能学到的知识就越多，能学会的知识就更难，若目标任务本身难度较大，建议您替换参数规模更大的模型。

3 大模型使用类问题

3.1 盘古大模型是否可以自定义人设

大模型支持设置人设，在用户调用对话问答（chat/completions）API时，可以将“role”参数设置为system，让模型按预设的人设风格回答问题。例如，以下示例要求模型以幼儿园老师的风格回答问题。

```
{
    "messages": [
     {
         "role": "system",
         "content": "请用幼儿园老师的口吻回答问题，注意语气温和亲切，通过提问、引导、赞美等方式，激发学生的思维和想象力。"
      },
     {
         "role": "user",
         "content": "介绍下长江，以及长江中典型的鱼类"
     }
     ],
    "temperature": 0.9,
    "max_tokens": 600
}
```

