模型评估基础

CS2916 大语言模型

飲水思源 愛國榮校

https://plms.ai/teaching/index.html

课程要求

□ 评估基础知识

- ■ 掌握评估基准的概念和构建的方法
- ■ 理解评估基准对AI技术发展的重要意义
- ■ 掌握常见的自动评价指标
- ■ 理解元评估概念和评估可靠性度量方法
- ■ 理解评估的可解释性

□ 大模型的评估

- ■ 掌握预训练阶段评估的方法
- ■ 掌握对齐阶段评估的方法

评估在机器学习中的定位

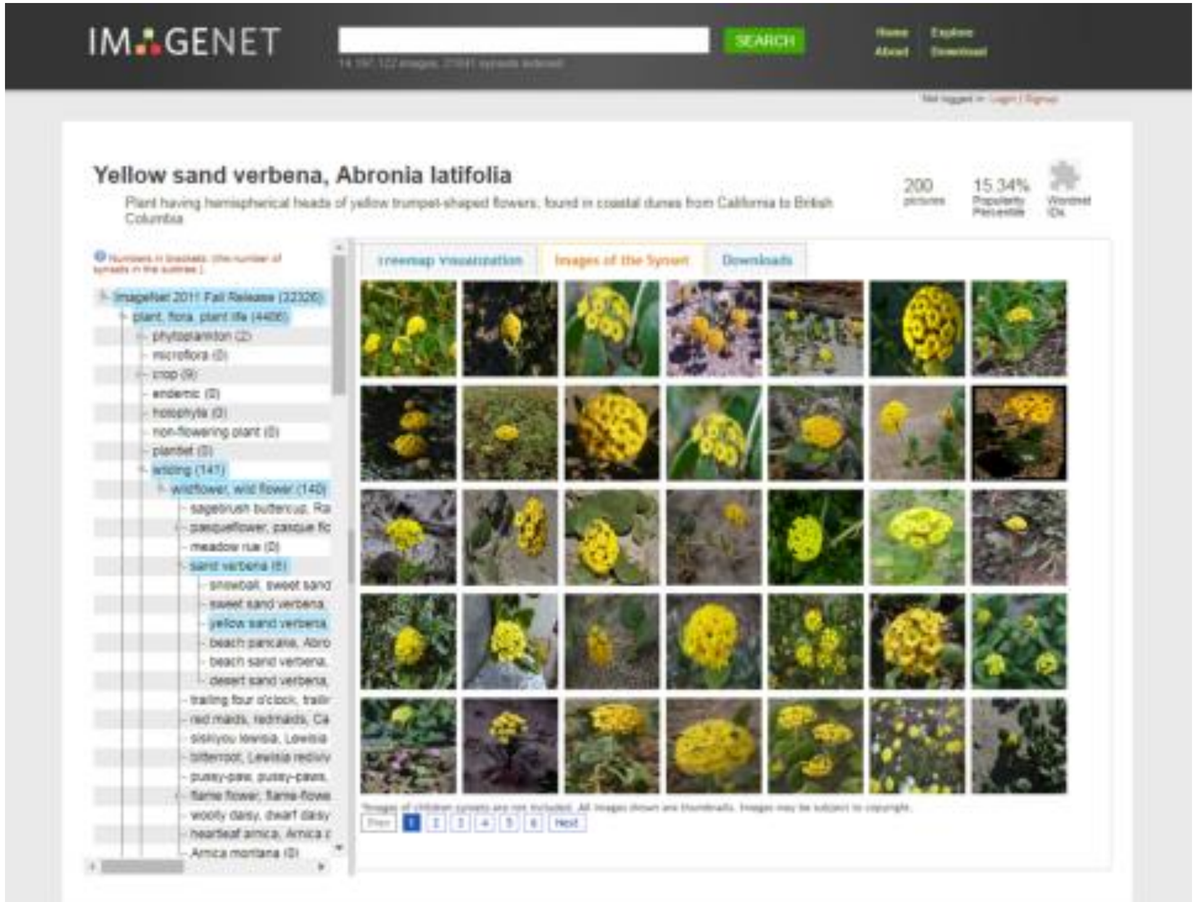
□ 机器学习推动了进步

评估在机器学习中的定位

- □ 机器学习推动了进步
- □ 评估基准（Evaluation Benchmark）设定方向
  - ■ What’ s next: 哪些任务是重要的?
  - ■ What’ s left: 我们现在的处境如何，在任务之内达到了吗?

例子：ImageNet

- □ 特点：
  - ■ 超过1400万标注的图片
  - ■ 超过两万个不同的类别
- □ 培育模型
  - ■ AlexNet, ResNet



14,197,122 images, 21841 synsets indexed

Not logged in. Login | Signup

Yellow sand verbena, Abronia latifolia

Plant having hemispherical heads of yellow trumpet-shaped flowers; found in coastal dunes from California to British Columbia

200 pictures

15.34% Popularity Percentile

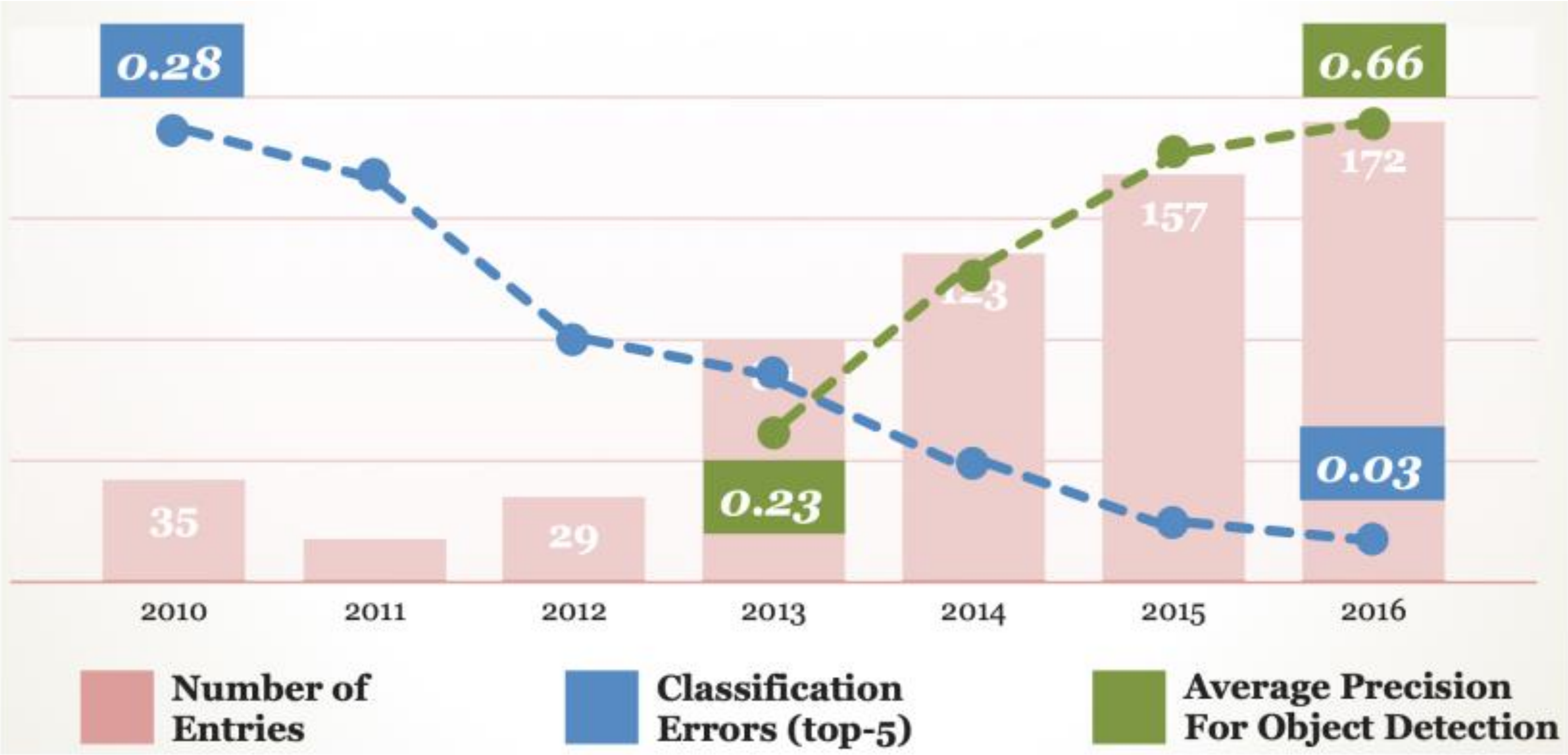
Wordnet IDs

Numbers in brackets: (the number of synsets in the subtree).

- ImageNet 2011 Fall Release (32326)
  - plant, flora, plant life (4486)
    - phytoplankton (2)
    - microflora (0)
    - crop (9)
    - endemic (0)
    - holophyte (0)
    - non-flowering plant (0)
    - plantlet (0)
    - wilding (147)
      - wildflower, wild flower (140)
        - sagebrush buttercup, Ra
        - pasqueflower, pasque flo
        - meadow rue (0)
        - sand verbena (6)
          - snowball, sweet sand
          - sweet sand verbena,
          - yellow sand verbena,
          - beach pancake, Abro
          - beach sand verbena,
          - desert sand verbena,
        - trailing four o'clock, trail
        - red maids, redmaids, Ca
        - siskiyou lewisia, Lewisia
        - bitterroot, Lewisia rediviv
        - pussy-paw, pussy-paws,
        - flame flower, flame-flowe
        - woolly daisy, dwarf daisy
        - heartleaf arnica, Arnica c
        - Arnica montana (0)

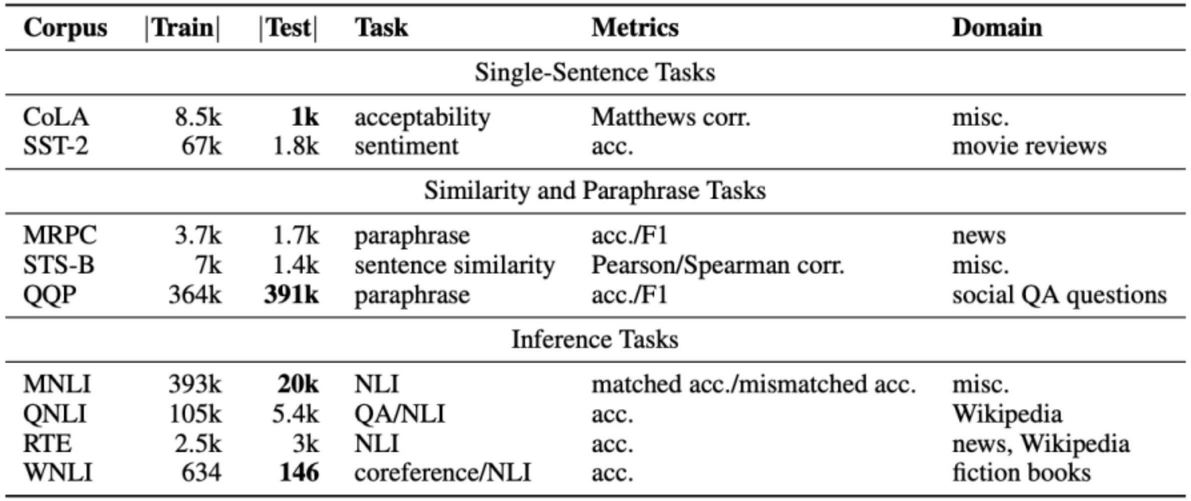
Images of children synsets are not included. All images shown are thumbnails. Images may be subject to copyright.

例子：ImageNet


0.28
0.66
172
157
123
35
29
0.23
0.03
2010
2011
2012
2013
2014
2015
2016
🟥 Number of Entries
🟦 Classification Errors (top-5)
🟩 Average Precision For Object Detection

例子：GLUE

- □ 特点：
  - ■ 多个任务组成
  - ■ 分类、理解型任务
- □ 培育模型
  - ■ BERT等大模型



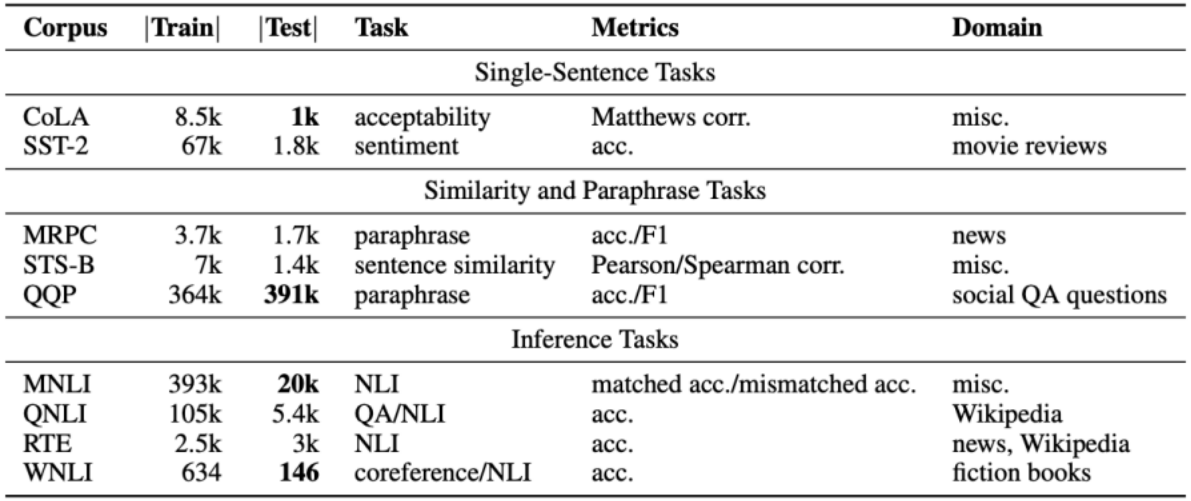
Corpus	**|Train|**	**|Test|**	**Task**	**Metrics**	**Domain**
			Single-Sentence Tasks		
CoLA	8.5k	**1k**	acceptability	Matthews corr.	misc.
SST-2	67k	1.8k	sentiment	acc.	movie reviews
			Similarity and Paraphrase Tasks		
MRPC	3.7k	1.7k	paraphrase	acc./F1	news
STS-B	7k	1.4k	sentence similarity	Pearson/Spearman corr.	misc.
QQP	364k	**391k**	paraphrase	acc./F1	social QA questions
			Inference Tasks		
MNLI	393k	**20k**	NLI	matched acc./mismatched acc.	misc.
QNLI	105k	5.4k	QA/NLI	acc.	Wikipedia
RTE	2.5k	3k	NLI	acc.	news, Wikipedia
WNLI	634	**146**	coreference/NLI	acc.	fiction books

GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding, Wang et al.2018

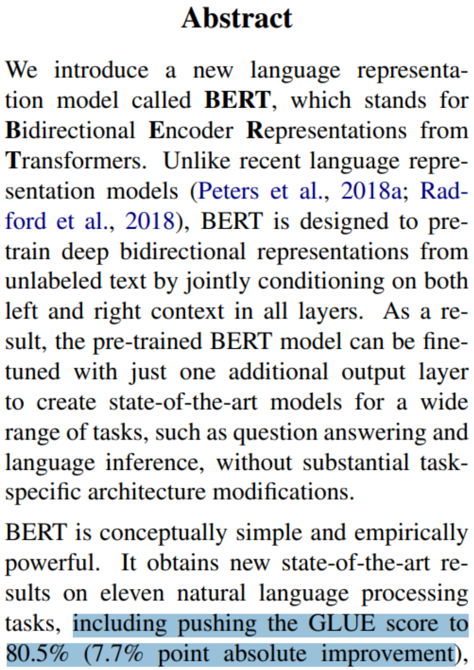
BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, Devlin et al.2018

例子：GLUE

- □ 特点：
  - ■ 多个任务组成
  - ■ 分类、理解型任务
- □ 培育模型
  - ■ BERT等大模型



Corpus	|Train|	|Test|	Task	Metrics	Domain
			Single-Sentence Tasks		
CoLA	8.5k	**1k**	acceptability	Matthews corr.	misc.
SST-2	67k	1.8k	sentiment	acc.	movie reviews
			Similarity and Paraphrase Tasks		
MRPC	3.7k	1.7k	paraphrase	acc./F1	news
STS-B	7k	1.4k	sentence similarity	Pearson/Spearman corr.	misc.
QQP	364k	**391k**	paraphrase	acc./F1	social QA questions
			Inference Tasks		
MNLI	393k	**20k**	NLI	matched acc./mismatched acc.	misc.
QNLI	105k	5.4k	QA/NLI	acc.	Wikipedia
RTE	2.5k	3k	NLI	acc.	news, Wikipedia
WNLI	634	**146**	coreference/NLI	acc.	fiction books



Abstract

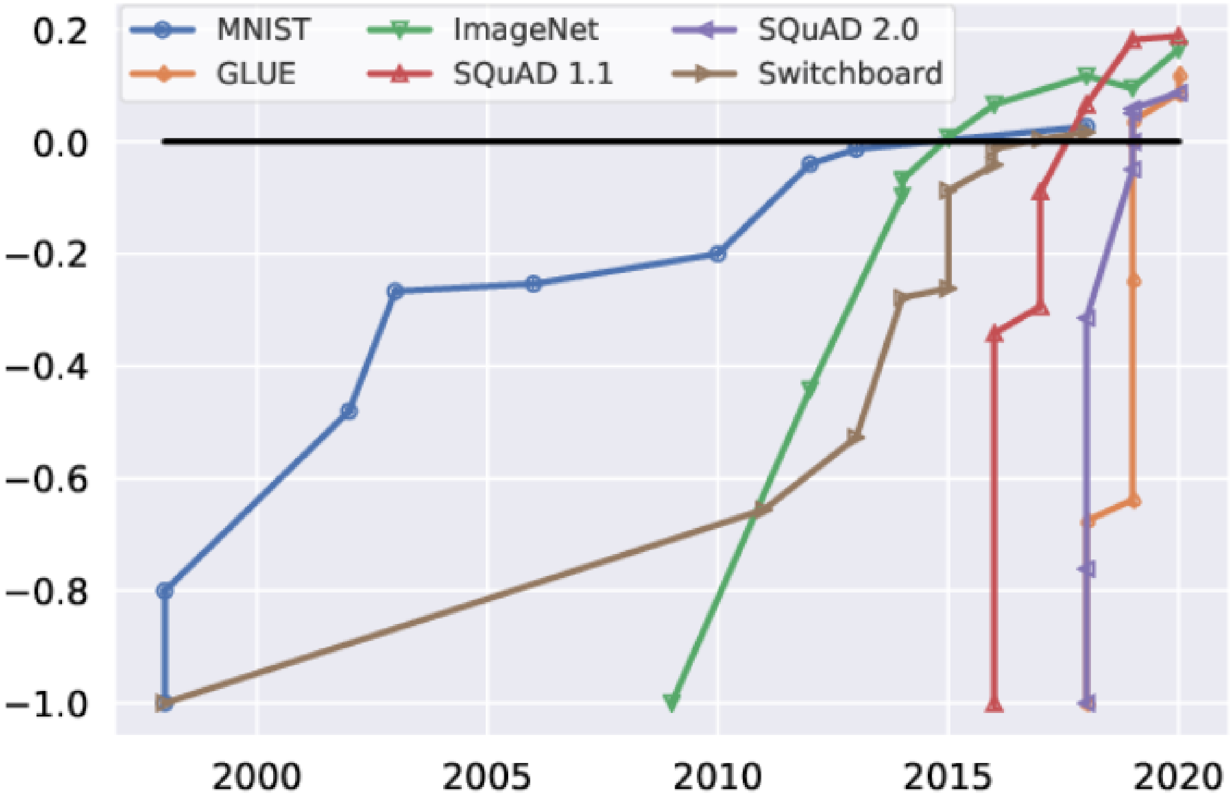
We introduce a new language representation model called **BERT**, which stands for **B**idirectional **E**ncoder **R**epresentations from **T**ransformers. Unlike recent language representation models (Peters et al., 2018a; Radford et al., 2018), BERT is designed to pretrain deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. As a result, the pre-trained BERT model can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, such as question answering and language inference, without substantial task-specific architecture modifications.

BERT is conceptually simple and empirically powerful. It obtains new state-of-the-art results on eleven natural language processing tasks, including pushing the GLUE score to 80.5% (7.7% point absolute improvement),

GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding, Wang et al.2018

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, Devlin et al.2018

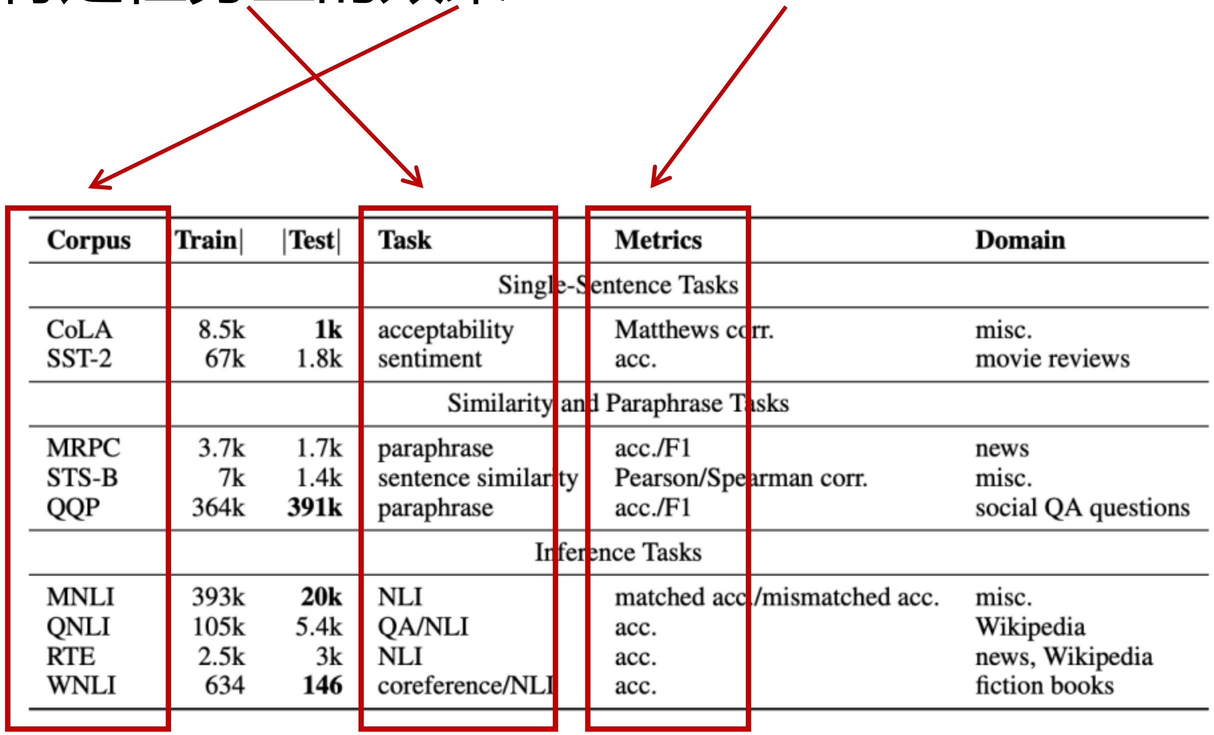
几个流行评估基准 性能饱和情况



如何构建一个评估基准?

评估基准（Evaluation Benchmark)

- 是一种用于评价和比较不同系统、算法或方法性能的工具或标准。它通常包括一系列**预定义的任务**、**数据集**、**评估指标**和**评估协议**，用于系统地测量和比较不同方法在特定任务上的效果



Corpus	|Train|	|Test|	Task	Metrics	Domain
			Single-Sentence Tasks		
CoLA	8.5k	**1k**	acceptability	Matthews corr.	misc.
SST-2	67k	1.8k	sentiment	acc.	movie reviews
			Similarity and Paraphrase Tasks		
MRPC	3.7k	1.7k	paraphrase	acc./F1	news
STS-B	7k	1.4k	sentence similarity	Pearson/Spearman corr.	misc.
QQP	364k	**391k**	paraphrase	acc./F1	social QA questions
			Inference Tasks		
MNLI	393k	**20k**	NLI	matched acc./mismatched acc.	misc.
QNLI	105k	5.4k	QA/NLI	acc.	Wikipedia
RTE	2.5k	3k	NLI	acc.	news, Wikipedia
WNLI	634	**146**	coreference/NLI	acc.	fiction books

评估基准构建挑战

□ 如何知道我们应该预定义哪些任务?

□ 如何以可扩展的方式构建合适的评估数据集?

□ **如何构建可靠的、自动化的评估方式?**

□ 如何构建一种公平、透明、可靠的评估协议?

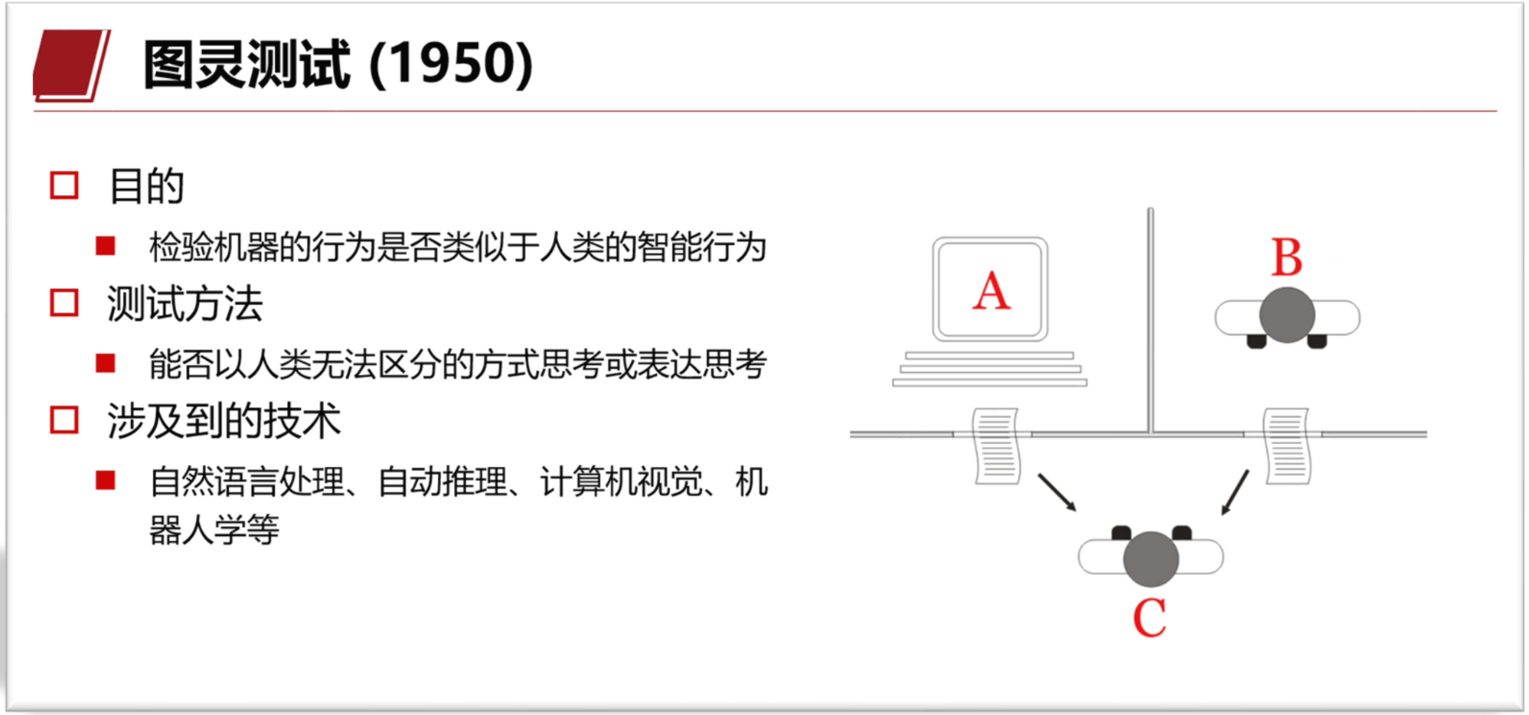
评估方法的分类

□ 参与方式

- 单模型评估
- 成对模型对比评估
- 多模型排序评估

评估方法的分类

- □ 参与方式
  - ■ 单模型评估
  - ■ 成对模型对比评估
  - ■ 多模型排序评估



图灵测试 (1950)

- □ 目的
  - ■ 检验机器的行为是否类似于人类的智能行为
- □ 测试方法
  - ■ 能否以人类无法区分的方式思考或表达思考
- □ 涉及到的技术
  - ■ 自然语言处理、自动推理、计算机视觉、机器人学等

评估方法的分类

- 直接程度
  - 内部评估（Intrinsic）：直接评估系统的某个内部特征或输出，而不是最终的任务性能
  - 外部评估（Extrinsic)：这种评估方法直接测量系统完成最终目标任务的能力，而不仅仅是其内部组件的性能

评估方法的分类

- 直接程度
  - 内部评估：直接评估系统的某个内部特征或输出，而不是最终的任务性能
  - 外部评估：这种评估方法直接测量系统完成最终目标任务的能力，而不仅仅是其内部组件的性能

通过困惑度（PPL）评估语言模型？通过数学推理任务来评估语言模型？

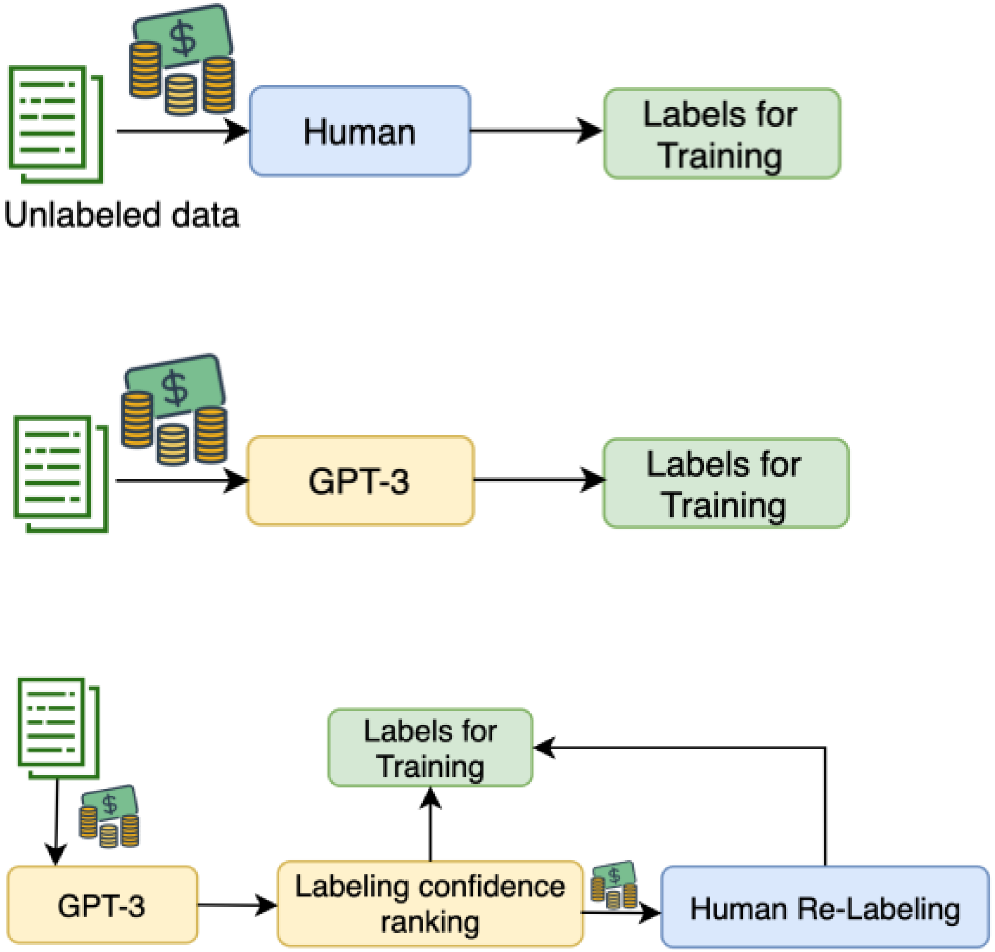
评估方法的分类

- □ 自动化程度
  - ■ 人工评估
  - ■ 自动评估

特性	自动化评估	人工评估
优点	- 高效，能快速处理大量数据 - 评估结果具有一致性，减少人为误差 - 易于扩展，适用于大规模数据集和复杂模型 - 成本效益高，尤其是长期来看	- 能深入理解复杂的业务需求和数据的细微差别 - 评估过程和标准具有高度灵活性 - 能识别模型的偏差和不公平性问题
缺点	- 灵活性有限，可能无法完全满足复杂的业务需求 - 可能只关注特定的性能指标，忽略模型的其他潜在问题	- 效率低，难以处理大量数据或复杂模型 - 成本高，尤其是需要专业知识的评估 - 可能存在一致性差和个人偏见的问题

评估方法的分类

- 自动化程度
  - 人工评估
  - 自动评估



Want To Reduce Labeling Cost? GPT-3 Can Help (EMNLP 2021)

Is ChatGPT better than Human Annotators? Huang et al.2023

评估的形式化描述

- x：测试实例（sample）
- $\tilde{y}$：系统预测 (system prediction)
- y：参考（reference)
  - 标签（Label)
  - 序列（Sequence)

场景分析：面向“标签”的评估

- **设计一个评估垃圾邮件过滤模型性能的指标**
  - x : 邮件
  - $\tilde{y}$: 系统预测标签
  - y : 正确标签
- 评估指标
  - **?**

场景分析： 面向“标签”的评估

- □ **设计一个评估垃圾邮件过滤模型性能的指标**
  - x : 邮件
  - $\tilde{y}$: 系统预测标签
  - y : 正确标签
- □ 评估指标
  - 准确率(Accuracy)：衡量的是分类器**正确分类**的邮件数**占总邮件数**的比例

场景分析

□ **设计一个评估垃圾邮件过滤模型的指标**

- x : 邮件
- $\tilde{y}$: 系统预测标签
- y : 正确标签

□ 评估指标

- **准确率(Accuracy)**：衡量的是分类器**正确分类**的邮件数**占总邮件数**的比例

场景分析

- □ **设计一个评估垃圾邮件过滤模型的指标**
  - ■ x : 邮件
  - ■ $\tilde{y}$: 系统预测标签
  - ■ y : 正确标签
- □ 评估指标
  - ■ **准确率(Accuracy)**：衡量的是分类器**正确分类**的邮件数**占总邮件数**的比例

如果有100个邮件，里面有10封垃圾邮件，即使模型全部预测为正常邮件，准确率也是90%

场景分析

- □ **设计一个评估垃圾邮件过滤模型的指标**
  - x：邮件
  - $\tilde{y}$：系统预测标签
  - y：正确标签
- □ 评估指标
  - **准确率(Accuracy)**：衡量的是分类器**正确分类**的邮件数**占总邮件数**的比例

垃圾邮件的错分有两种情况: “漏网” 与 “误伤”，无法反映出来

场景分析

- **设计一个评估垃圾邮件过滤模型的指标**
  - x：邮件
  - $\tilde{y}$：系统预测标签
  - y：正确标签

		Predicted	
		0	1
Actual	0	TN	FP
	1	FN	TP

- 评估指标
  - **准确率(Accuracy)**：衡量的是分类器正确分类的邮件数占总邮件数的比例
  - **精准率(Precision):** 被标记为垃圾邮件的邮件中，实际为垃圾邮件的比例
    - TP/(TP+FP) => 0/0 => 直接定义为0

场景分析

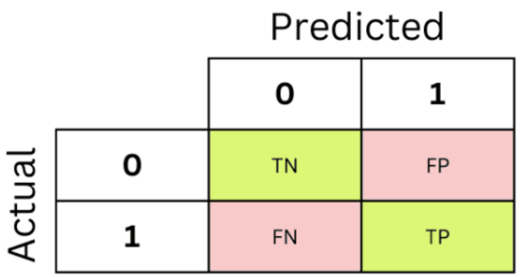
- □ **设计一个评估垃圾邮件过滤模型的指标**
  - ■ x：邮件
  - ■ $\tilde{y}$：系统预测标签
  - ■ y：正确标签

		Predicted	
		0	1
Actual	0	TN	FP
	1	FN	TP

- □ 评估指标
  - ■ **准确率(Accuracy)**：衡量的是分类器正确分类的邮件数占总邮件数的比例
  - ■ **精准率(Precision):** 被标记为垃圾邮件的邮件中，实际为垃圾邮件的比例
  - ■ **召回率(Recall):** 所有实际垃圾邮件中，被正确识别出来的比例
    - □ TP/(TP+FN) = 0/10

场景分析

- □ **设计一个评估垃圾邮件过滤模型的指标**
  - ■ x : 邮件
  - ■ $\tilde{y}$: 系统预测标签
  - ■ y : 正确标签



		Predicted	
		0	1
Actual	0	TN	FP
	1	FN	TP

- □ 评估指标
  - ■ **准确率(Accuracy)**：衡量的是分类器正确分类的邮件数占总邮件数的比例
  - ■ **精准率(Precision)**: 被标记为垃圾邮件的邮件中，实际为垃圾邮件的比例
  - ■ **召回率(Recall):** 所有实际垃圾邮件中，被正确识别出来的比例
  - ■ **F1分数(F1 Score)**：精确率和召回率的调和平均值 (2 P*R/(P+R))

场景分析： 面向“序列”的评估

□ 设计文本摘要评估系统

x : 输入长文本

随着春风渐暖，全国各地的景色也迎来了年度最为绚烂的时刻。从北至南，无数花朵竞相绽放，为大地披上了一层缤纷的色彩。在东部的山城，樱花树下，人们纷纷驻足，用镜头记录下这短暂而美丽的瞬间。西部的草原上，野花如海，吸引了众多摄影爱好者和旅行者前来探访。

城市公园内，各种春花争奇斗艳，市民在花海中漫步，享受春天的温暖和生机。儿童在花丛中嬉戏，家长们则拿出手机，记录下这些美好的时光。此外，一些城市还举办了花展和植物节，吸引了大批市民和游客前来观赏和体验。在乡村，桃花、李花、梨花等竞相开放，整个乡村像是被一层粉嫩的轻纱覆盖，吸引了许多城市居民前来赏花和体验乡村生活。农民们在花间忙碌，为即将到来的收获季节做准备。

这个春季，无论是城市还是乡村，大自然都以其无与伦比的美丽，为人们带来了欢乐和希望。人们纷纷走出家门，融入这片绚烂之中，感受春天的魅力。

$\tilde{y}$: 系统生成摘要

全国各地春色绚烂，樱花、野花竞开，城市到乡村处处花海。人们纷纷出门享受春光，赏花摄影，感受季节之美。

y : 人写的摘要

春季将全国装点得色彩斑斓，从城市到乡村，樱花与野花竞相绽放。民众踏出家门，沉浸在花海之中，欣赏并捕捉这季节的美好。

场景分析： 面向“序列”的评估

□ 设计文本摘要评估系统

$$u(x,y,\tilde{y})=?$$

x : 输入长文本

随着春风渐暖，全国各地的景色也迎来了年度最为绚烂的时刻。从北至南，无数花朵竞相绽放，为大地披上了一层缤纷的色彩。在东部的山城，樱花树下，人们纷纷驻足，用镜头记录下这短暂而美丽的瞬间。西部的草原上，野花如海，吸引了众多摄影爱好者和旅行者前来探访。

城市公园内，各种春花争奇斗艳，市民在花海中漫步，享受春天的温暖和生机。儿童在花丛中嬉戏，家长们则拿出手机，记录下这些美好的时光。此外，一些城市还举办了花展和植物节，吸引了大批市民和游客前来观赏和体验。在乡村，桃花、李花、梨花等竞相开放，整个乡村像是被一层粉嫩的轻纱覆盖，吸引了许多城市居民前来赏花和体验乡村生活。农民们在花间忙碌，为即将到来的收获季节做准备。

这个春季，无论是城市还是乡村，大自然都以其无与伦比的美丽，为人们带来了欢乐和希望。人们纷纷走出家门，融入这片绚烂之中，感受春天的魅力。

$\tilde{y}$: 系统生成摘要

全国各地春色绚烂，樱花、野花竞开，城市到乡村处处花海。人们纷纷出门享受春光，赏花摄影，感受季节之美。

y : 人写的摘要

春季将全国装点得色彩斑斓，从城市到乡村，樱花与野花竞相绽放。民众踏出家门，沉浸在花海之中，欣赏并捕捉这季节的美好。

Exact Match (EM)

- **特点：**
  - **优势：**High precision: if metric is 1 => we have a good sequence
  - **不足：**Low recall: if the metric is not 1, then we still may have a good hypothesis
- **应用**
  - 自动问答

$$u(x, y, \tilde{y}) = \mathrm{I}(y == \tilde{y})$$

Word error rate (WER)

- **特点：**
  - **优势：** Relaxes exact match
  - **不足：** Semantically similar words ignored
- **应用：**
  - 语音识别
  - 机器翻译

$$u(x, y, \tilde{y}) = \frac{\delta(y, \tilde{y})}{|y|}$$

$\delta(y, \tilde{y})$ word edit distance between y and $\tilde{y}$

$|y|$ is the length of y

Perplexity (PPL)

- **特点：**
  - **优势：** Relaxes exact match
  - **不足：** Semantically similar words ignored
- **应用：**
  - 语言模型

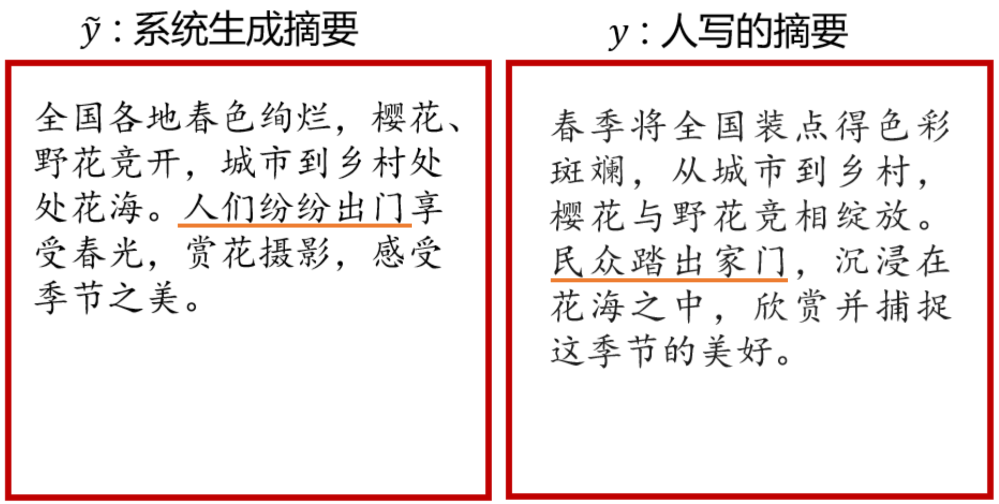
$$u(x, y, \tilde{y}) = \exp\left(-\frac{1}{y}\sum_{i=1}^{|y|}\log_{P_\theta}(y_i|y_{1:i-1})\right)$$

BLEU/ROUGE

- **特点：**
  - 计算机器生成的文本和参考文本之间的n-gram（连续的n个词）重合度来工作
  - **优势：**Relaxes exact match
  - **不足：**Semantically similar words ignored
- **应用：**
  - 机器翻译、文本摘要

BLEU/ROUGE

- **特点：**
  - 计算机器生成的文本和参考文本之间的n-gram（连续的n个词）重合度来工作
  - **优势：**Relaxes exact match
  - **不足：**Semantically similar words ignored
- **应用：**
  - 机器翻译、文本摘要



$\tilde{y}$: 系统生成摘要

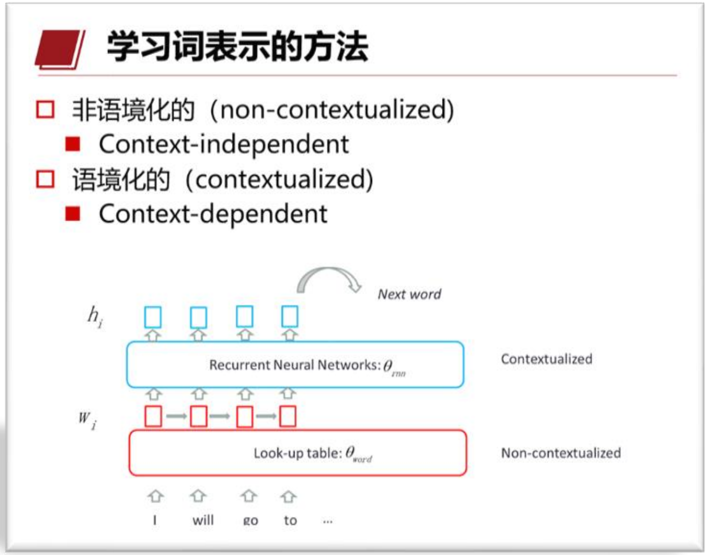
全国各地春色绚烂，樱花、野花竞开，城市到乡村处处花海。人们纷纷出门享受春光，赏花摄影，感受季节之美。

y : 人写的摘要

春季将全国装点得色彩斑斓，从城市到乡村，樱花与野花竞相绽放。民众踏出家门，沉浸在花海之中，欣赏并捕捉这季节的美好。

BERTScore

- □ **特点：**
  - 引入大模型进行语义泛化
  - **优势：**支持语义泛化
- □ **应用：**
  - 机器翻译、文本摘要



学习词表示的方法

- □ 非语境化的（non-contextualized)
  - Context-independent
- □ 语境化的（contextualized)
  - Context-dependent

Next word
h_i
Recurrent Neural Networks: θ_{rnn}
Contextualized
w_i
Look-up table: θ_{word}
Non-contextualized
I
will
go
to
...

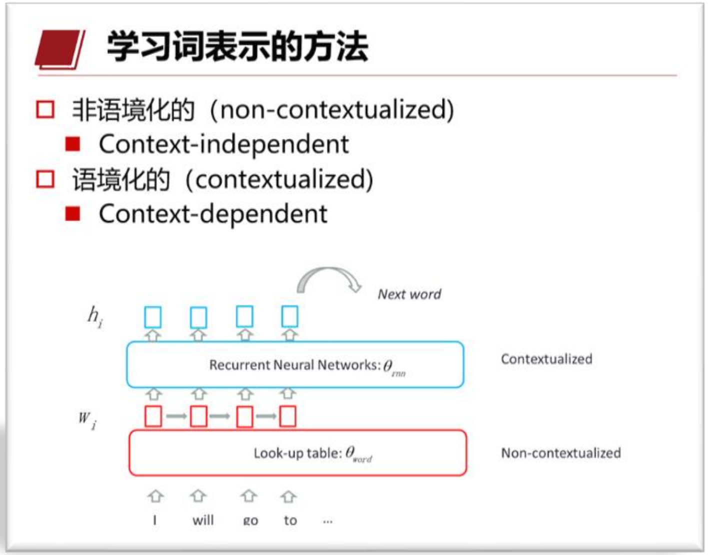
BERTScore

□ 特点：

- ■ 引入大模型进行语义泛化
- ■ **优势**：支持语义泛化

□ 应用：

- ■ 机器翻译、文本摘要



学习词表示的方法

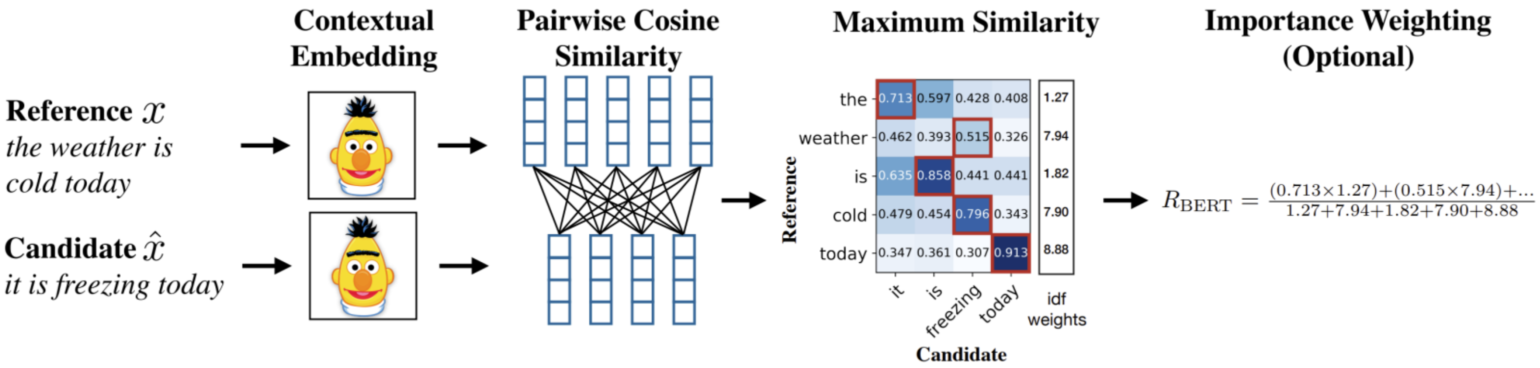
□ 非语境化的（non-contextualized）

■ Context-independent

□ 语境化的（contextualized）

■ Context-dependent

Next word
h_i
Recurrent Neural Networks: θ_{rnn}
Contextualized
w_i
Look-up table: θ_{word}
Non-contextualized
I
will
go
to
...



Reference \ Candidate	it	is	freezing	today	idf weights
the	0.713	0.597	0.428	0.408	1.27
weather	0.462	0.393	0.515	0.326	7.94
is	0.635	0.858	0.441	0.441	1.82
cold	0.479	0.454	0.796	0.343	7.90
today	0.347	0.361	0.307	0.913	8.88

$$R_{\mathrm{BERT}} = \frac{(0.713\times1.27)+(0.515\times7.94)+\ldots}{1.27+7.94+1.82+7.90+8.88}$$

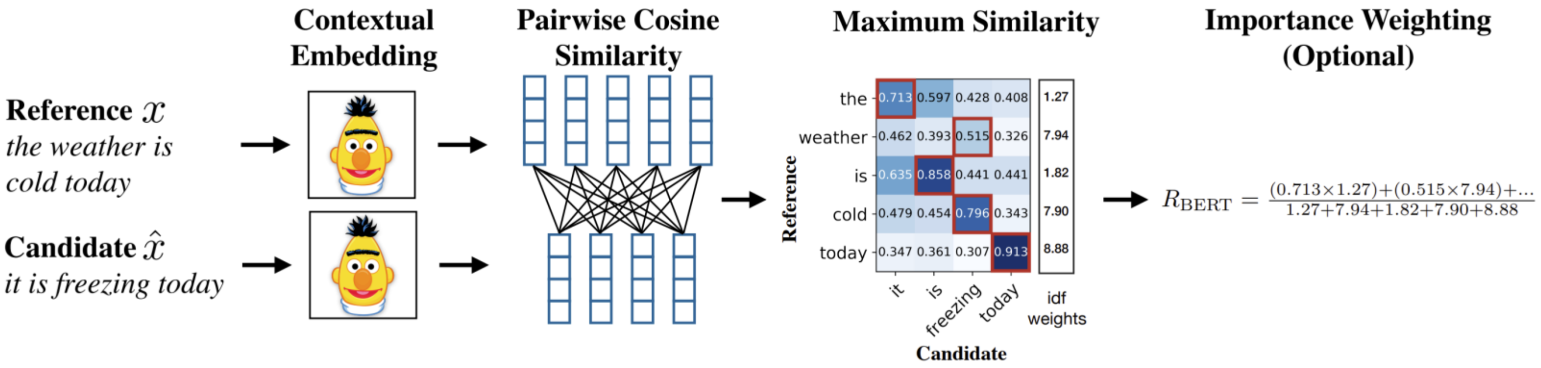
BERTScore

□ 特点：

- 引入大模型进行语义泛化
- **优势**：支持语义泛化

□ 应用：

- 机器翻译、文本摘要



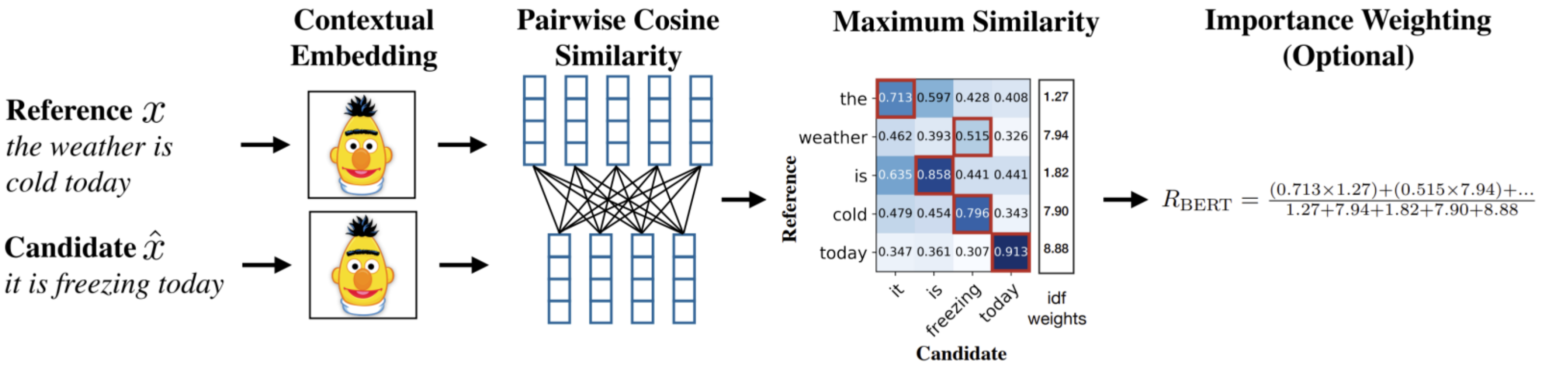
Reference \ Candidate	it	is	freezing	today	idf weights
the	0.713	0.597	0.428	0.408	1.27
weather	0.462	0.393	0.515	0.326	7.94
is	0.635	0.858	0.441	0.441	1.82
cold	0.479	0.454	0.796	0.343	7.90
today	0.347	0.361	0.307	0.913	8.88

$$R_{\mathrm{BERT}} = \frac{(0.713\times1.27)+(0.515\times7.94)+\ldots}{1.27+7.94+1.82+7.90+8.88}$$

有哪些不足？

BERTScore

- 局限性
  - 需要依赖于Reference



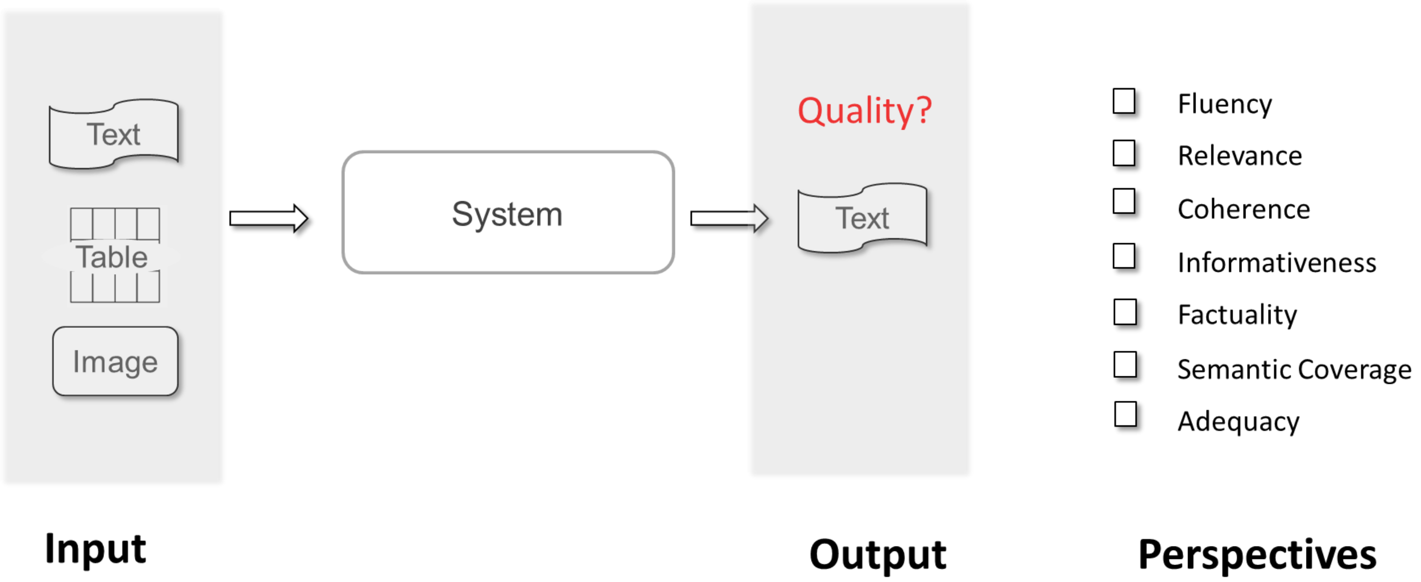
	it	is	freezing	today	idf weights
the	0.713	0.597	0.428	0.408	1.27
weather	0.462	0.393	0.515	0.326	7.94
is	0.635	0.858	0.441	0.441	1.82
cold	0.479	0.454	0.796	0.343	7.90
today	0.347	0.361	0.307	0.913	8.88

$$R_{\text{BERT}} = \frac{(0.713\times1.27)+(0.515\times7.94)+\ldots}{1.27+7.94+1.82+7.90+8.88}$$

BERTScore

□ 局限性

- 需要依赖于Reference
- 评估的维度单一



Text

Table

Image

System

Quality?

Text

□ Fluency
□ Relevance
□ Coherence
□ Informativeness
□ Factuality
□ Semantic Coverage
□ Adequacy

Input

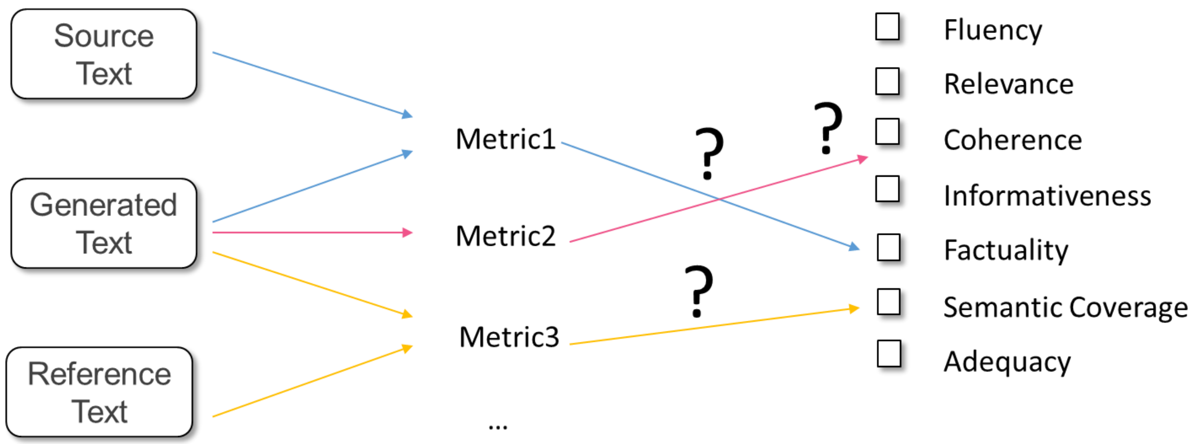
Output

Perspectives

BERTScore

□ 局限性

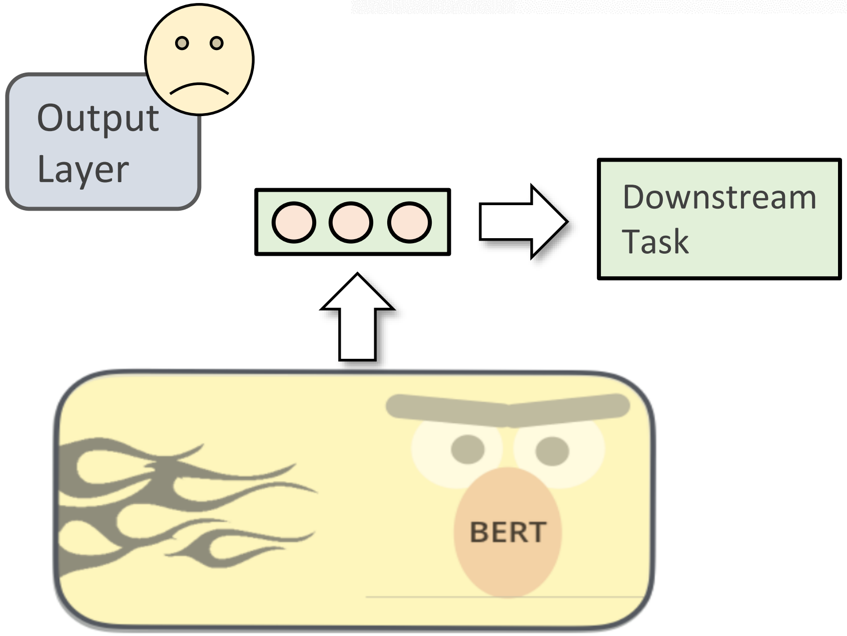
- 需要依赖于Reference
- 评估的维度单一
- 忽略了对“x” (source) 的使用



BERTScore

- 局限性
  - 需要依赖于Reference
  - 评估的维度单一
  - 忽略了对“x” (source) 的使用
  - 没有充分利用大模型

BERTScore
MoverScore
BLEURT
COMET
…



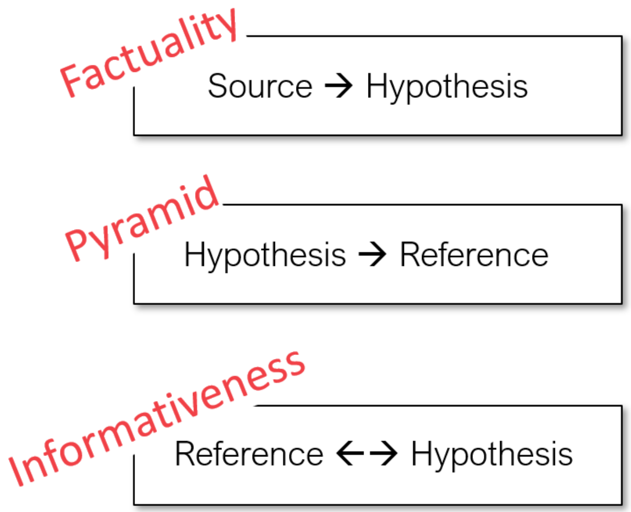
BARTScore: Evaluation as Generation

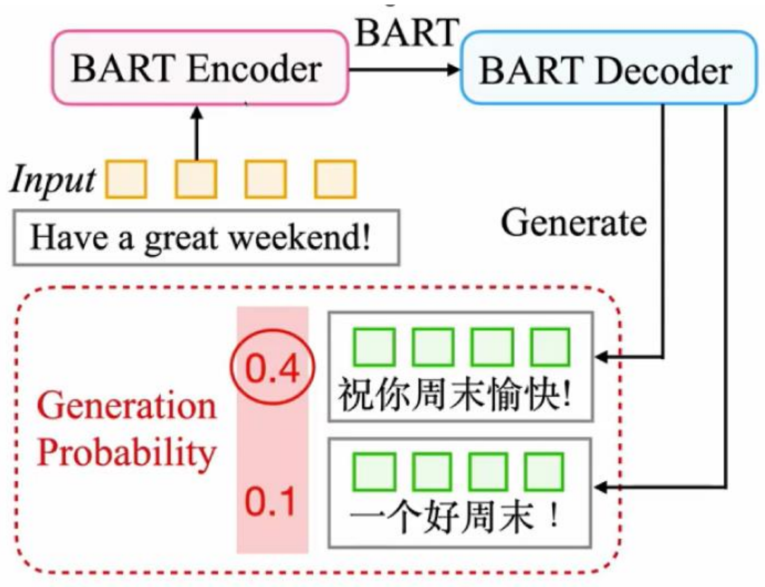
BARTScore: Evaluation as Generation

□ 定义

$$BARTScore = \sum_{t=1}^{m} w_t \log p(y_t|y_{<t}, x, \theta)$$

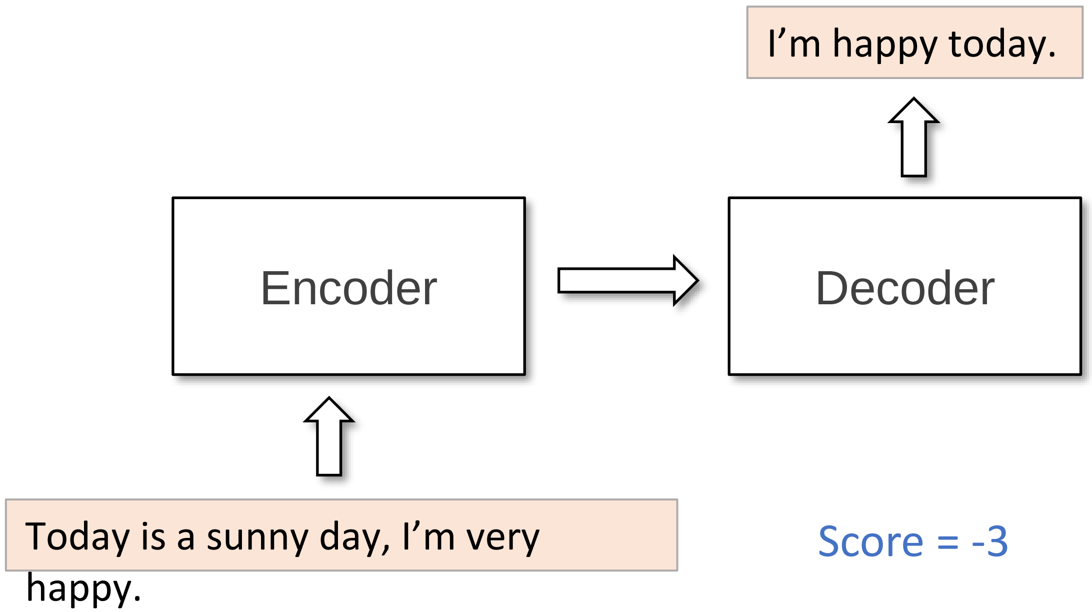
□ 支持多种使用





BARTScore: Evaluation as Generation

- Original



I’m happy today.

Encoder

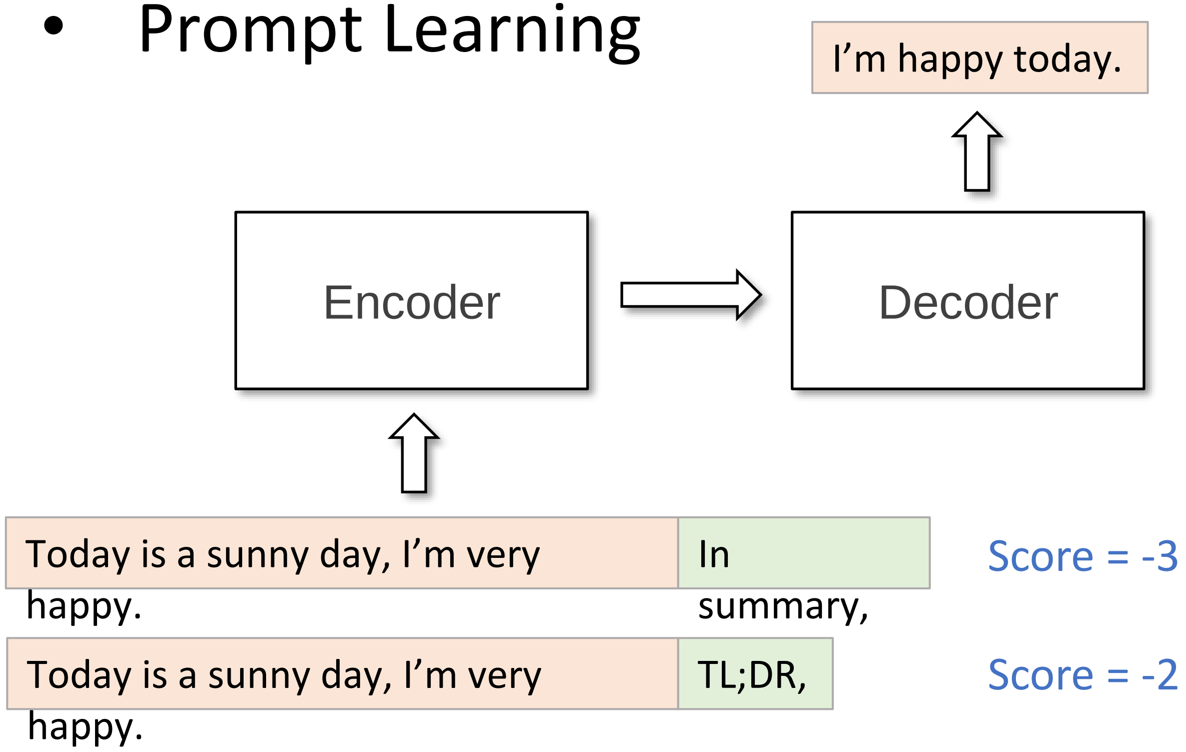
Decoder

Today is a sunny day, I’m very happy.

Score = -3

Final score = -3

- Prompt Learning



I’m happy today.

Encoder

Decoder

Today is a sunny day, I’m very happy. In summary,

Score = -3

Today is a sunny day, I’m very happy. TL;DR,

Score = -2

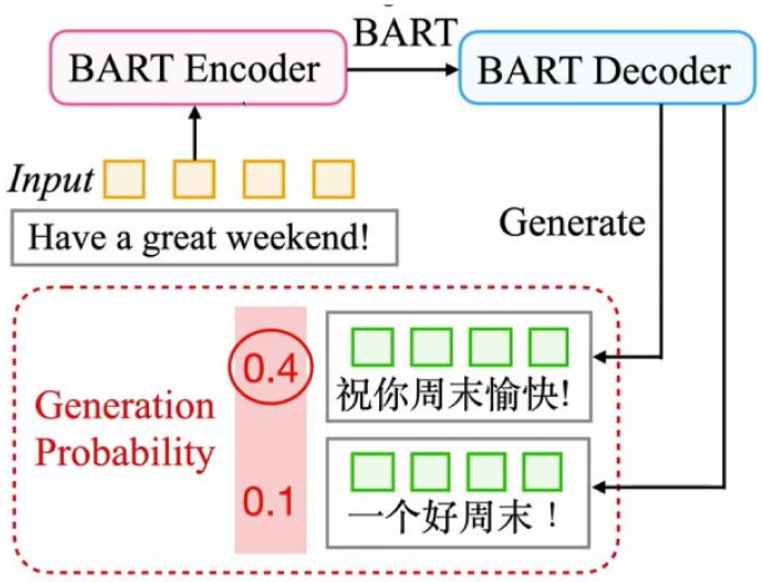
Final score = $\frac{(-3)+(-2)}{2} = -2.5$

BARTScore: Evaluation as Generation

- □ 定义

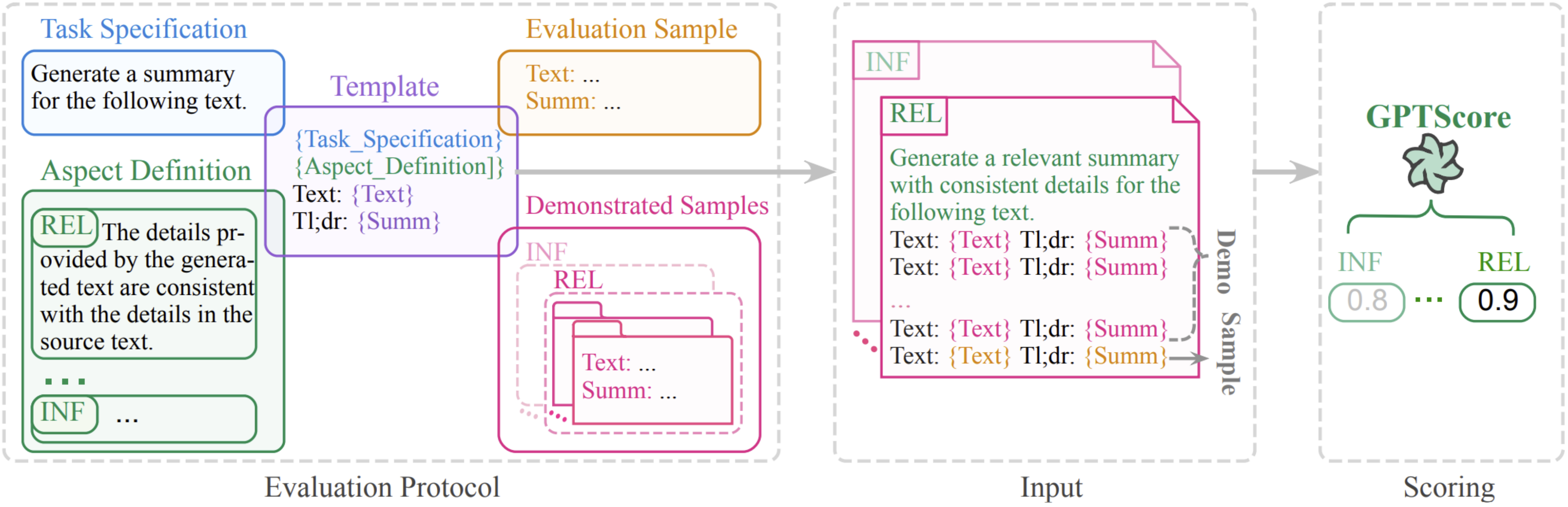
$$BARTScore = \sum_{t=1}^{m} w_t \log p(y_t | y_{<t}, x, \theta)$$

- □ 支持多种使用
- □ 不足之处?



GPTScore

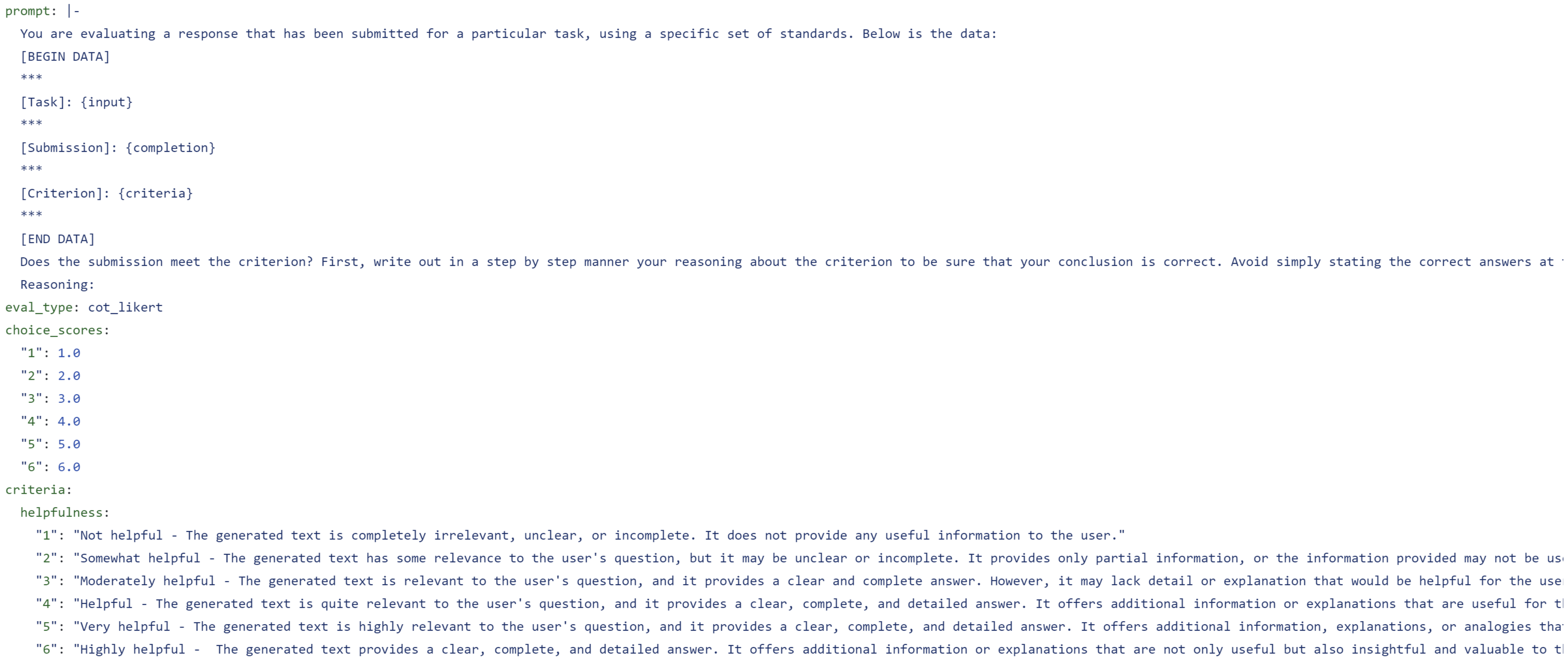
- Evaluation
  - How to evaluate a model as you desire?



Task Specification
Generate a summary for the following text.
Template
{Task_Specification}
{Aspect_Definition]}
Text: {Text}
Tl;dr: {Summ}
Evaluation Sample
Text: ...
Summ: ...
Aspect Definition
REL The details provided by the generated text are consistent with the details in the source text.
...
INF ...
Demonstrated Samples
INF
REL
Text: ...
Summ: ...
Evaluation Protocol
INF
REL
Generate a relevant summary with consistent details for the following text.
Text: {Text} Tl;dr: {Summ}
Text: {Text} Tl;dr: {Summ}
...
Text: {Text} Tl;dr: {Summ}
Text: {Text} Tl;dr: {Summ}
Demo
Sample
Input
GPTScore
INF
REL
0.8
...
0.9
Scoring

ChatGPT Score

- Evaluation
  - How to evaluate a model as you desire?



```
prompt: |-
  You are evaluating a response that has been submitted for a particular task, using a specific set of standards. Below is the data:
  [BEGIN DATA]
  ***
  [Task]: {input}
  ***
  [Submission]: {completion}
  ***
  [Criterion]: {criteria}
  ***
  [END DATA]
  Does the submission meet the criterion? First, write out in a step by step manner your reasoning about the criterion to be sure that your conclusion is correct. Avoid simply stating the correct answers at
  Reasoning:
eval_type: cot_likert
choice_scores:
  "1": 1.0
  "2": 2.0
  "3": 3.0
  "4": 4.0
  "5": 5.0
  "6": 6.0
criteria:
  helpfulness:
    "1": "Not helpful - The generated text is completely irrelevant, unclear, or incomplete. It does not provide any useful information to the user."
    "2": "Somewhat helpful - The generated text has some relevance to the user's question, but it may be unclear or incomplete. It provides only partial information, or the information provided may not be us
    "3": "Moderately helpful - The generated text is relevant to the user's question, and it provides a clear and complete answer. However, it may lack detail or explanation that would be helpful for the use
    "4": "Helpful - The generated text is quite relevant to the user's question, and it provides a clear, complete, and detailed answer. It offers additional information or explanations that are useful for t
    "5": "Very helpful - The generated text is highly relevant to the user's question, and it provides a clear, complete, and detailed answer. It offers additional information, explanations, or analogies tha
    "6": "Highly helpful -  The generated text provides a clear, complete, and detailed answer. It offers additional information or explanations that are not only useful but also insightful and valuable to t
```

评估方法的可靠性

评估文本“铁三角”

Text

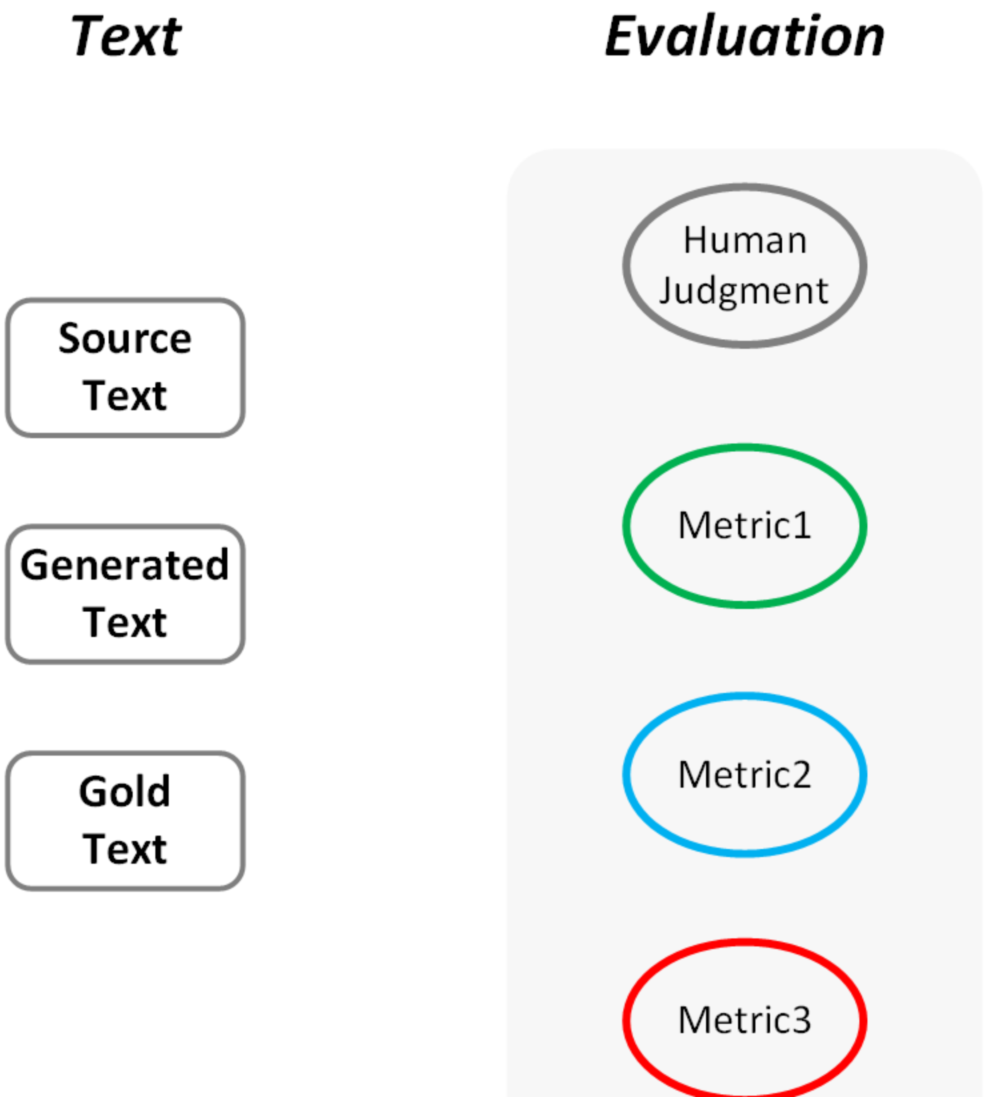
Source Text

Generated Text (Hypothesis, System output)

Gold Text (Reference)

Document
Summary

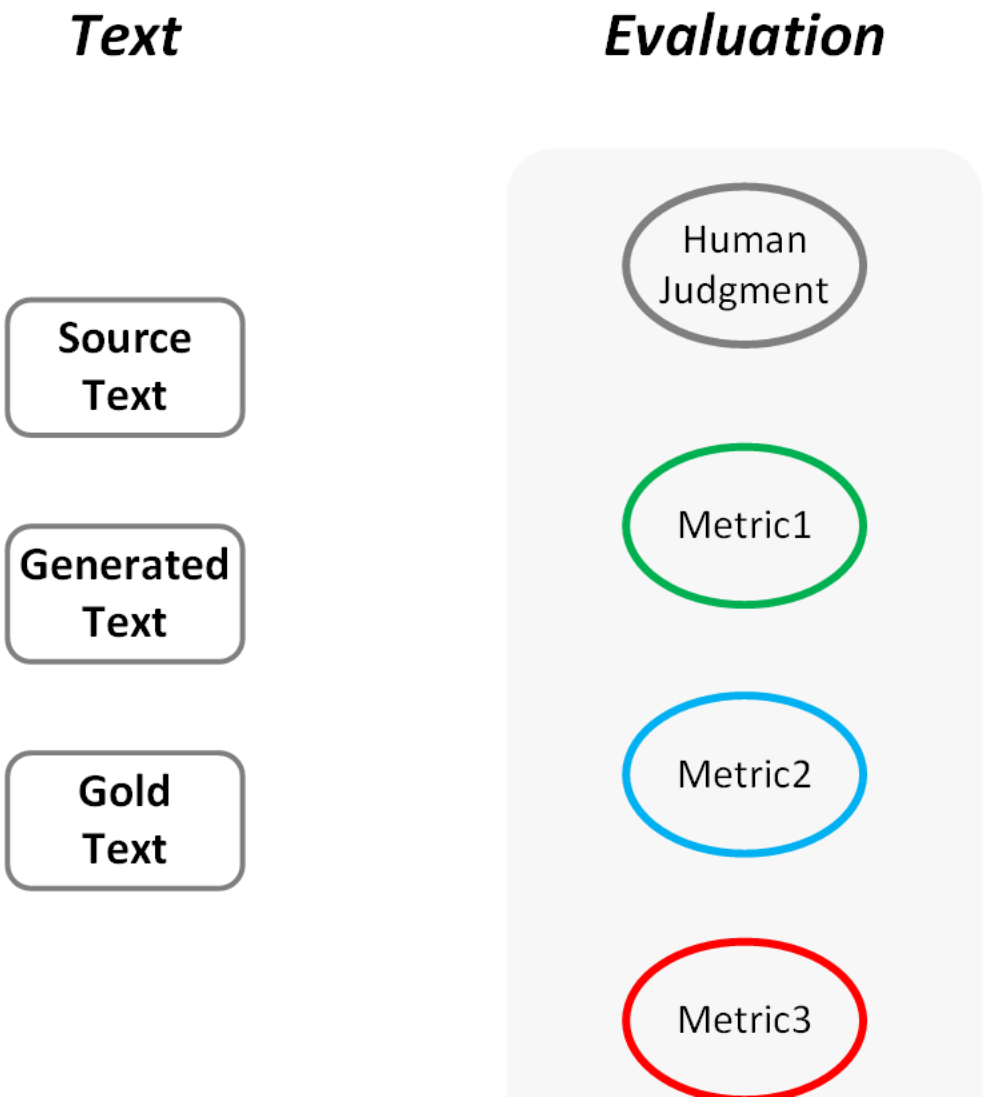
评估



Purpose of Evaluation:

- Assess the quality of *generated text* output from a given *system*

评估



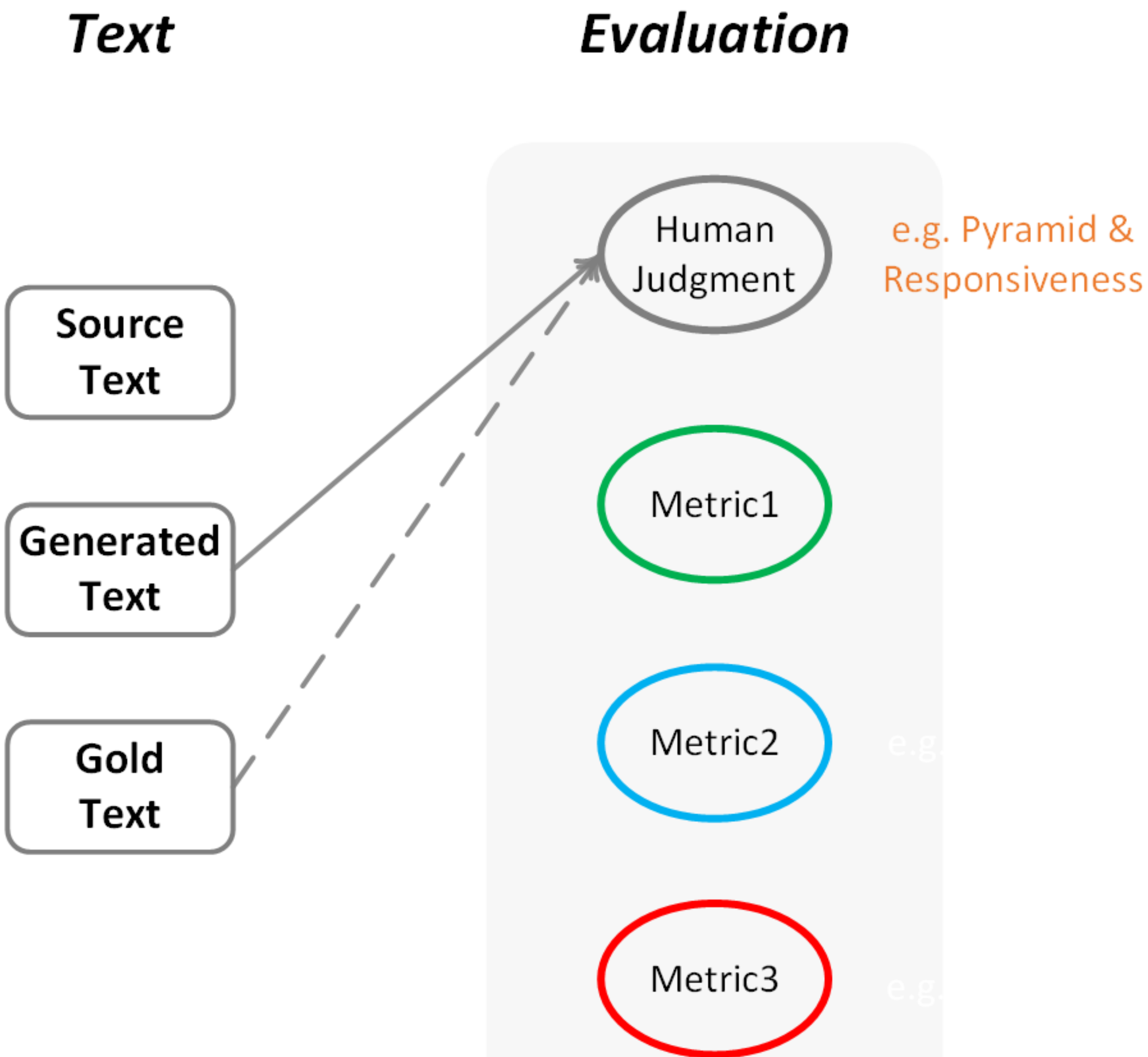
Purpose of Evaluation:

- Assess the quality of *generated text* output from a given *system*

Types of Evaluation:

- Human evaluation
- Automated evaluation

评估



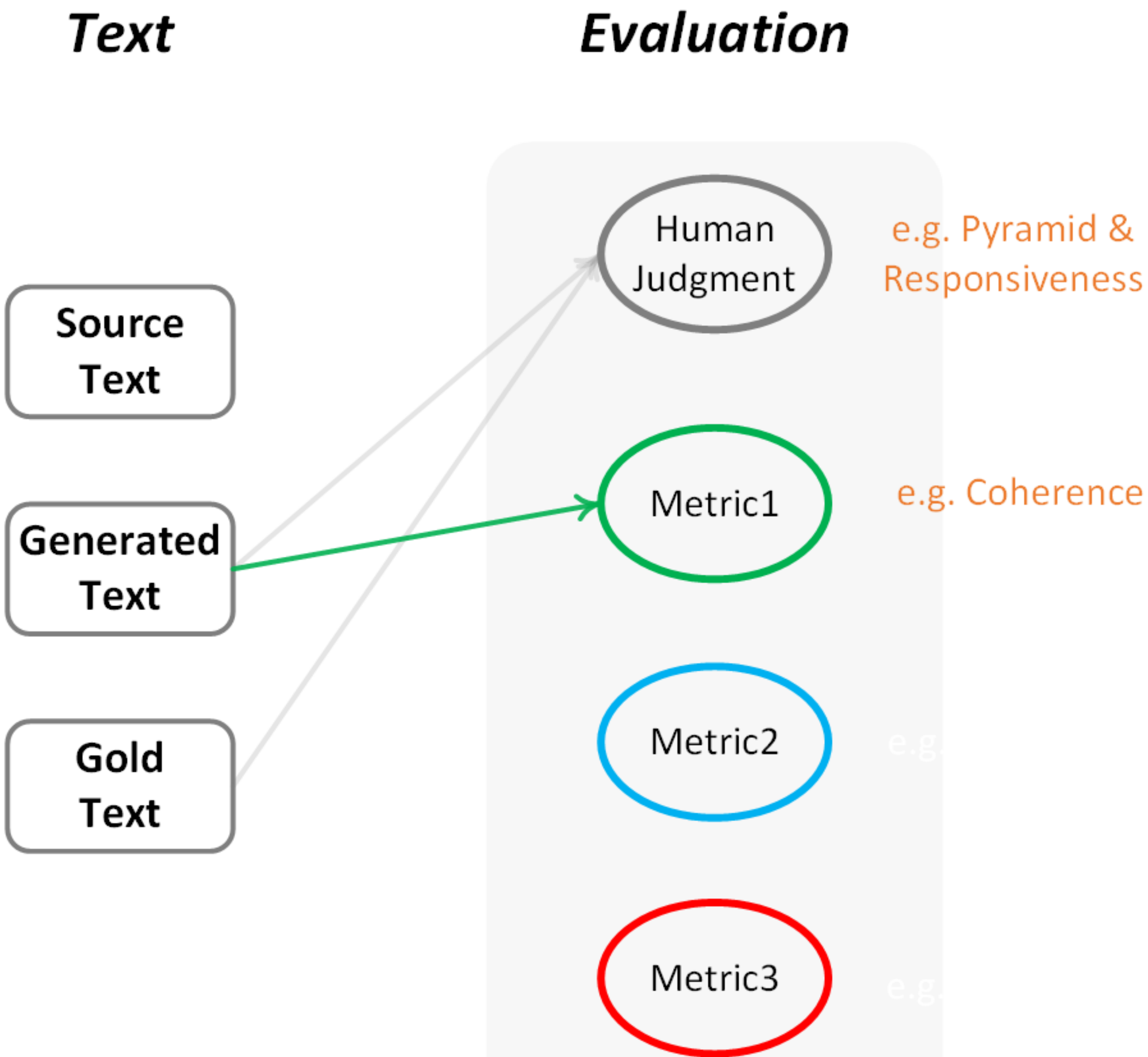
Purpose of Evaluation:

- Assess the quality of *generated text* output from a given *system*

Types of Evaluation:

- Human evaluation
- Automated evaluation

评估



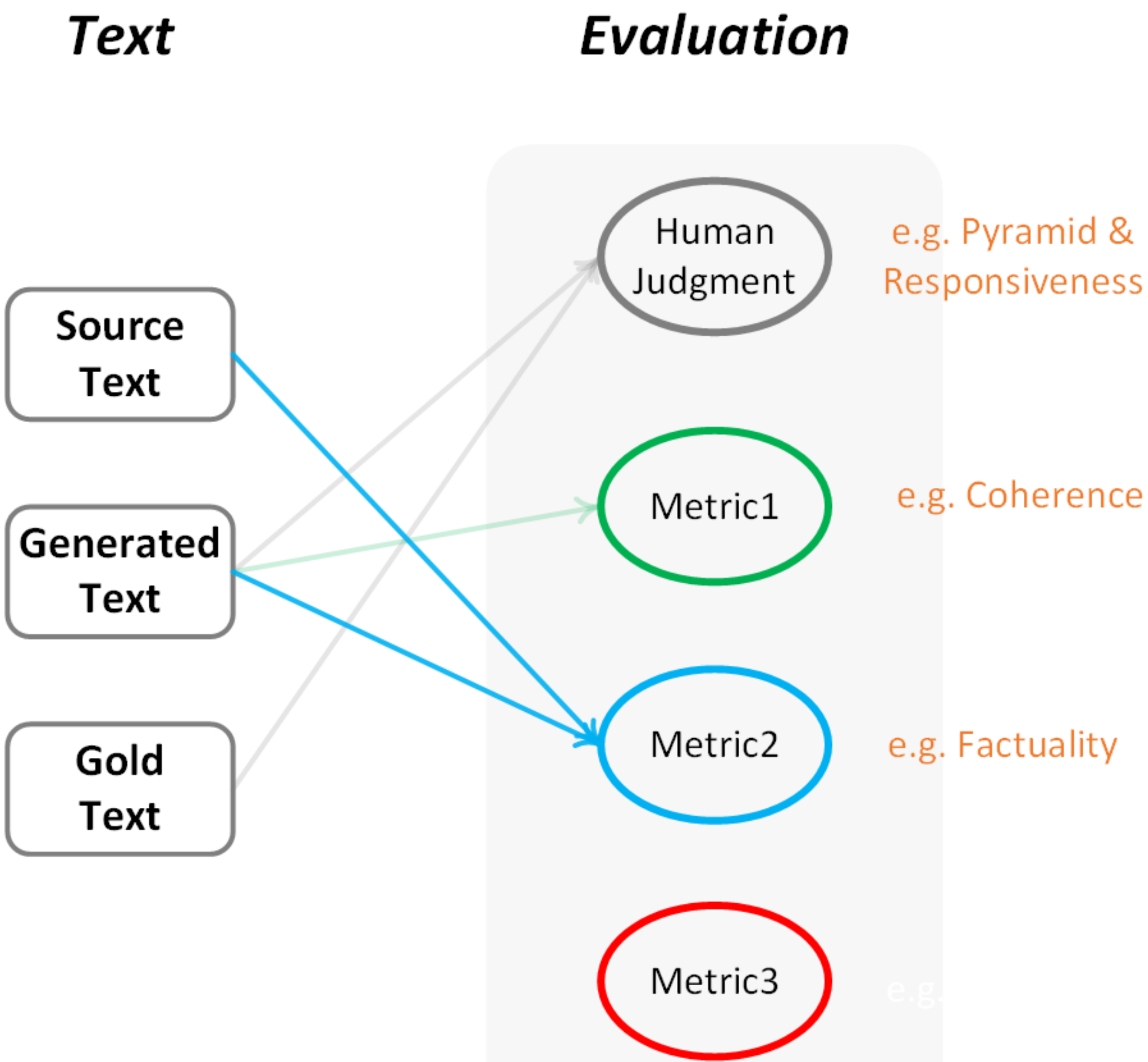
Purpose of Evaluation:

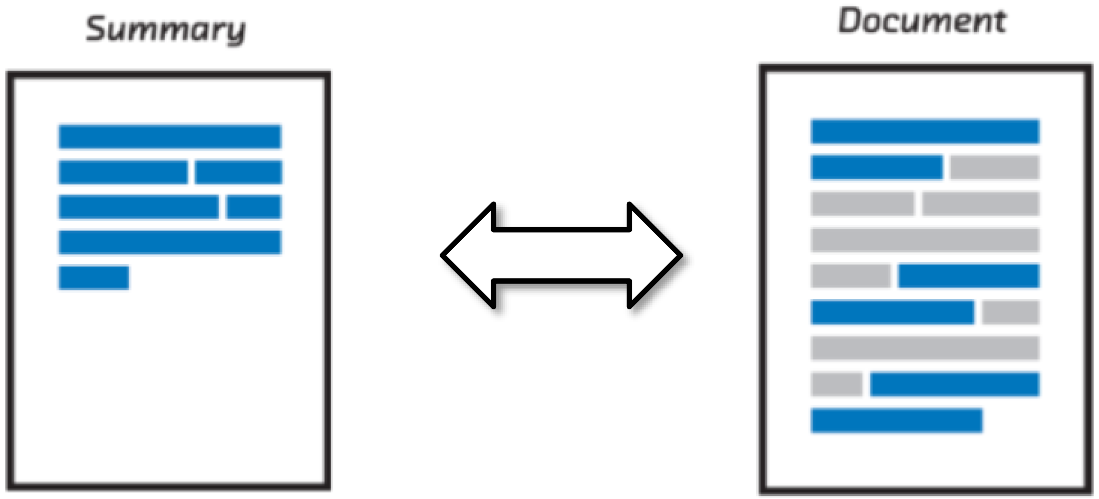
- Assess the quality of *generated text* output from a given *system*

Types of Evaluation:

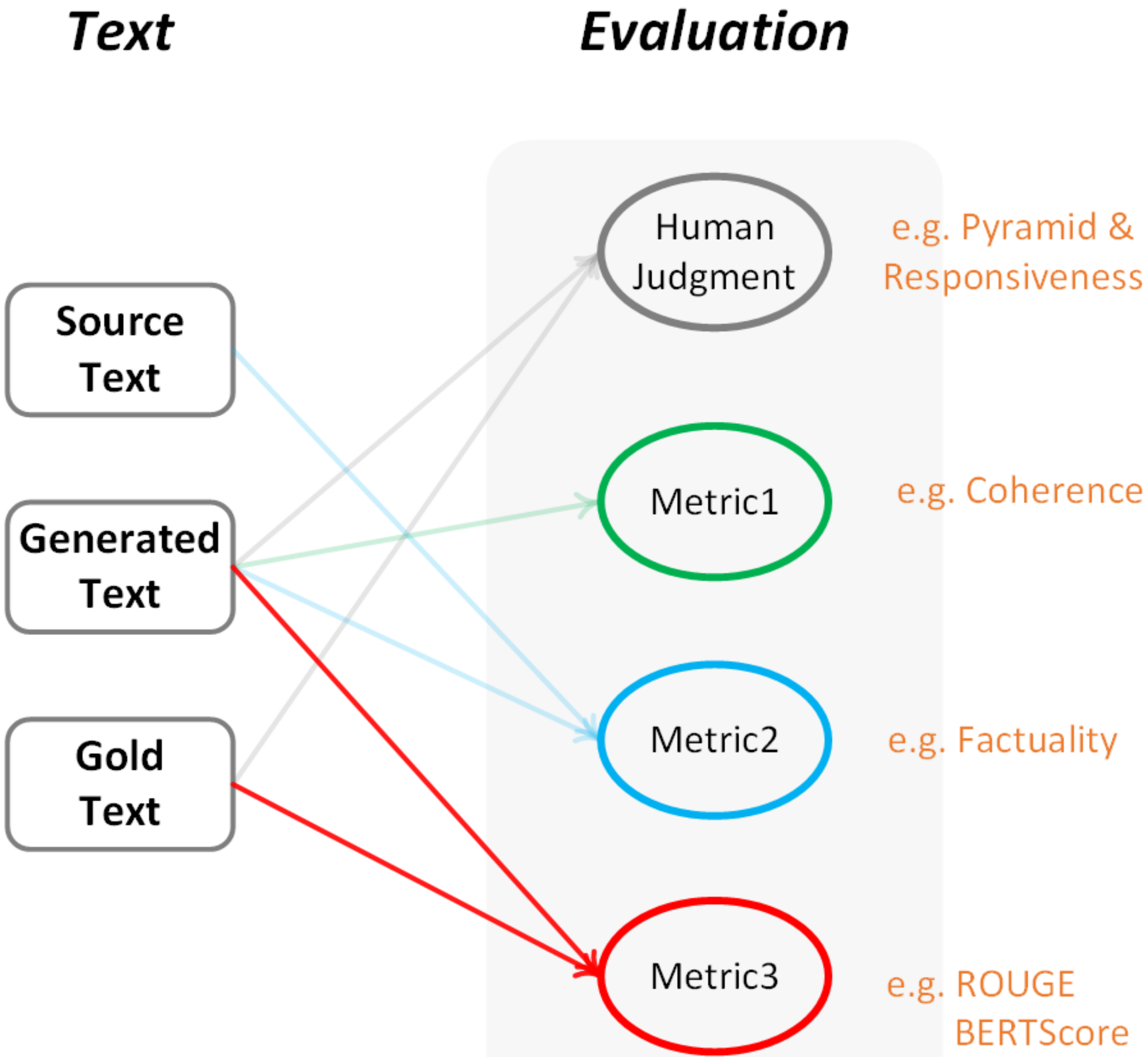
- Human evaluation
- Automated evaluation
  - Many metrics
  - Easy way to characterize them

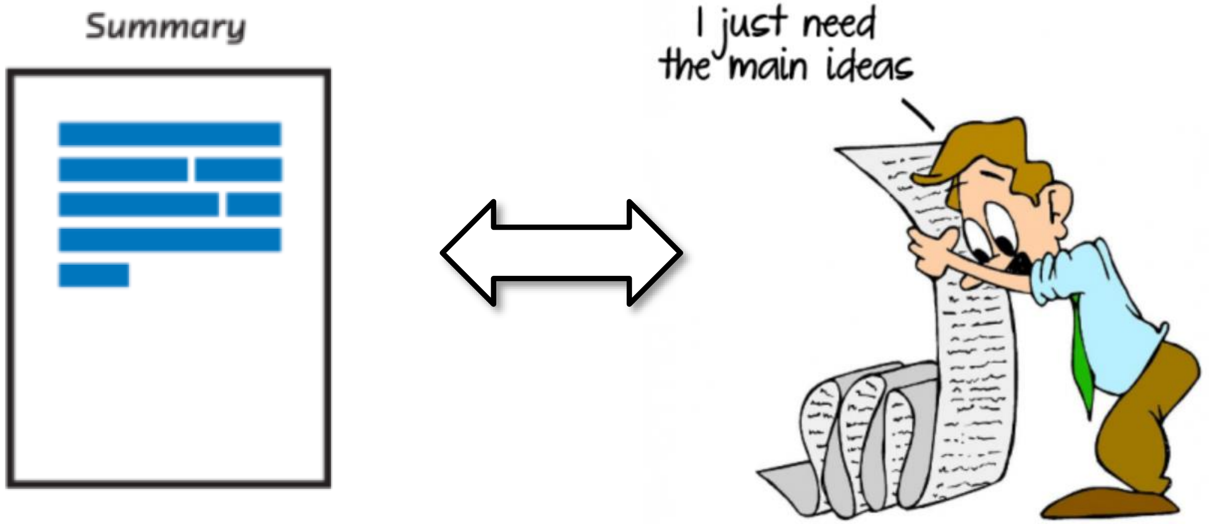
评估


Text
Evaluation
Source Text
Generated Text
Gold Text
Human Judgment
Metric1
Metric2
Metric3
e.g. Pyramid & Responsiveness
e.g. Coherence
e.g. Factuality

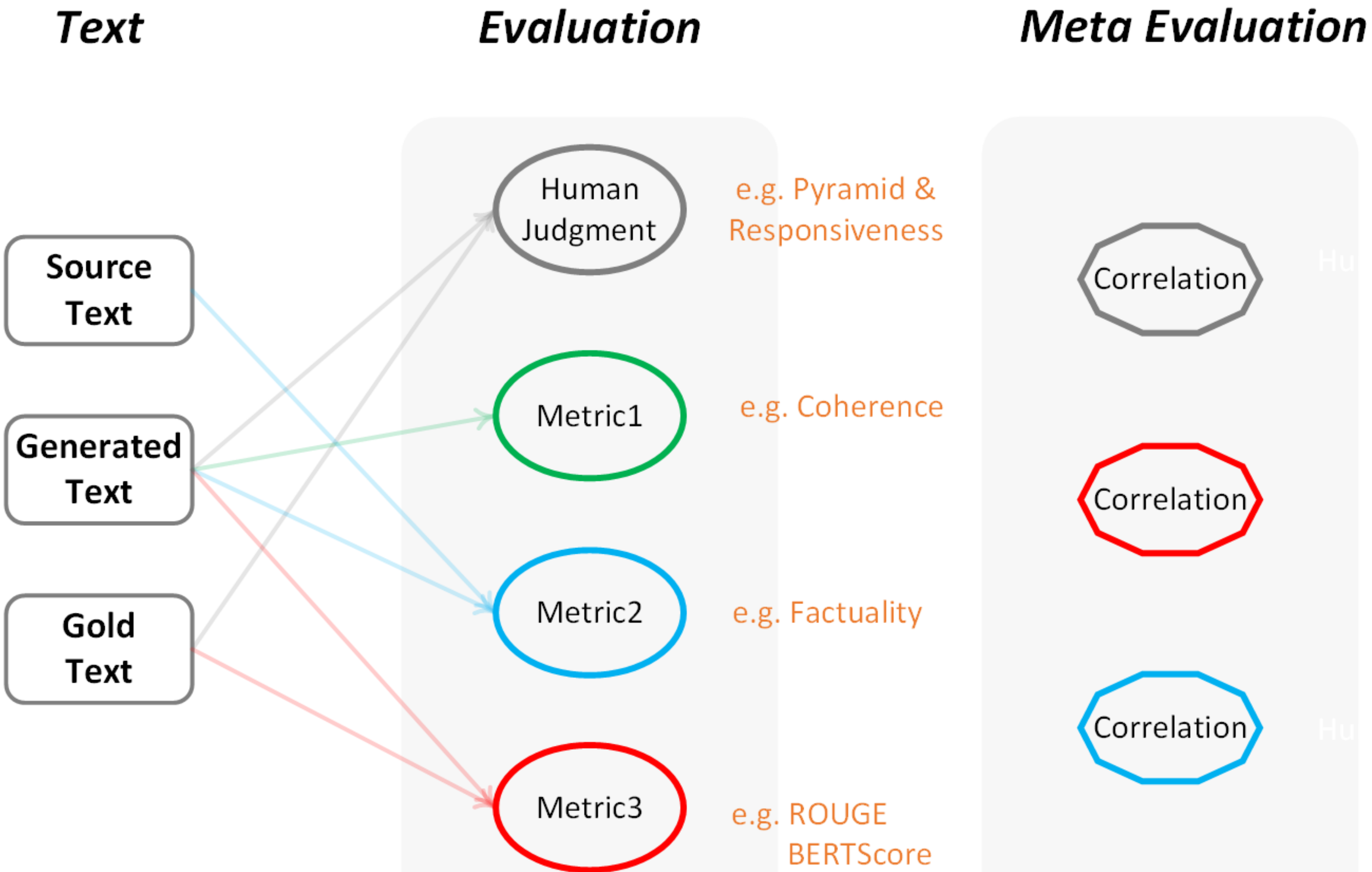

Summary
Document

评估


Text
Evaluation
Source Text
Generated Text
Gold Text
Human Judgment
Metric1
Metric2
Metric3
e.g. Pyramid & Responsiveness
e.g. Coherence
e.g. Factuality
e.g. ROUGE
BERTScore


Summary
I just need the main ideas

元评估



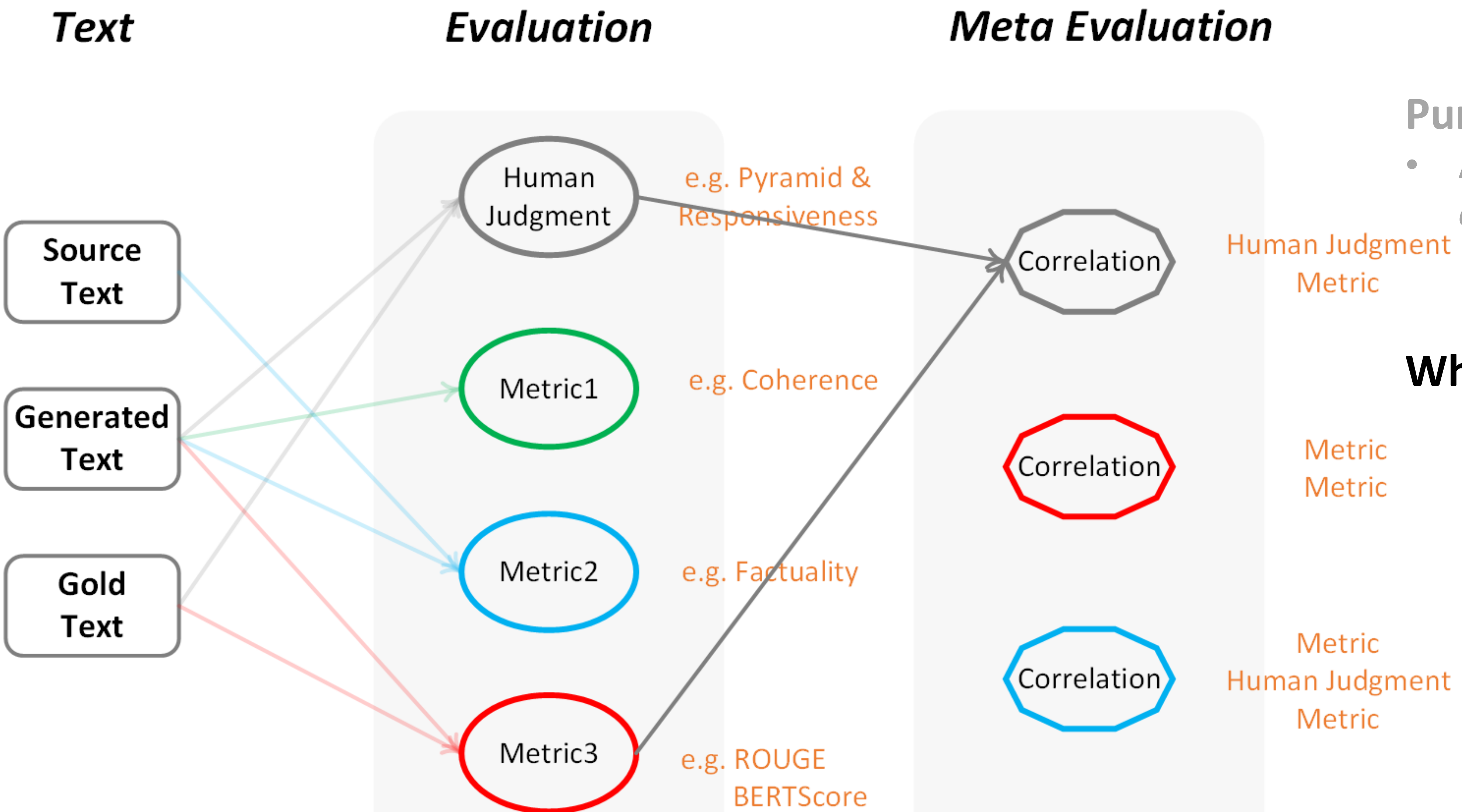
Purpose of Meta Eval:

- Assess the reliability of *automated metrics*

Purpose of Evaluation:

- Assess the quality of *systems*

元评估



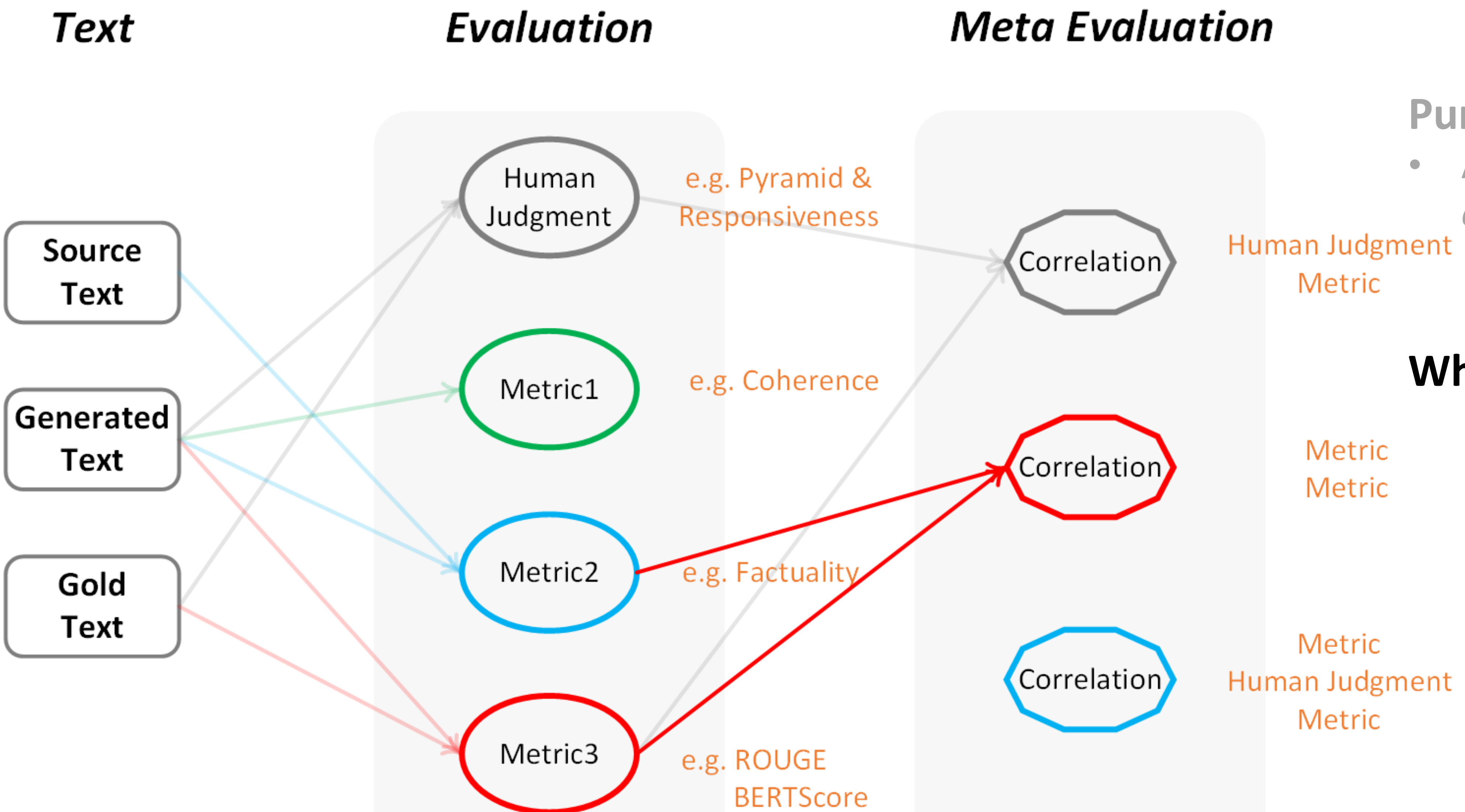
Purpose of Meta Eval:

- Assess the reliability of *automated metrics*

What's the reliability?

Reliability is defined as …

元评估

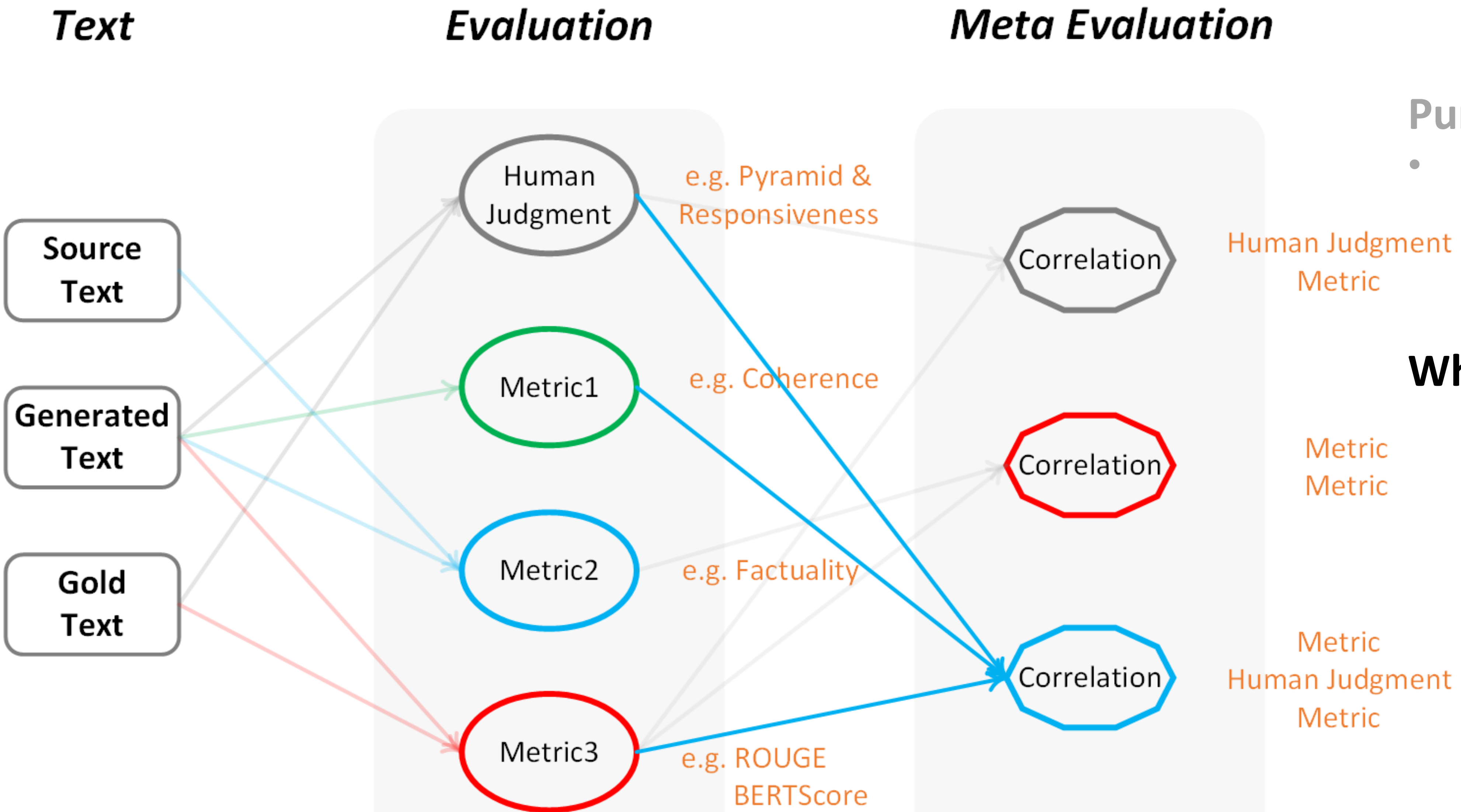


Purpose of Meta Eval:

- Assess the reliability of *automated metrics*

What's the reliability?

元评估

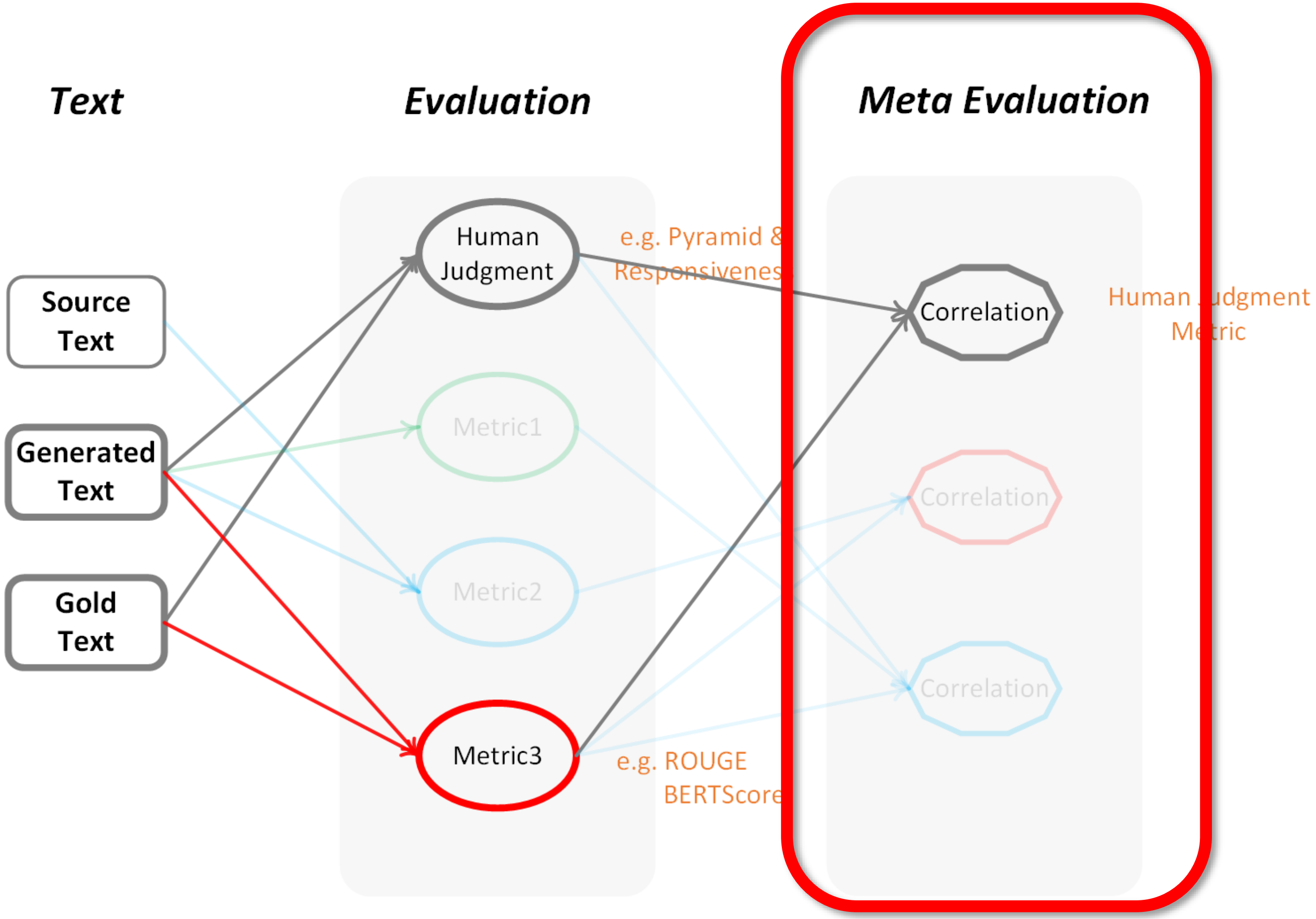


Purpose of Meta Eval:

- Assess the reliability of *automated metrics*

What's the reliability?

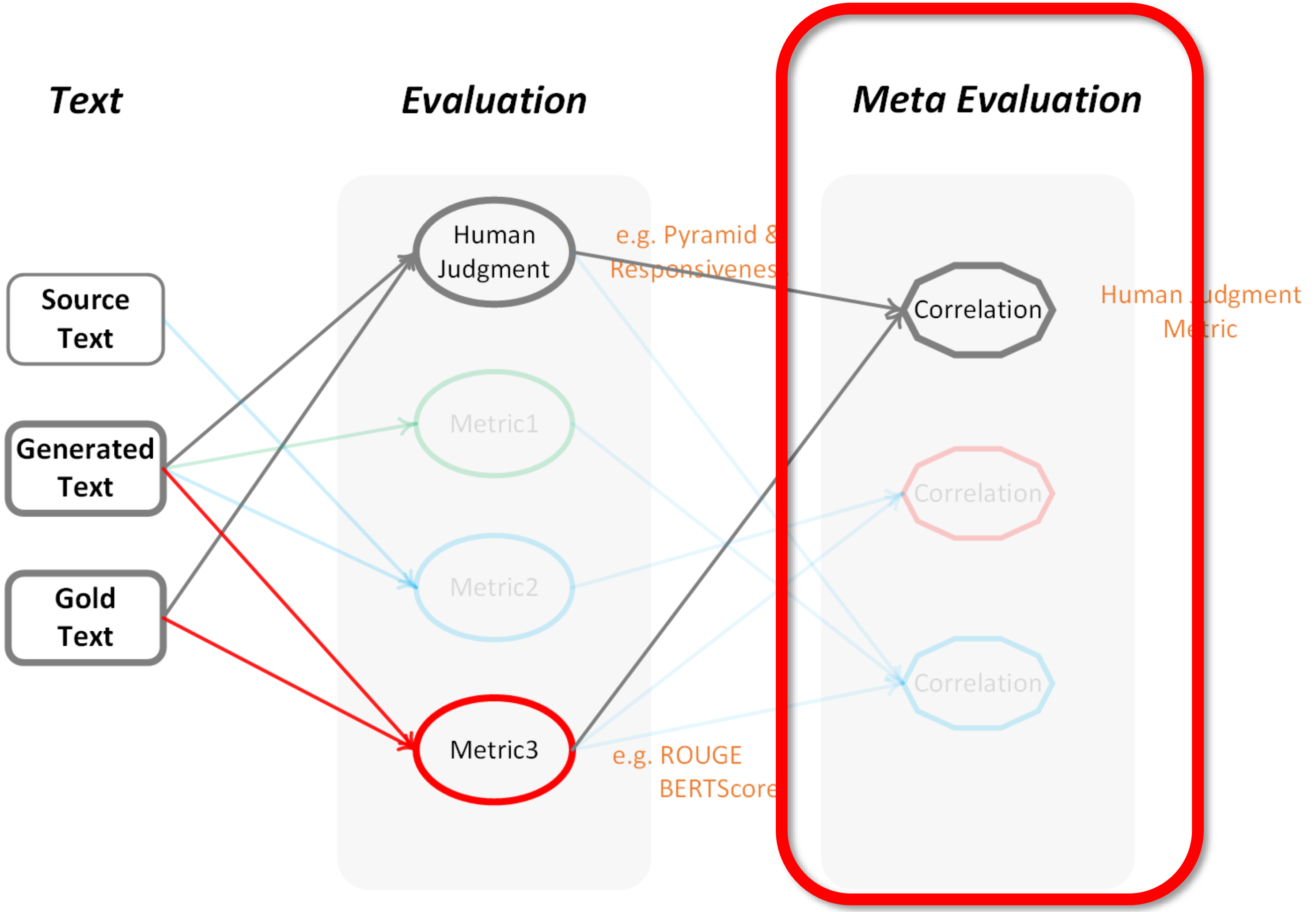
元评估



Meta evaluation dominates the direction of system optimization?

Evaluation metric -> systems

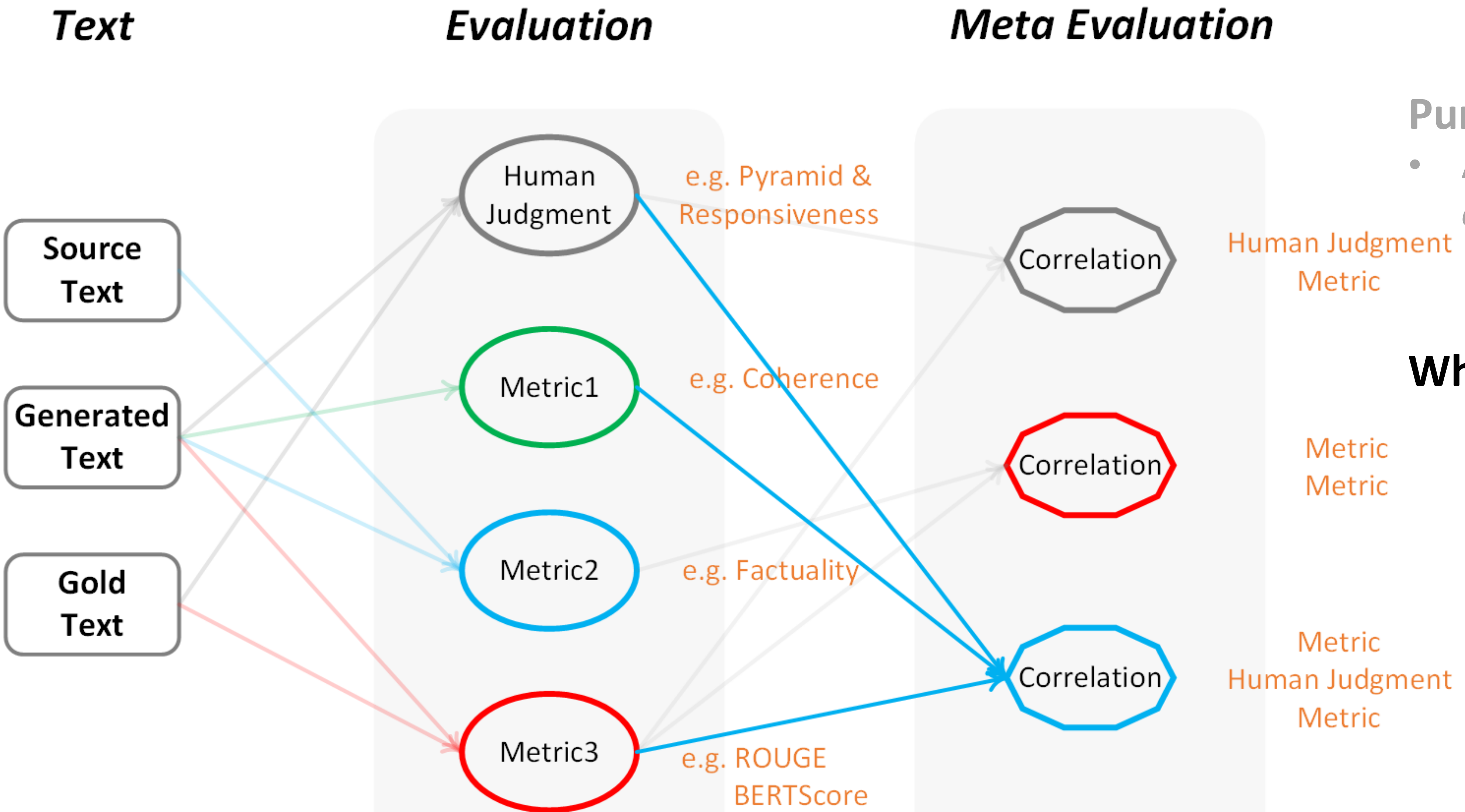
元评估



Meta evaluation dominates the direction of system optimization?

Meta eval. -> Evaluation metric -> systems

元评估挑战：数据集过时

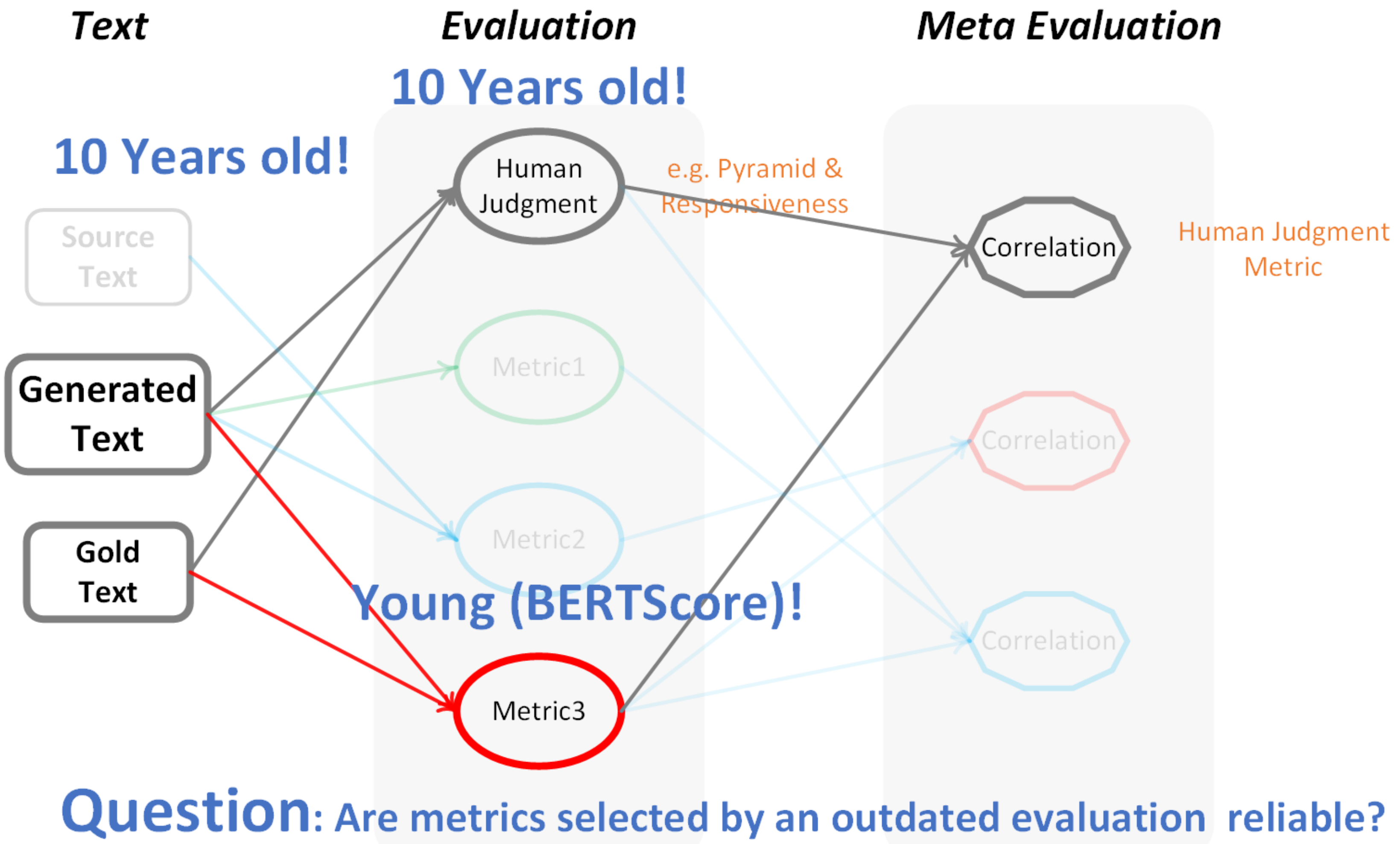


Purpose of Meta Eval:

- Assess the reliability of *automated metrics*

What's the reliability?

元评估挑战：数据集过时

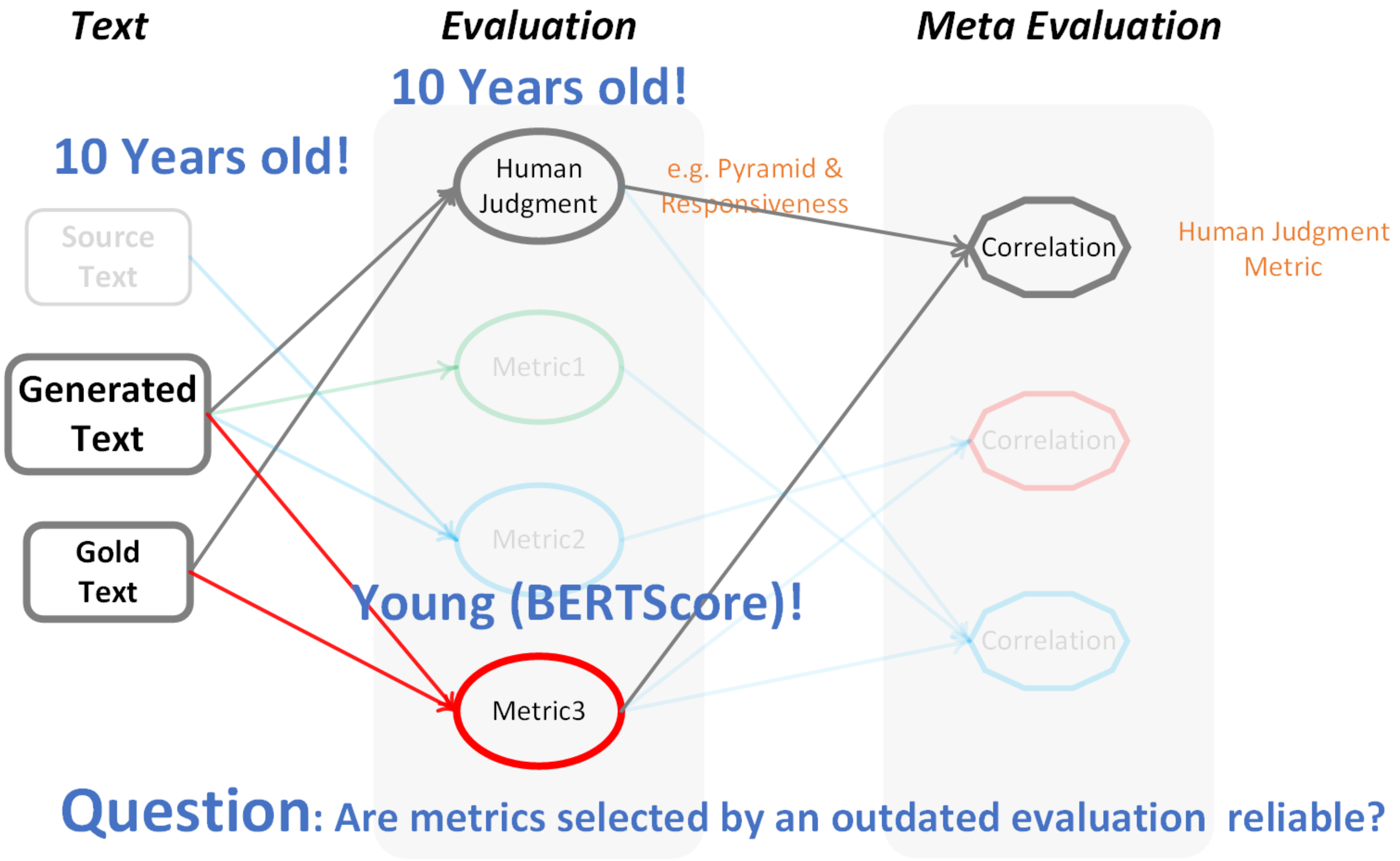


Question: Are metrics selected by an outdated evaluation reliable?

When calculating reliability (correlation), we need

- Metric judgment
- Human judgment

元评估挑战：数据集过时



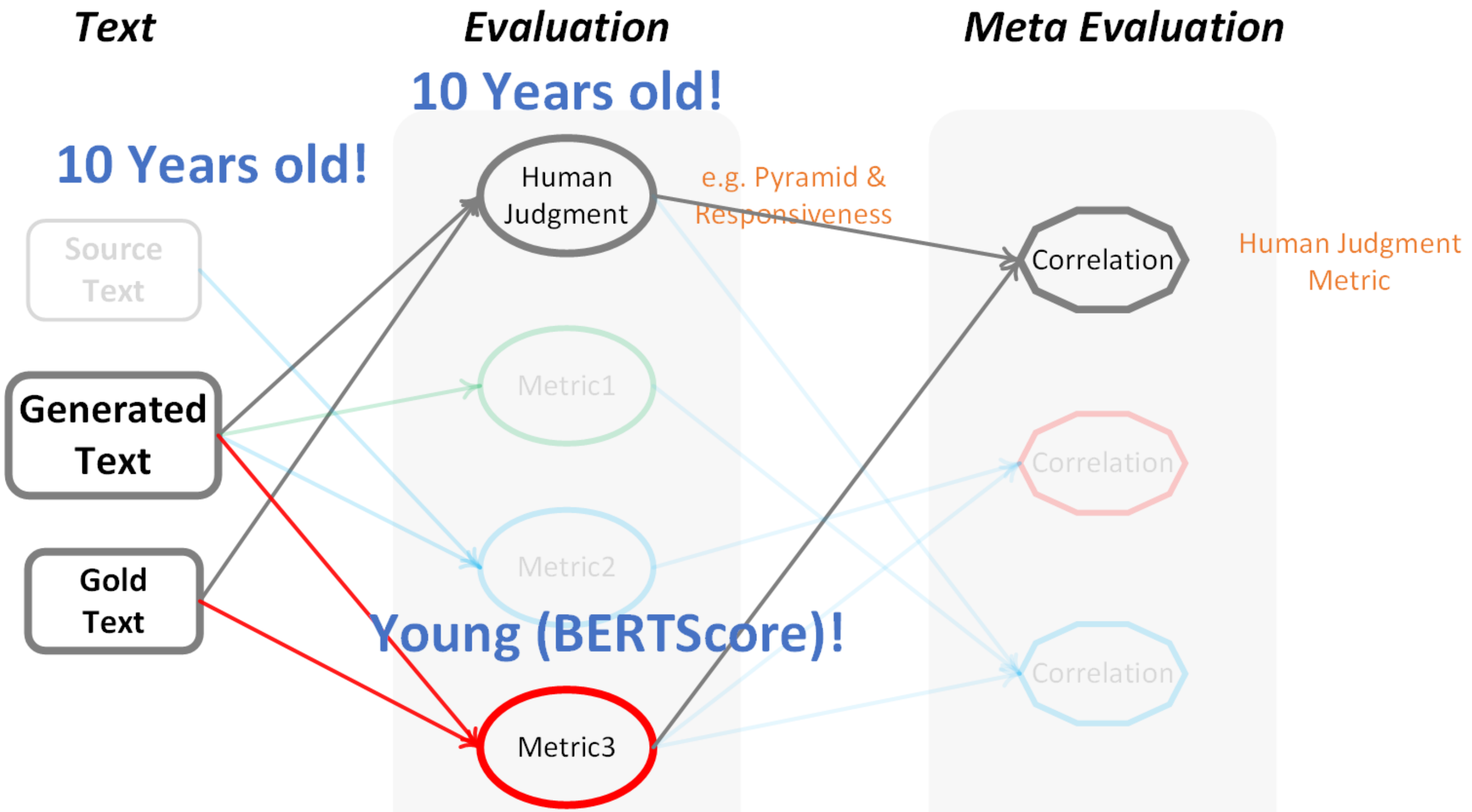
Question: Are metrics selected by an outdated evaluation reliable?

When calculating reliable score, we need

- Metric score
- Human judgment

outdated!

元评估挑战：数据集过时

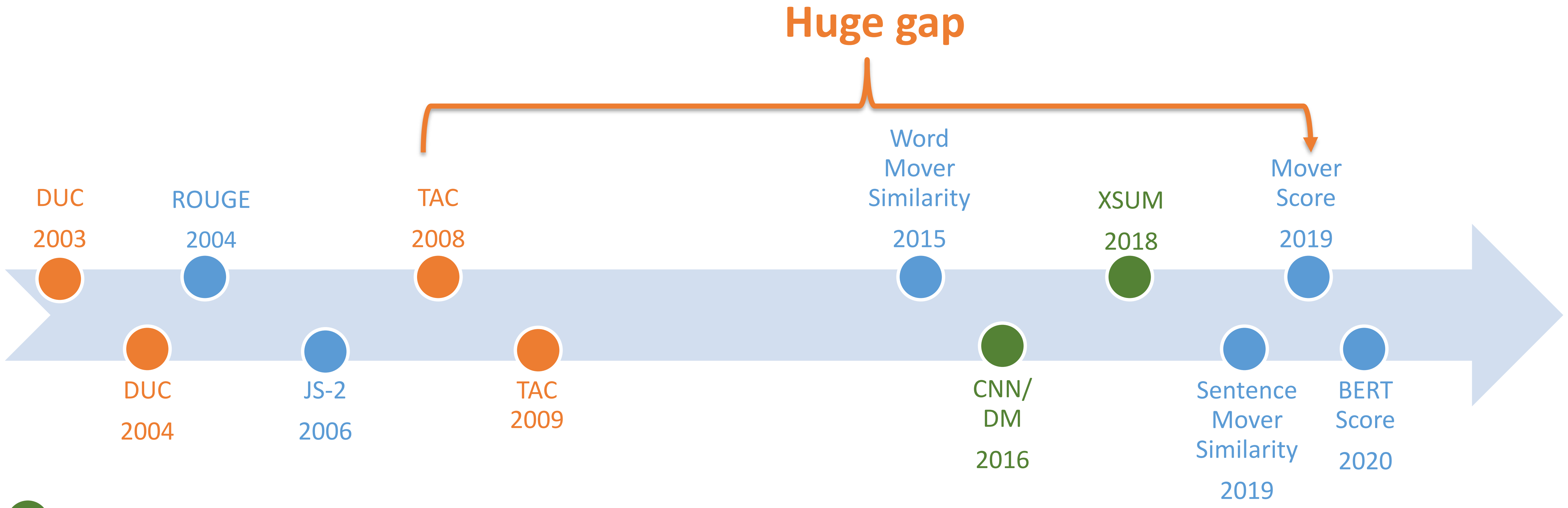


Question: Are metrics selected by an outdated evaluation reliable?

When calculating reliable score, we need

- Metric score
- Human judgment
  - Gold text
  - Generated text from diverse systems

outdated! (e.g. TAC)



Huge gap

DUC 2003

DUC 2004

ROUGE 2004

JS-2 2006

TAC 2008

TAC 2009

Word Mover Similarity 2015

CNN/ DM 2016

XSUM 2018

Mover Score 2019

Sentence Mover Similarity 2019

BERT Score 2020

● Summarization dataset

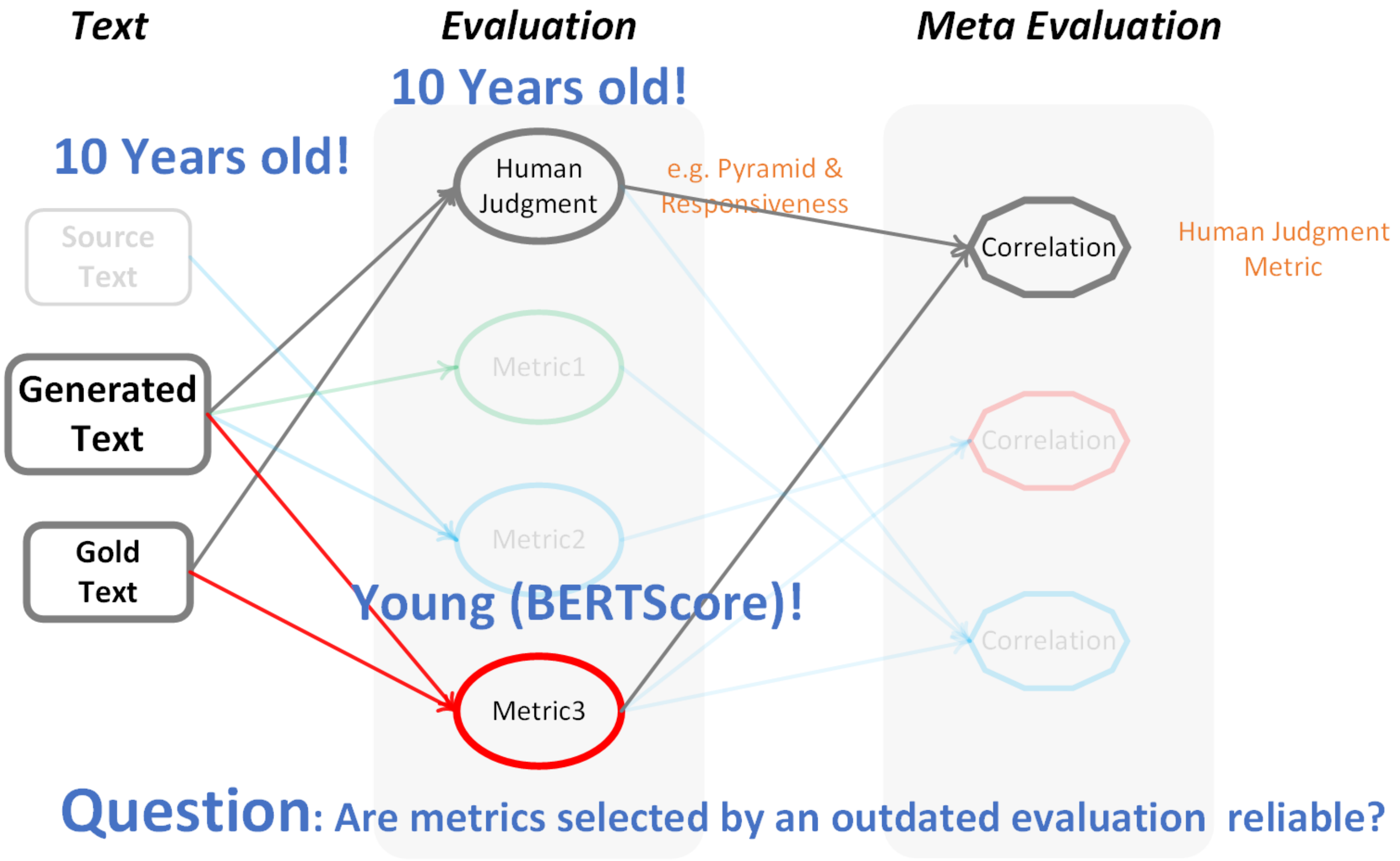
● Meta-evaluation dataset

● Automatic evaluation metric

元评估数据集发展

Can meta-evaluation on old datasets be trusted?

思考：元评估其它挑战?



Question: Are metrics selected by an outdated evaluation reliable?

When calculating reliable score, we need

- Metric score
- Human judgment
  - Gold text
  - Generated text from diverse systems