**AIGC**

**数据标注的重要性**

|  |
| --- |
| 数据标注是AI认识世界的起点 |

|  |
| --- |
| **数据标注**的本质是将现实世界信息结构化、数字化，充分发挥数据信息的价值。  作为底层服务，数据标注贯穿大模型全生命周期（训练测试、评估验证和应用迭代） |

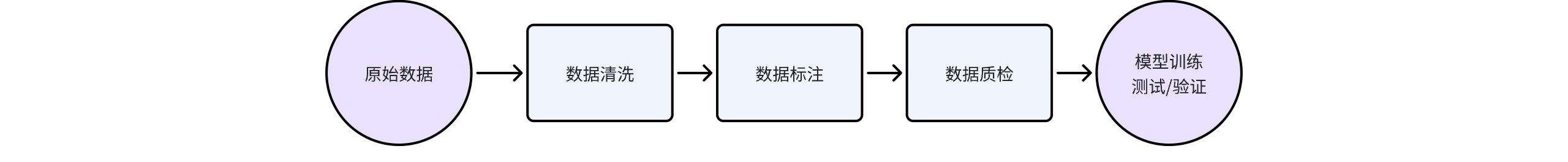
数据标注中的二八定律：80%时间在准备数据，模型训练和部署仅占20%

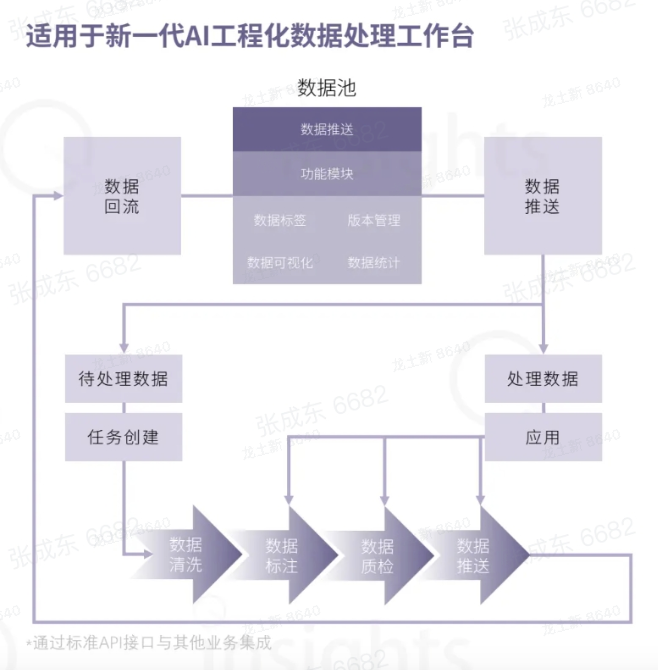
自动化标注：

合成数据：无版权问题，Sam Altman说“未来所有数据都将变成合成数据”

标注角色细分：AI训练师、模型精调师、指令工程师等

高质量数据需求长期且持续





**数据处理管线pipeline**

**模型训练**

* 训练配置文件[name].yaml

|  |
| --- |
| YAML experiment:  exp\_dir: experiments/  exp\_name: pku\_cws  seed: 42  task: word-segmentation  dataset:  data\_file:  train: https://modelscope.cn/api/v1/datasets/dingkun/chinese\_word\_segmentation\_pku/repo?Revision=master&FilePath=train.txt  dev: https://modelscope.cn/api/v1/datasets/dingkun/chinese\_word\_segmentation\_pku/repo?Revision=master&FilePath=dev.txt  test: https://modelscope.cn/api/v1/datasets/dingkun/chinese\_word\_segmentation\_pku/repo?Revision=master&FilePath=test.txt  data\_type: conll  preprocessor:  type: sequence-labeling-preprocessor  max\_length: 256  tag\_scheme: BIES  data\_collator: SequenceLabelingDataCollatorWithPadding  model:  type: sequence-labeling-model  embedder:  model\_name\_or\_path: damo/nlp\_structbert\_word-segmentation\_chinese-base  dropout: 0.1  use\_crf: true  train:  max\_epochs: 30  dataloader:  batch\_size\_per\_gpu: 32  optimizer:  type: AdamW  lr: 2.0e-5  param\_groups:  - regex: crf  lr: 2.0e-1  lr\_scheduler:  type: LinearLR  start\_factor: 1.0  end\_factor: 0.0  total\_iters: 30  evaluation:  dataloader:  batch\_size\_per\_gpu: 64  metrics:  - type: ner-metric  - type: ner-dumper  model\_type: sequence\_labeling  dump\_format: conll |