东南大学 SOUTHEAST UNIVERSITY

计算机科学与工程学院
School of computer science and engineering

Seciton 2:
AI for Code Generation

# AutoMMLab: Automatically Generating Deployable Models from Language Instructions for Computer Vision Tasks
# 通过语言指令为计算机视觉任务自动生成可部署模型

AAAI'25


**Zekang Yang[1], Wang Zeng[1], Sheng Jin[2,1]*, Chen Qian[1], Ping Luo[2], Wentao Liu[1]***

[1]SenseTime Research and Tetras.AI
[2]The University of Hong Kong
{yangzekang, zengwang, jinsheng, qianchen, liuwentao}@tetras.ai pluo@cs.hku.hk

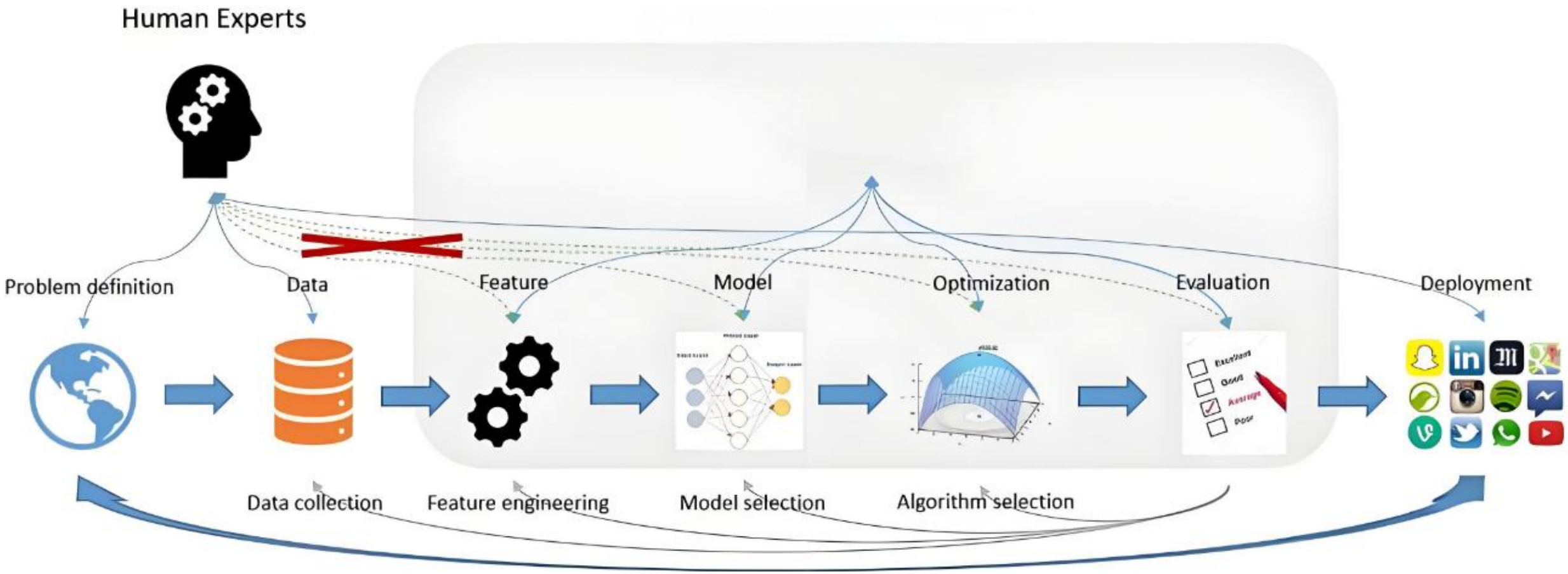

汇报人：王维龙 2025 年 9 月 18 日

# 提纲

# 提纲

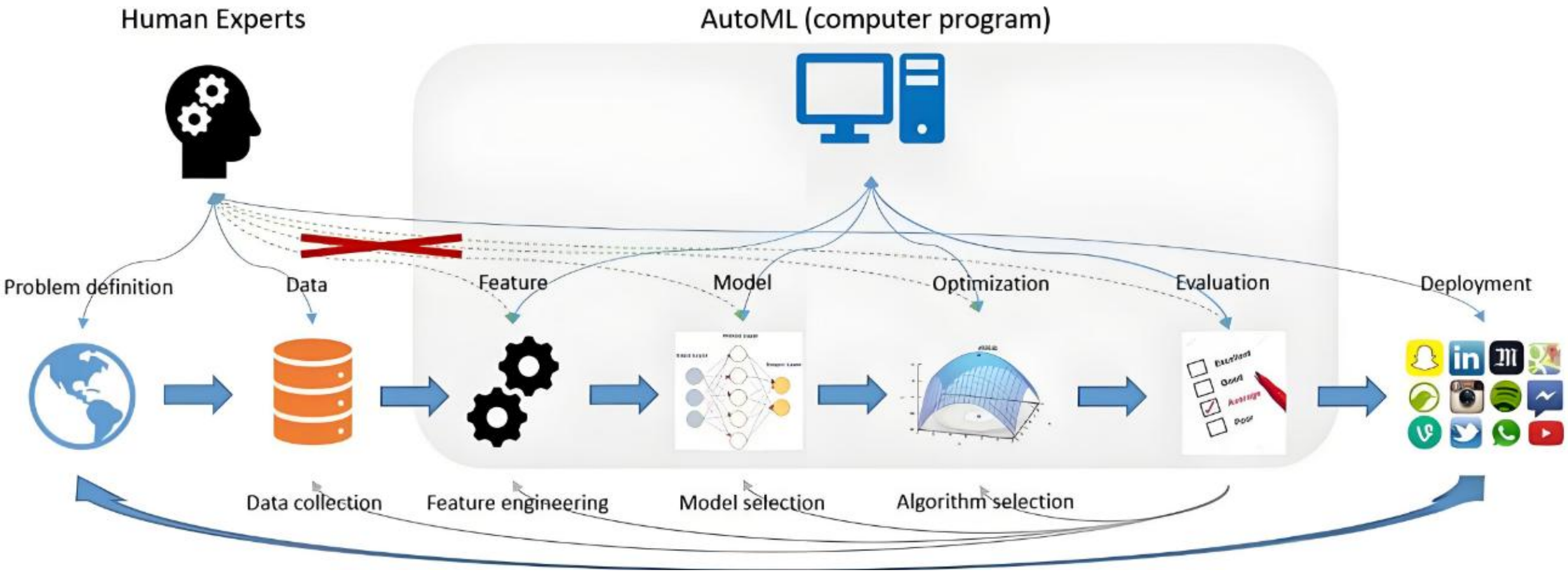
# 背景概况

## AI模型构建全流程



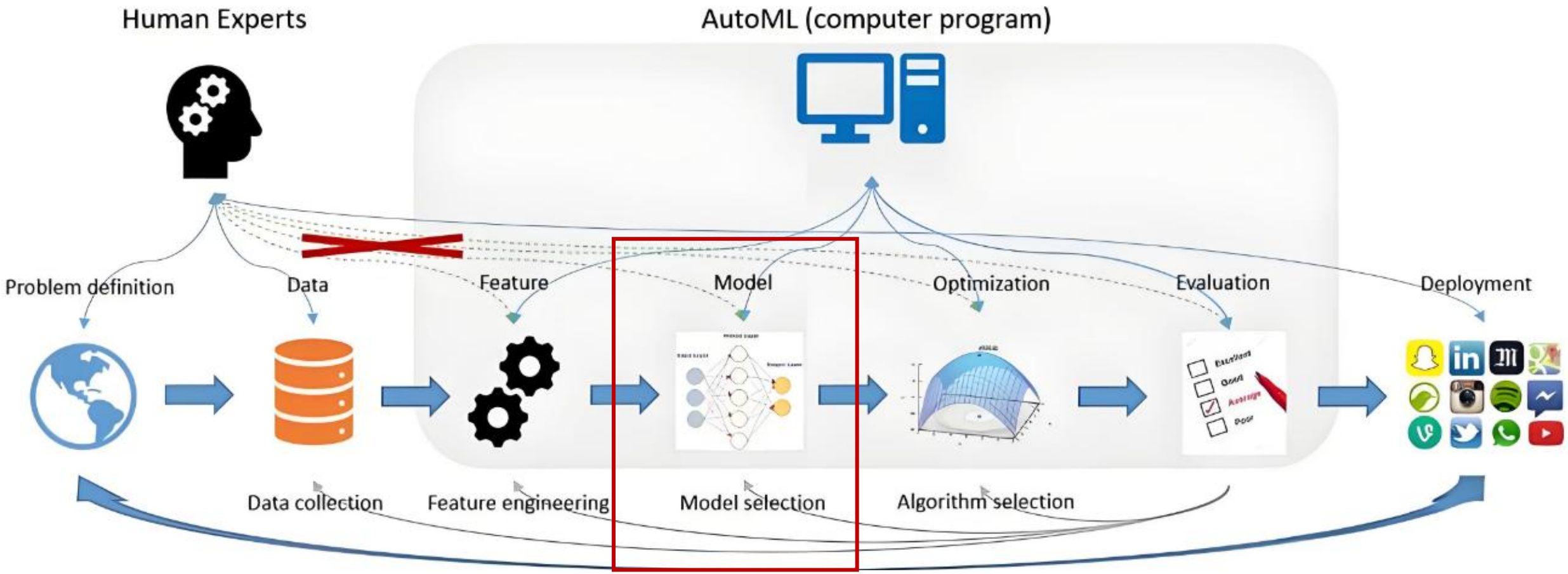
# 背景概况

## 自动化机器学习（Auto Machine Learning, AutoML）



# 背景概况

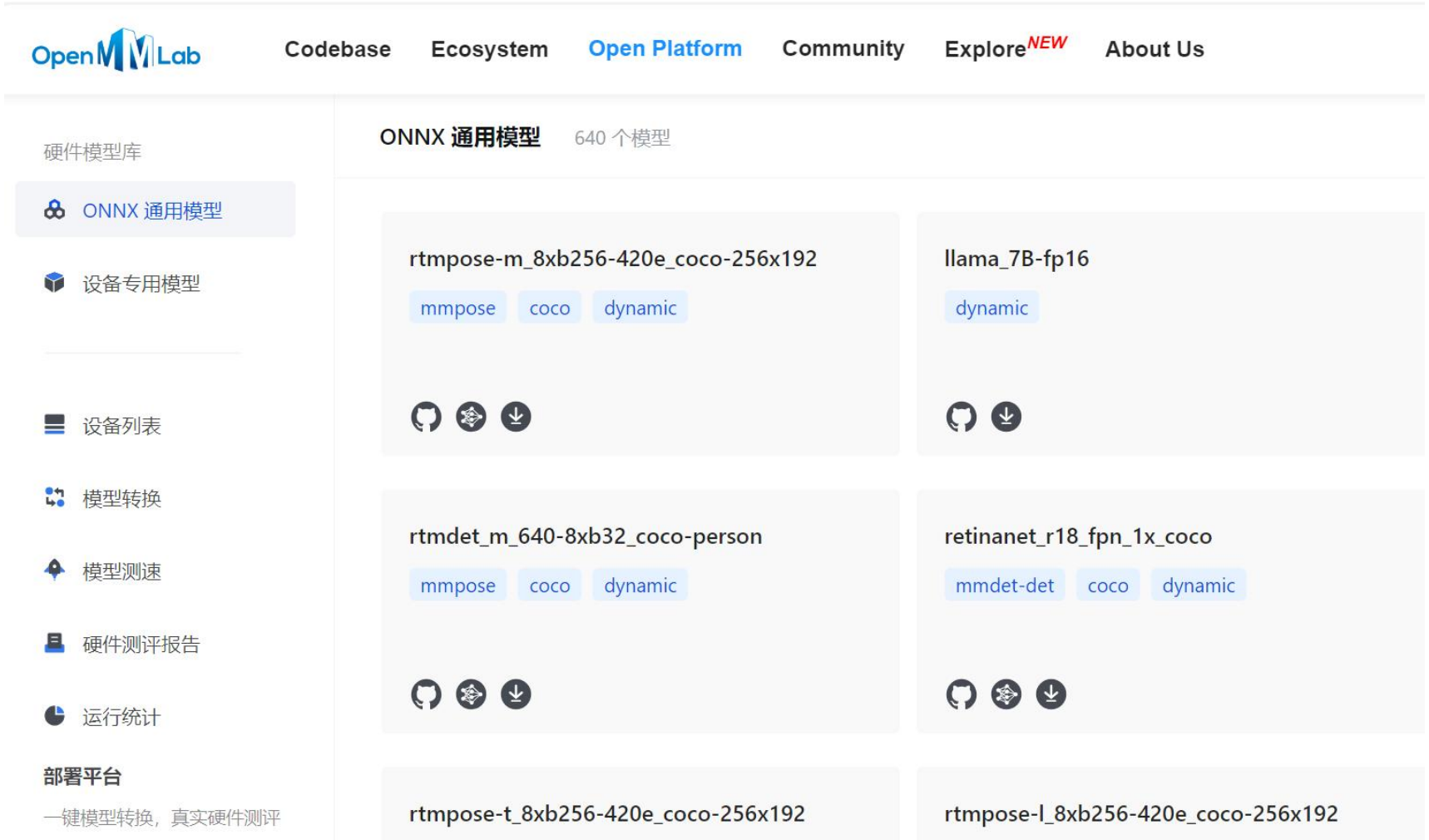
## 自动化机器学习（Auto Machine Learning, AutoML）



# 背景概况

## OpenMMLab

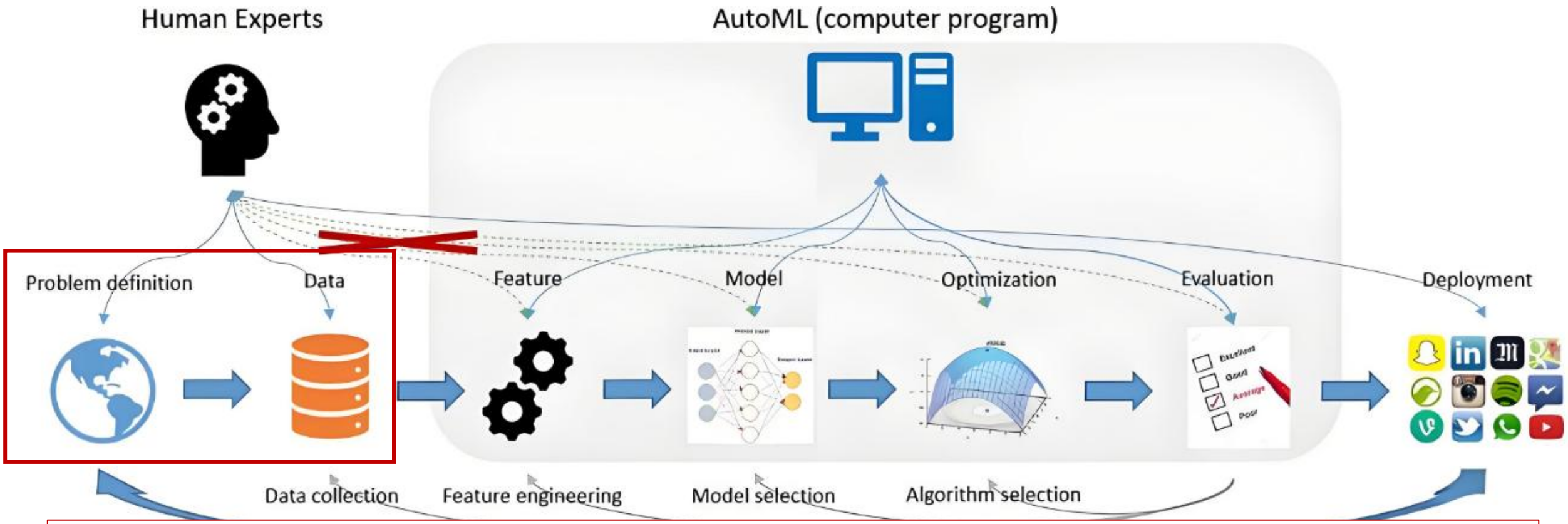
- 累计开源超过30个算法库、2400个预训练模型。
- 涵盖图像识别分类、目标检测、视频理解、预训练、多模态、AIGC等等计算机视觉任务。
- 在Github上累计获得十三万颗star，超过了Pytorch。




https://openmmlab.com/

# 背景概况

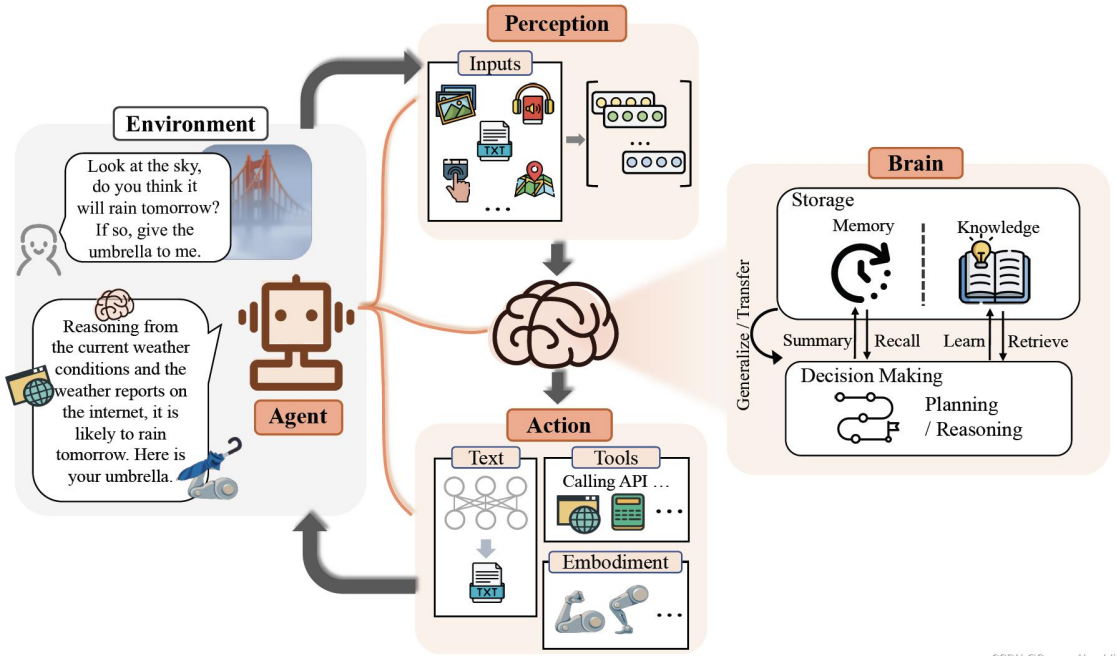
## 自动化机器学习（Auto Machine Learning, AutoML）




整个流程仍需要人工参与，并且要求具备一定专业知识。

# 背景概况

## 大语言模型（Large Language Model, LLM）

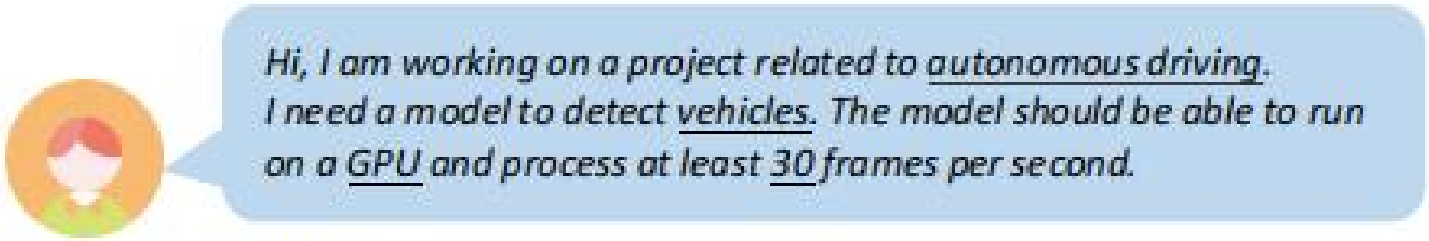



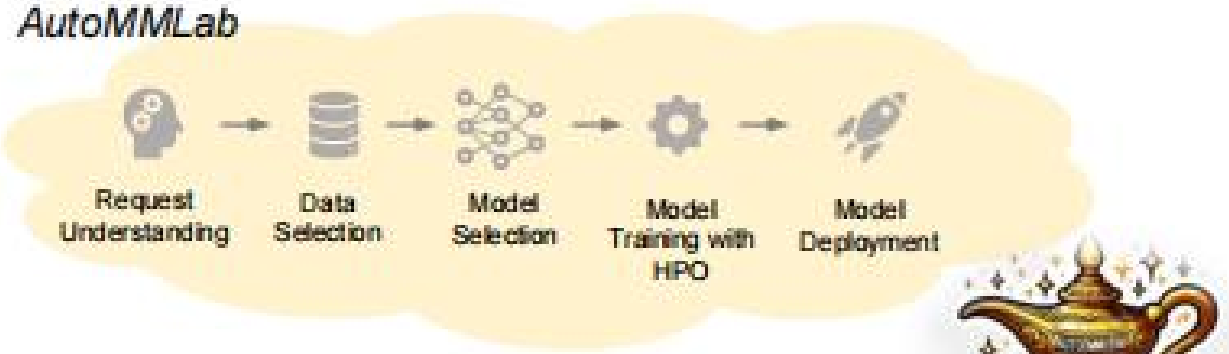

**LLM具备强大的理解、生成和使用工具的能力，为全自动化模型生成提供了新的求解思路。**

# 背景概况

AutoML + OpenMMLab

**Input:** Request (Data & Model & Deploy)

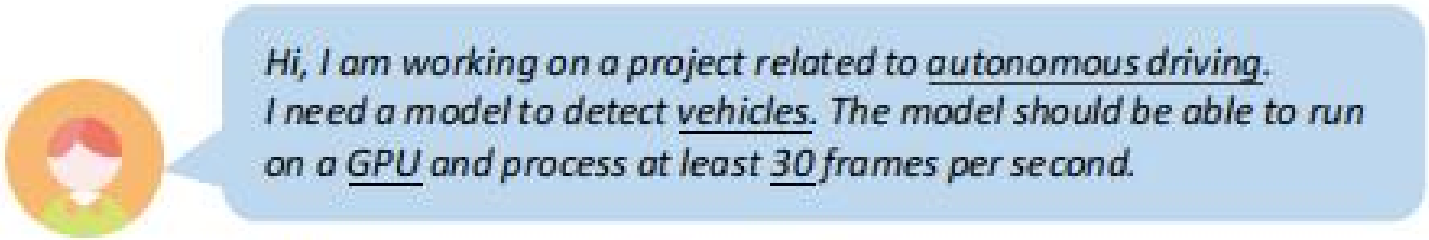


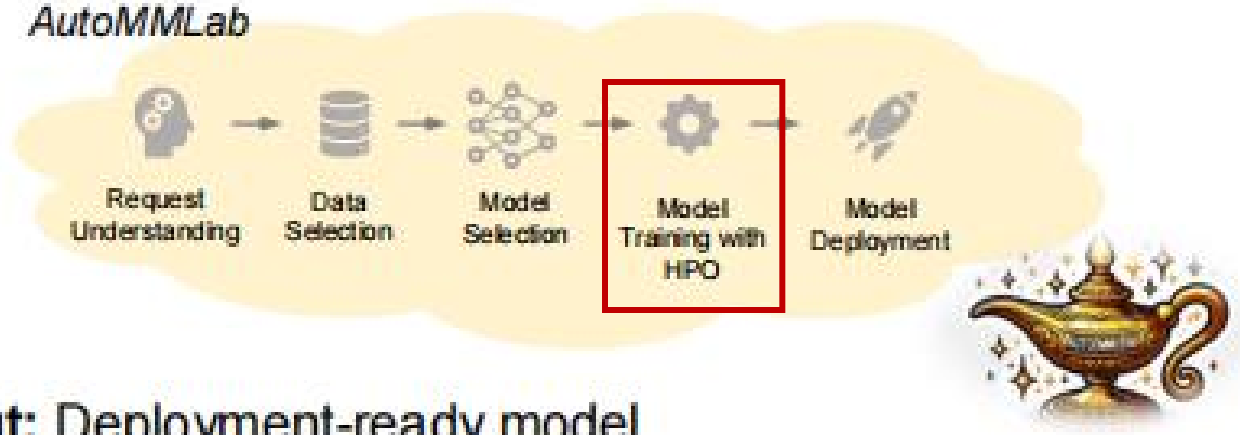




**Output:** Deployment-ready model

# 背景概况

**Input:** Request (Data & Model & Deploy)




**Output:** Deployment-ready model

# 提纲

1 研究背景

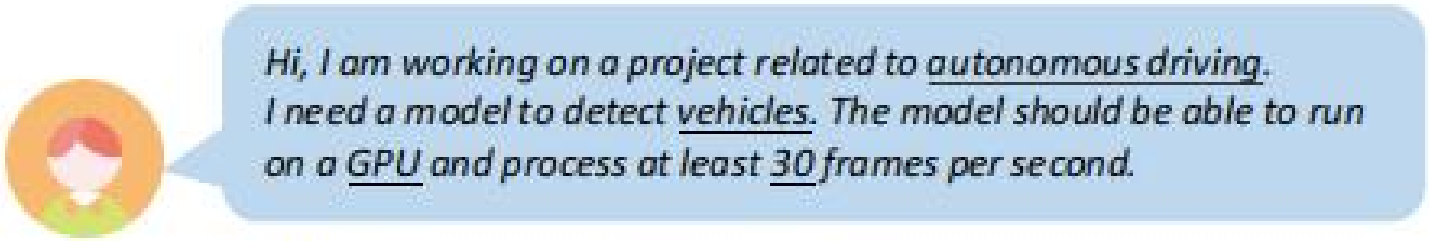
2 相关工作

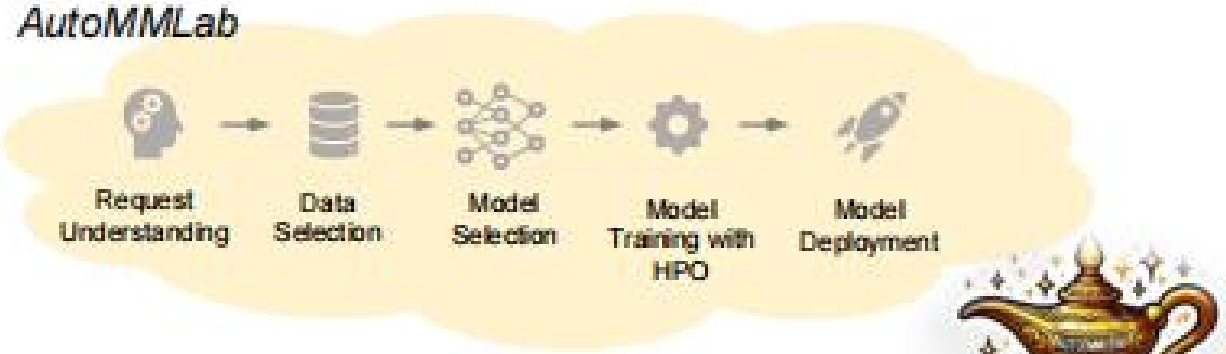
3 研究内容

4 实验评估

# 背景概况

**Input:** Request (Data & Model & Deploy)



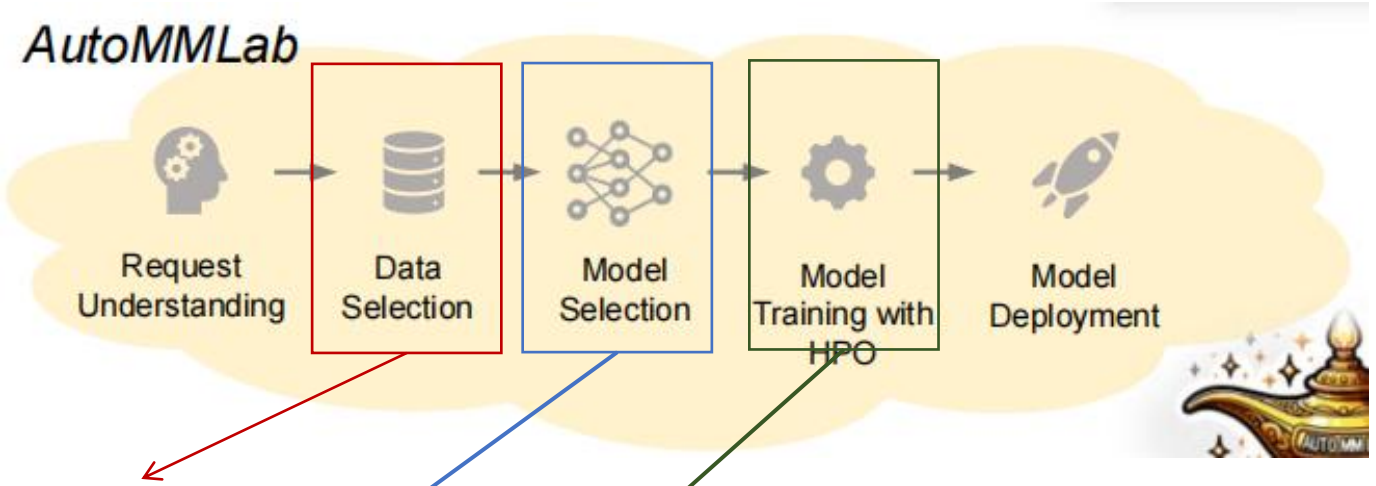




**Output:** Deployment-ready model

# 相关工作

## 基于LLM的AutoML

[1] Hong, S.; Lin, Y.; Liu, B.; Liu, B.; Wu, B.; Zhang, C.; Wei, C.; Li, D.; Chen, J.; Zhang, J.; et al. 2025. Data interpreter: An llm agent for data science. ACL 2025.
[2]Zheng, M.; Su, X.; You, S.; Wang, F.; Qian, C.; Xu, C.; and Albanie, S. 2023. Can GPT-4 Perform Neural Architecture Search? arXiv:2304.10970.
[3] Zhang, M. R.; Desai, N.; Bae, J.; Lorraine, J.; and Ba, J. 2023. Using large language models for hyperparameter optimization. In NeurIPS 2023 Foundation Models for Decision Making Workshop.

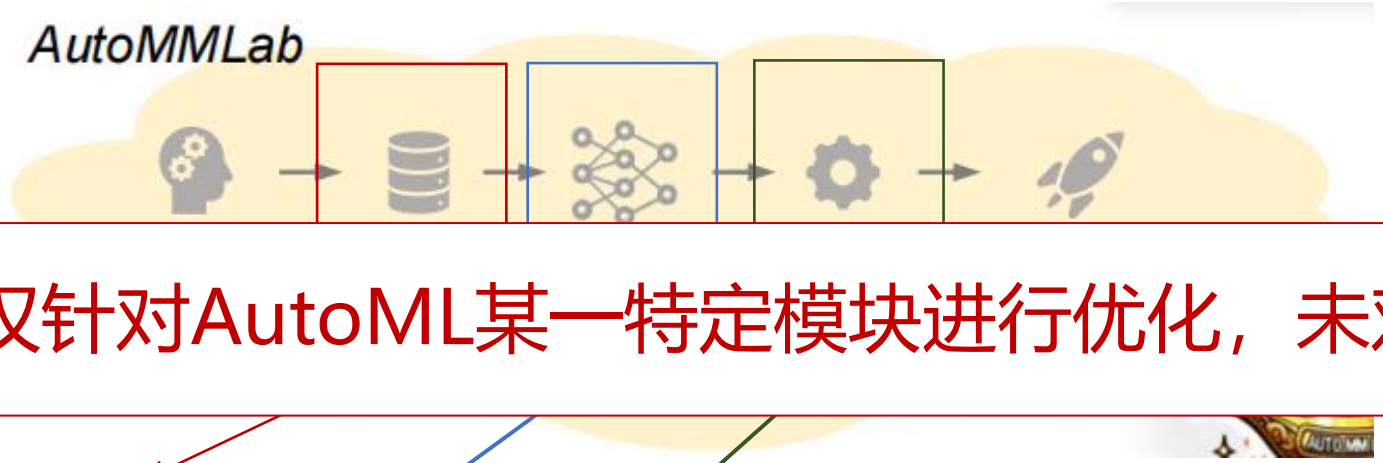



工作[1]仅考虑**数据的理解与分析**，利用LLM自动分析结构化数据，并且缺乏自适应能力。
工作[2]和仅考虑了**模型的架构**，重点面向神经网络架构搜索问题。
工作[3]仅考虑**超参数优化**，主要是证明了基于LLM的方法优于随机搜索和贝叶斯优化。

# 相关工作

## 基于LLM的AutoML

[1] Hong, S.; Lin, Y.; Liu, B.; Liu, B.; Wu, B.; Zhang, C.; Wei, C.; Li, D.; Chen, J.; Zhang, J.; et al. 2025. Data interpreter: An llm agent for data science. ACL 2025.
[2]Zheng, M.; Su, X.; You, S.; Wang, F.; Qian, C.; Xu, C.; and Albanie, S. 2023. Can GPT-4 Perform Neural Architecture Search? arXiv:2304.10970.
[3] Zhang, M. R.; Desai, N.; Bae, J.; Lorraine, J.; and Ba, J. 2023. Using large language models for hyperparameter optimization. In NeurIPS 2023 Foundation Models for Decision Making Workshop.

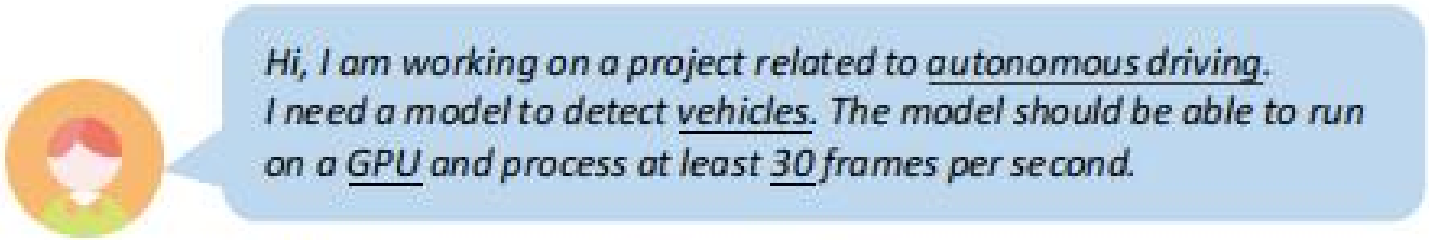


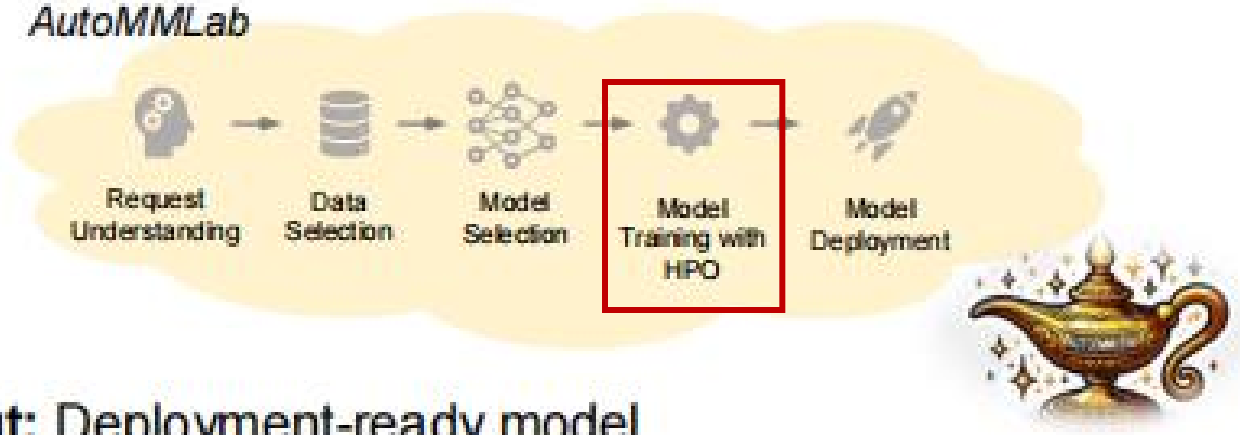

局限性1: 仅针对AutoML某一特定模块进行优化，未对全流程优化。

工作[1]仅考虑**数据的理解与分析**，利用LLM自动分析结构化数据，并且缺乏自适应能力。
工作[2]和仅考虑了**模型的架构**，重点面向神经网络架构搜索问题。
工作[3]仅考虑**超参数优化**，主要是证明了基于LLM的方法优于随机搜索和贝叶斯优化。

# 背景概况

**Input:** Request (Data & Model & Deploy)



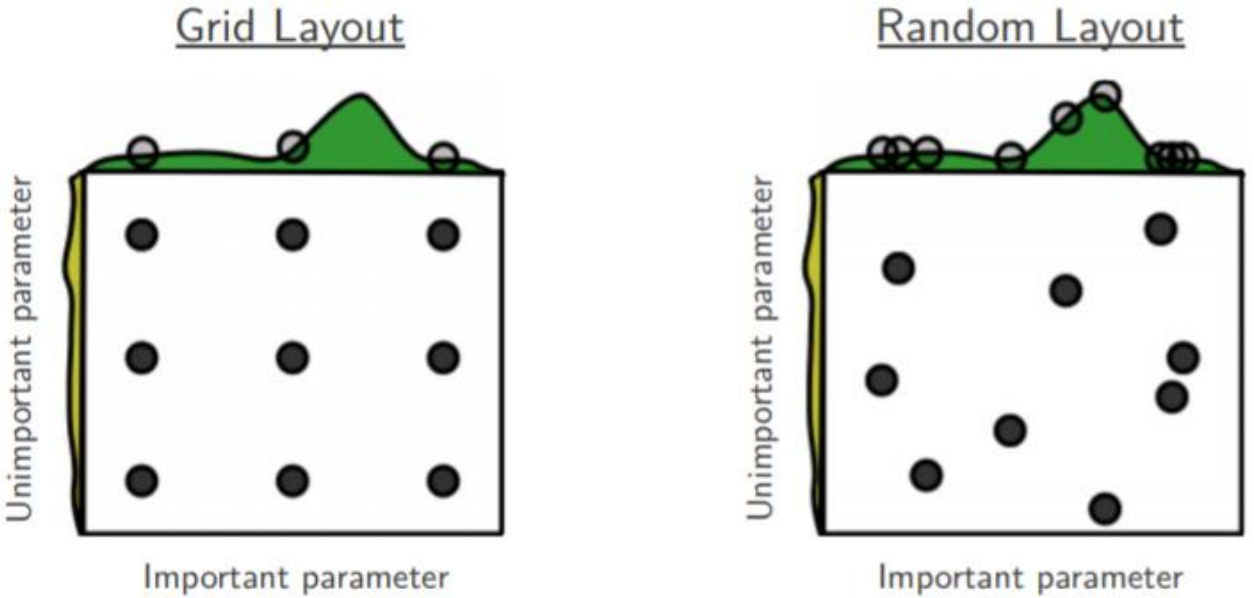




**Output:** Deployment-ready model

# 相关工作

## 超参数优化（HPO）

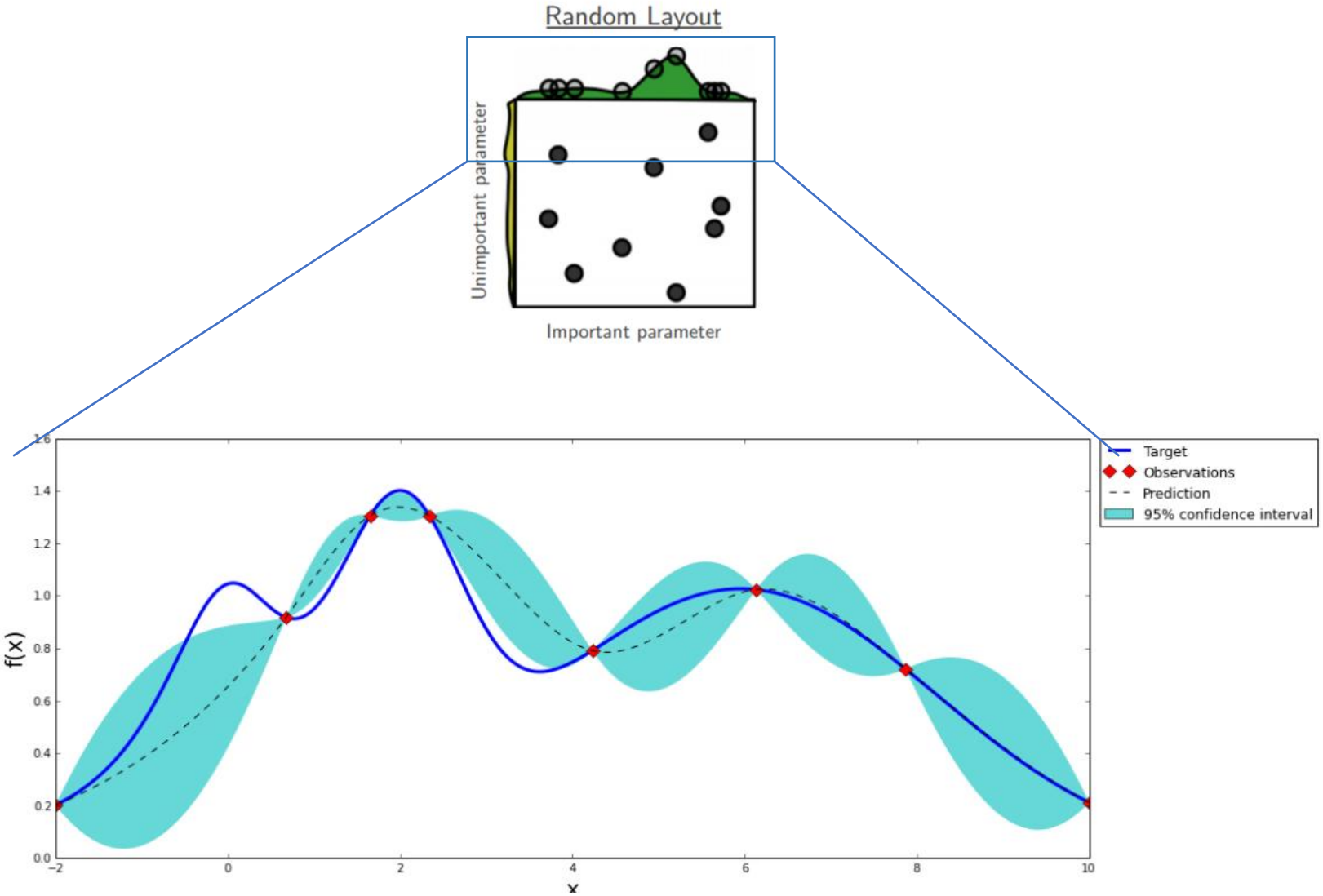
- 网格搜索
- 随机搜索



# 相关工作

## 超参数优化（HPO）

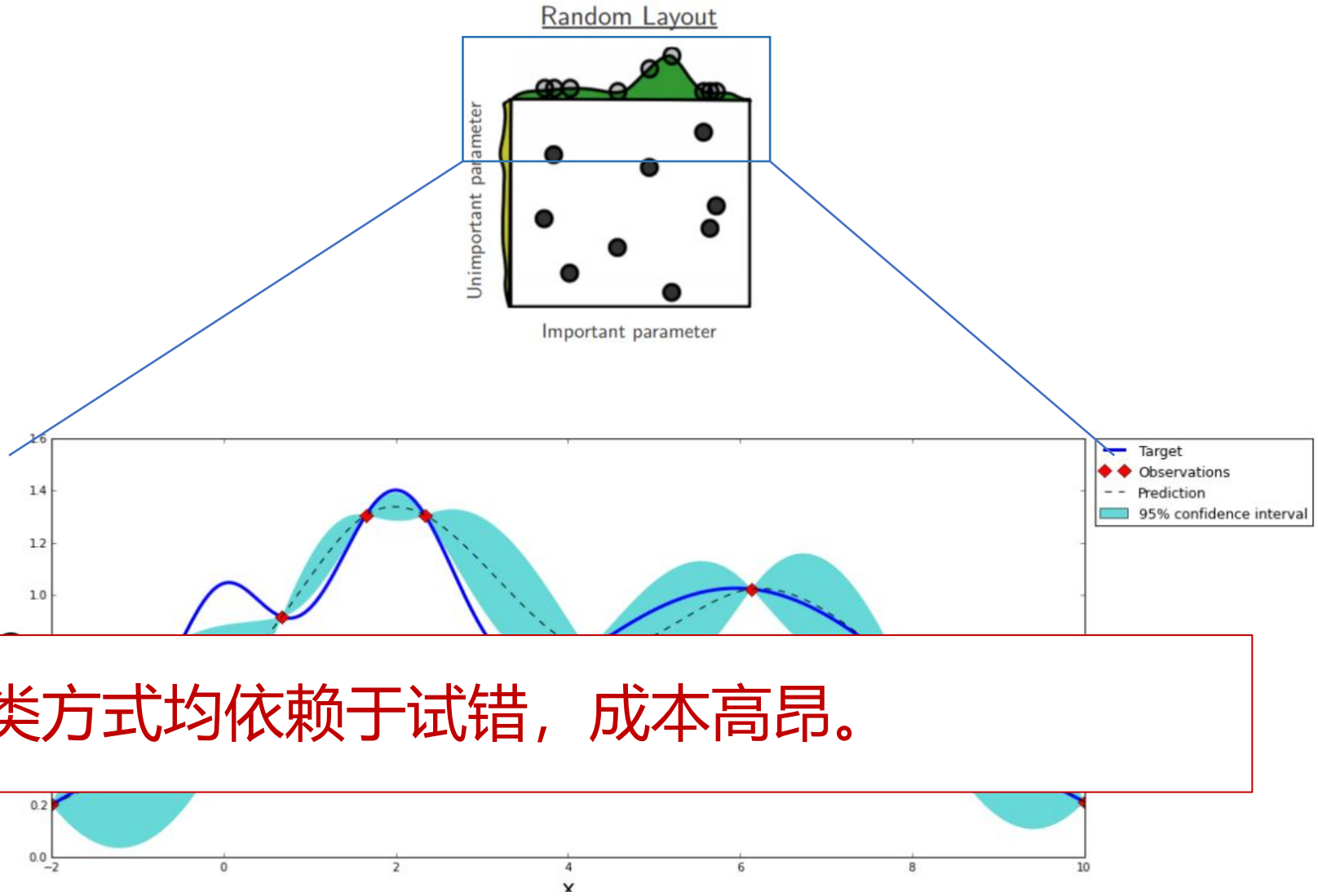
- 网格搜索
- 随机搜索
- **贝叶斯优化**
- **…**



# 局限性2

## 超参数优化（HPO）

- 网格搜索
- 随机搜索
- **贝叶斯优化**
- …

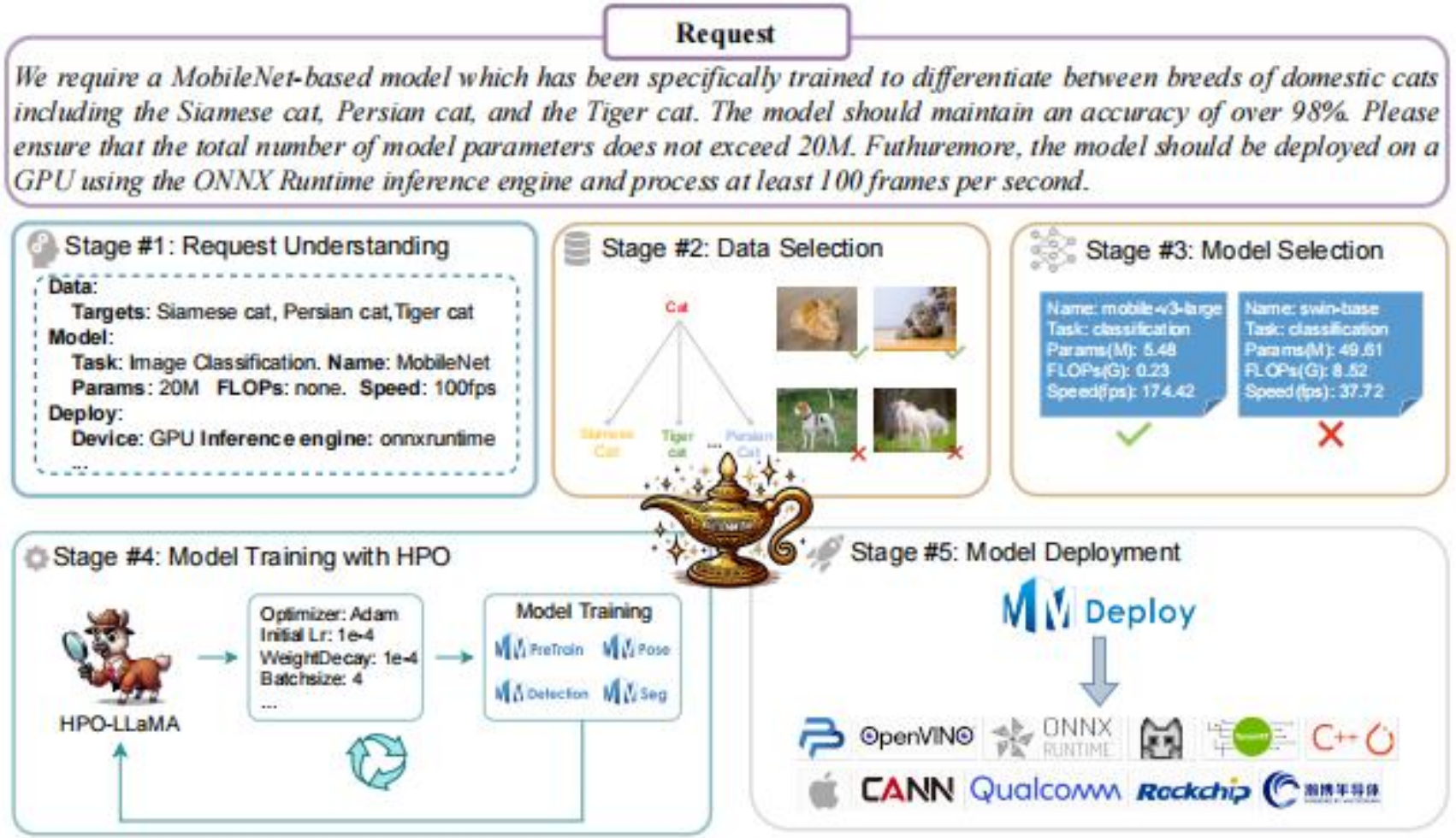



局限性2: 此类方式均依赖于试错，成本高昂。

# 提纲

# AutoMMLab

- 用户请求
- 请求理解
- 数据选择
- 模型选择
- 超参数优化和模型训练
- 模型部署



# AutoMMLab

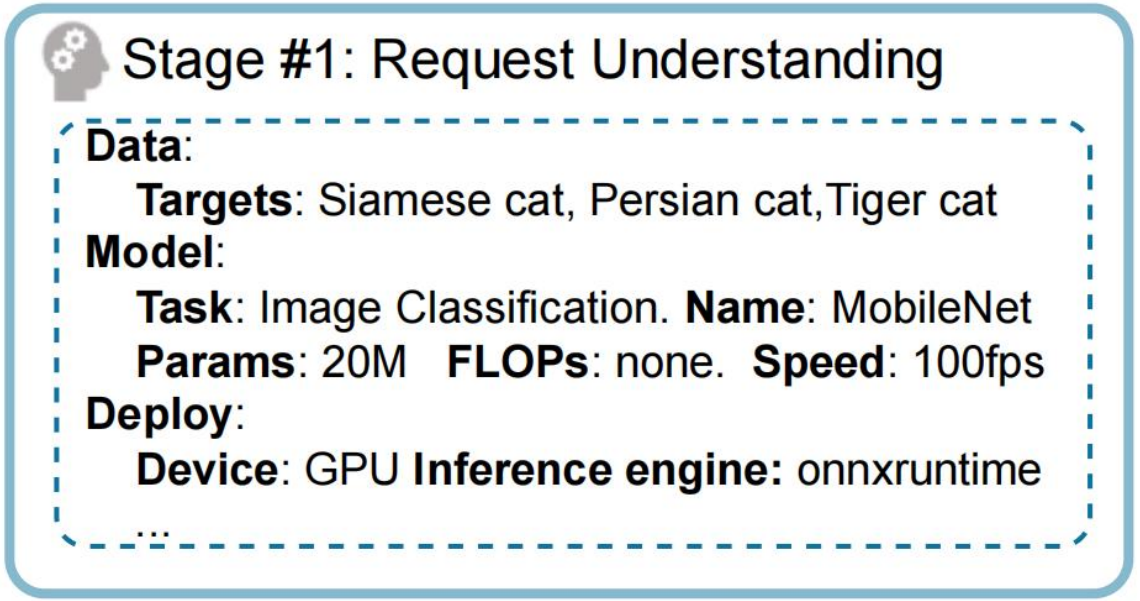
- 用户请求
- 请求理解
- 数据选择
- 模型选择
- 超参数优化和模型训练
- 模型部署

**Request**

*We require a MobileNet-based model which has been specifically trained to differentiate between breeds of domestic cats including the Siamese cat, Persian cat, and the Tiger cat. The model should maintain an accuracy of over 98%. Please ensure that the total number of model parameters does not exceed 20M. Futhuremore, the model should be deployed on a GPU using the ONNX Runtime inference engine and process at least 100 frames per second.*

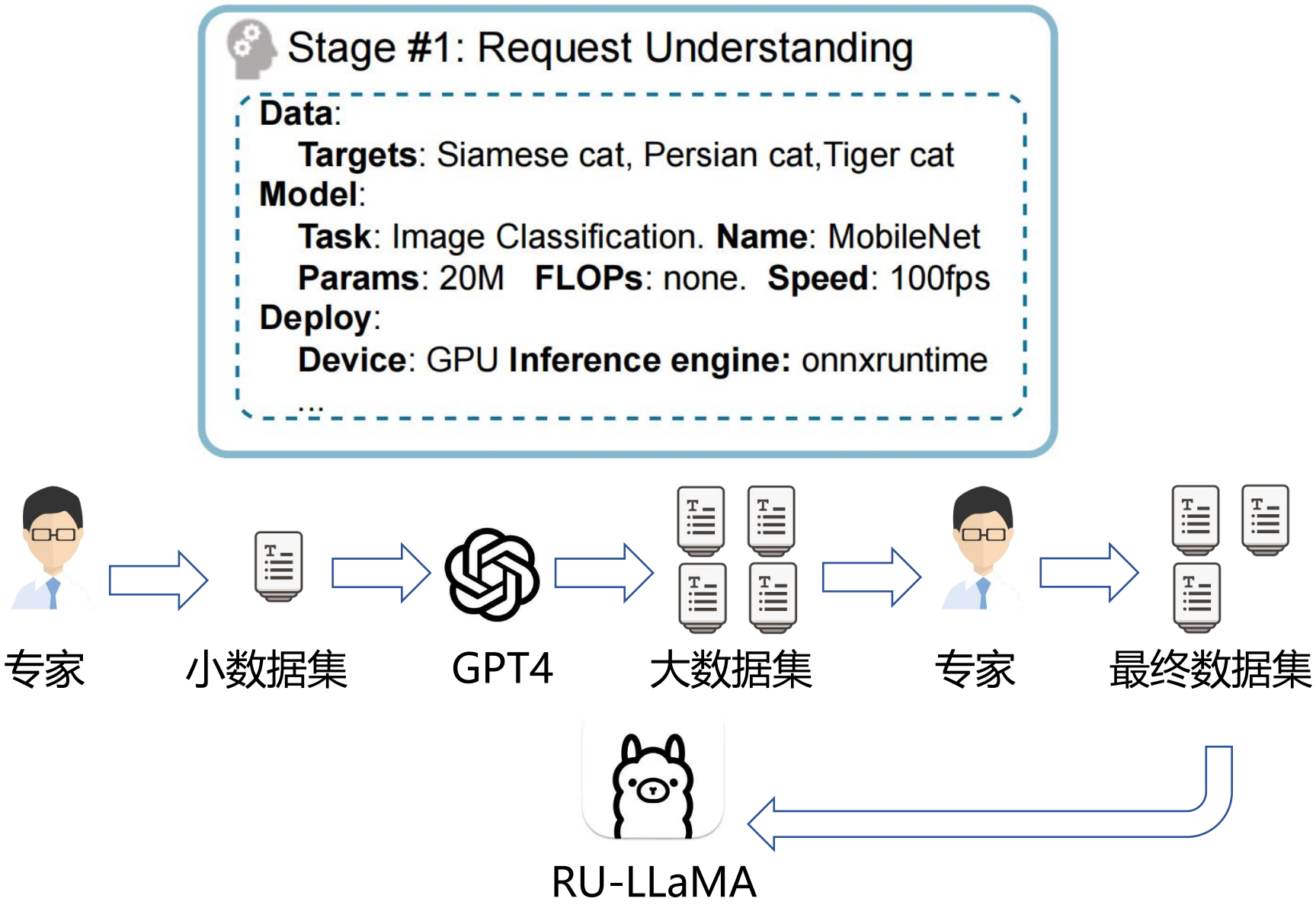
# AutoMMLab

- 用户请求
- 请求理解
- 数据选择
- 模型选择
- 超参数优化和模型训练
- 模型部署



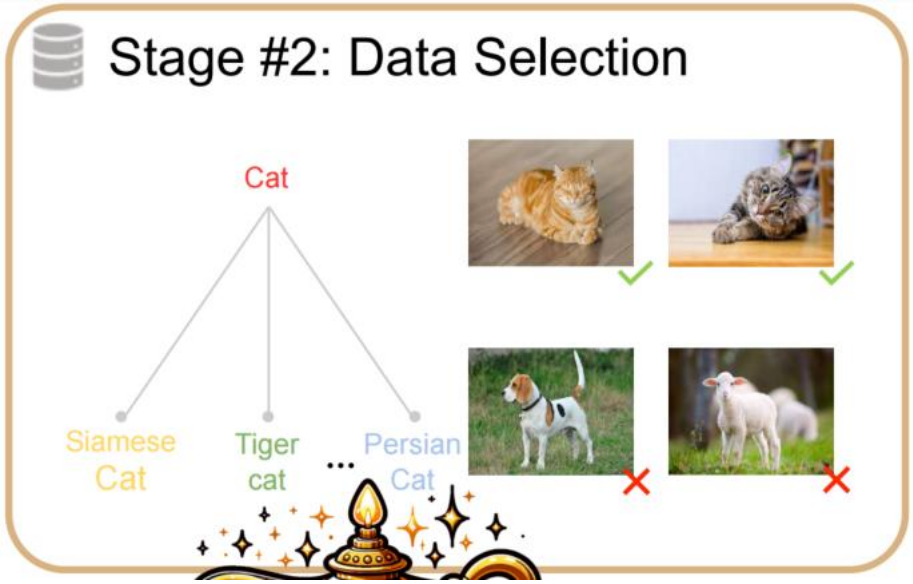
# RU-LLaMA

- 用户请求
- 请求理解
- 数据选择
- 模型选择
- 超参数优化和模型训练
- 模型部署



# AutoMMLab

- 用户请求
- 请求理解
- **数据选择**
- 模型选择
- 超参数优化和模型训练
- 模型部署



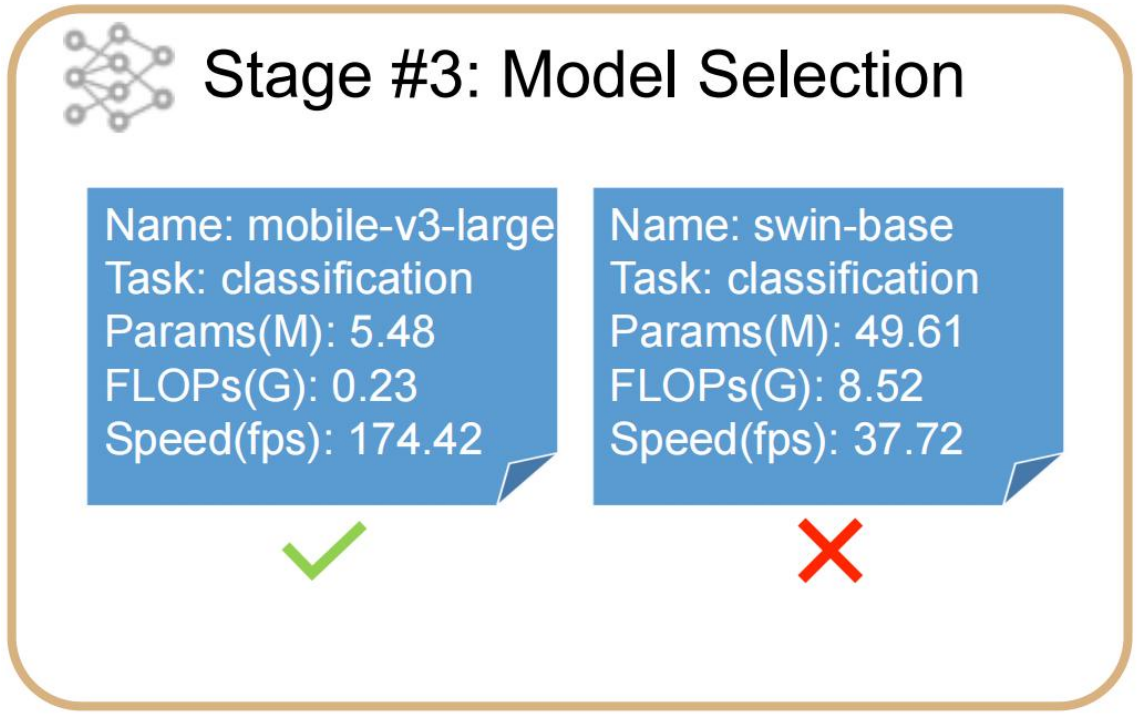

任务类型（名称）

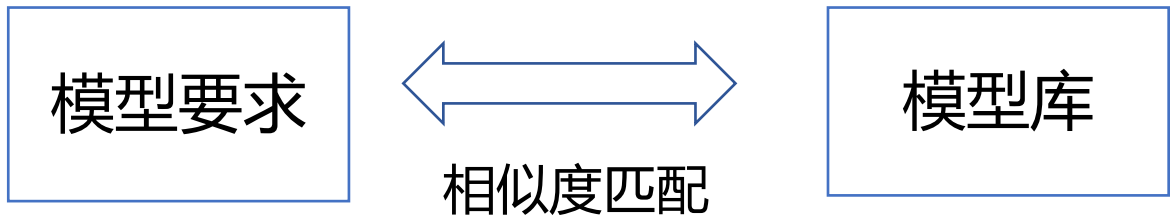
相似度匹配

数据集库

# AutoMMLab

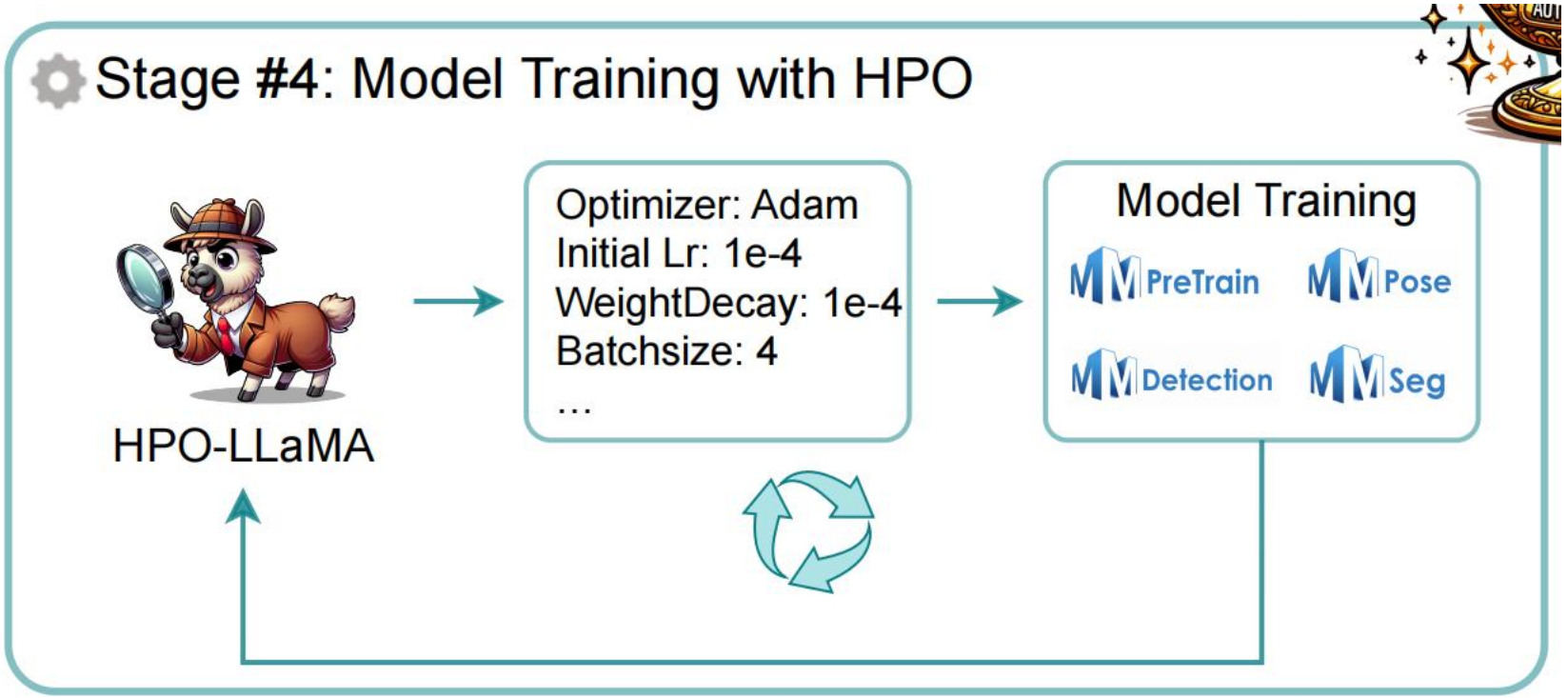
- 用户请求
- 请求理解
- 数据选择
- **模型选择**
- 超参数优化和模型训练
- 模型部署



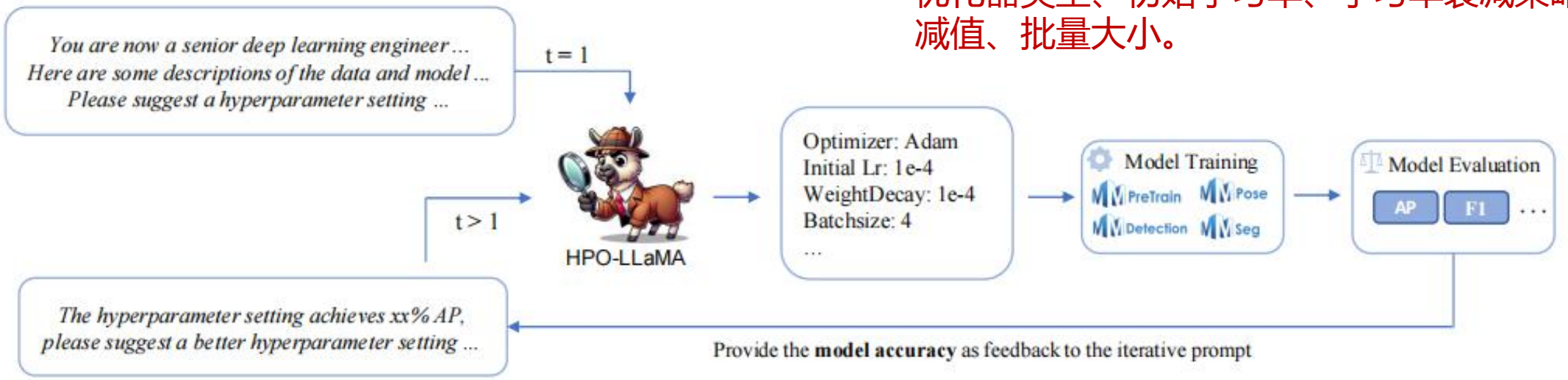
# AutoMMLab

- 用户请求
- 请求理解
- 数据选择
- 模型选择
- 超参数优化和模型训练
- 模型部署

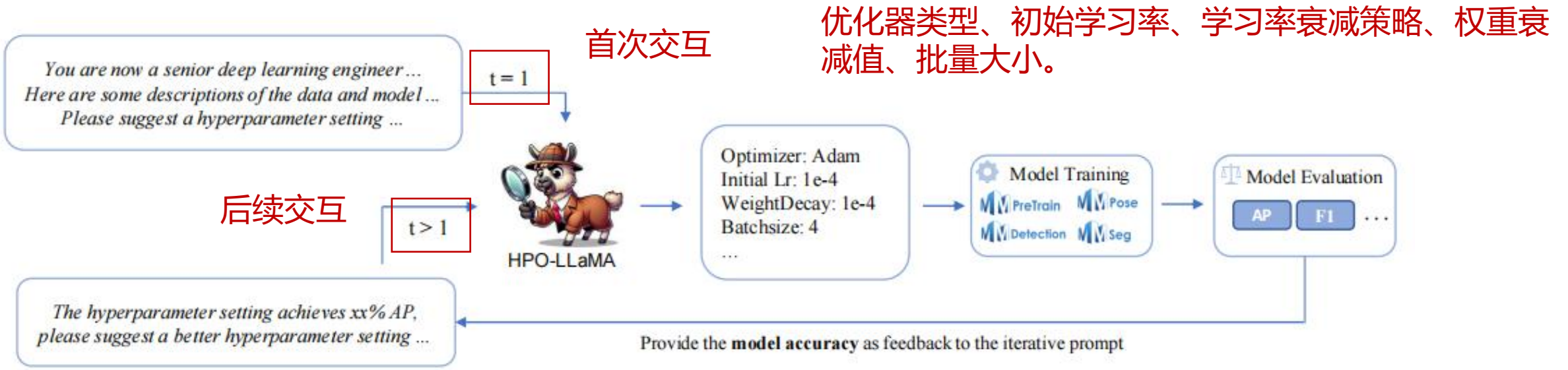


# HPO-LLaMA

优化器类型、初始学习率、学习率衰减策略、权重衰减值、批量大小。

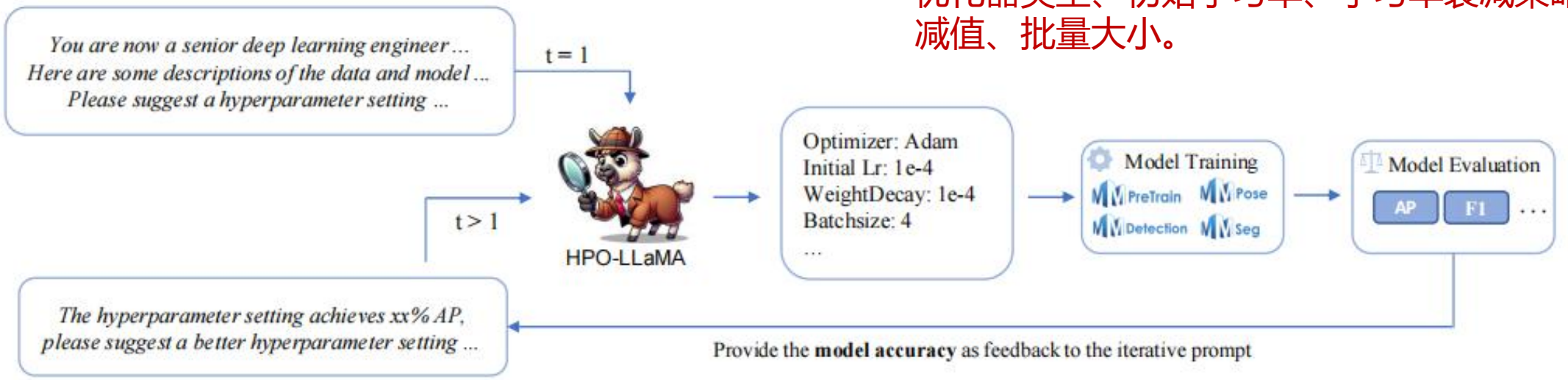


# HPO-LLaMA


You are now a senior deep learning engineer ...
Here are some descriptions of the data and model ...
Please suggest a hyperparameter setting ...
t = 1
首次交互
优化器类型、初始学习率、学习率衰减策略、权重衰减值、批量大小。
后续交互
t > 1
HPO-LLaMA
Optimizer: Adam
Initial Lr: 1e-4
WeightDecay: 1e-4
Batchsize: 4
...
Model Training
PreTrain
Pose
Detection
Seg
Model Evaluation
AP
F1
...
The hyperparameter setting achieves xx% AP, please suggest a better hyperparameter setting ...
Provide the model accuracy as feedback to the iterative prompt

# HPO-LLaMA

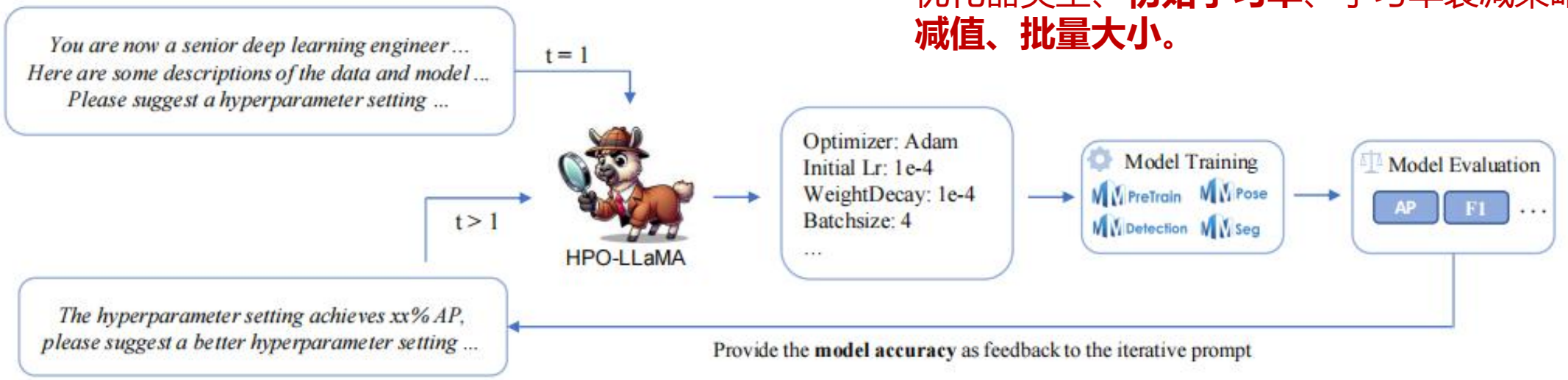
优化器类型、初始学习率、学习率衰减策略、权重衰减值、批量大小。




| | Classification | Object detection | Semantic segmentation | Keypoint detection |
|---|---|---|---|---|
| Optimizer types | SGD (Sutskever et al. 2013), Adam (Kingma and Ba 2017), AdamW (Loshchilov and Hutter 2019), RMSprop (Graves 2014) | | | |
| Learning rate decay strategies | MultiStepLR, ConsineAnnealingLR, StepLR, PolyLR | | | |
| Initial learning rate range | $[10^{-8}, 0.1]$ | | | |
| Weight decay range | $[10^{-5}, 0.1]$ | | | |
| Training iteration range | [2000,5000] | [4000,9000] | [2000,7000] | [2000,5000] |
| Batch size range | [1,64] | [1,16] | [2,8] | [2,64] |

# HPO-LLaMA

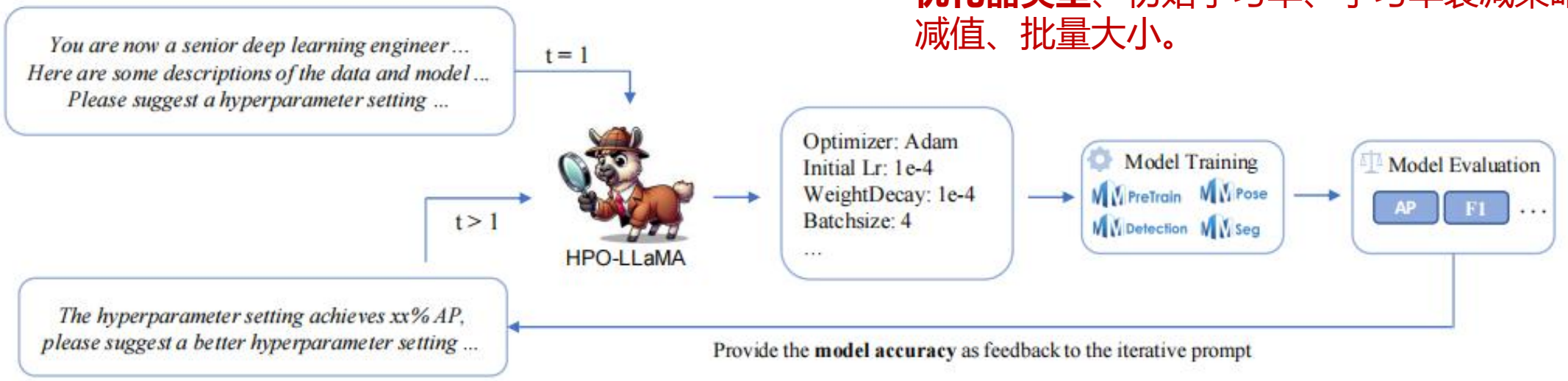
优化器类型、**初始学习率**、学习率衰减策略、**权重衰减值、批量大小。**

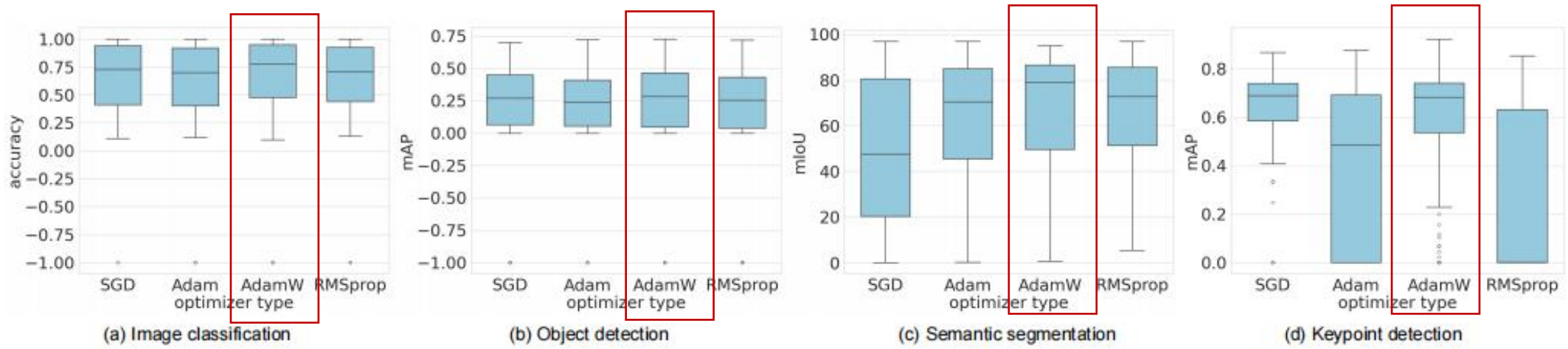



| | Classification | Object detection | Semantic segmentation | Keypoint detection |
|---|---|---|---|---|
| Initial learning rate | -0.22 | -0.49 | 0.17 | -0.43 |
| Weight decay | -0.11 | 0.01 | -0.14 | -0.07 |
| Training iteration | 0.09 | -0.01 | 0.10 | -0.01 |
| Batch size | 0.08 | 0.12 | 0.01 | -0.01 |

# HPO-LLaMA

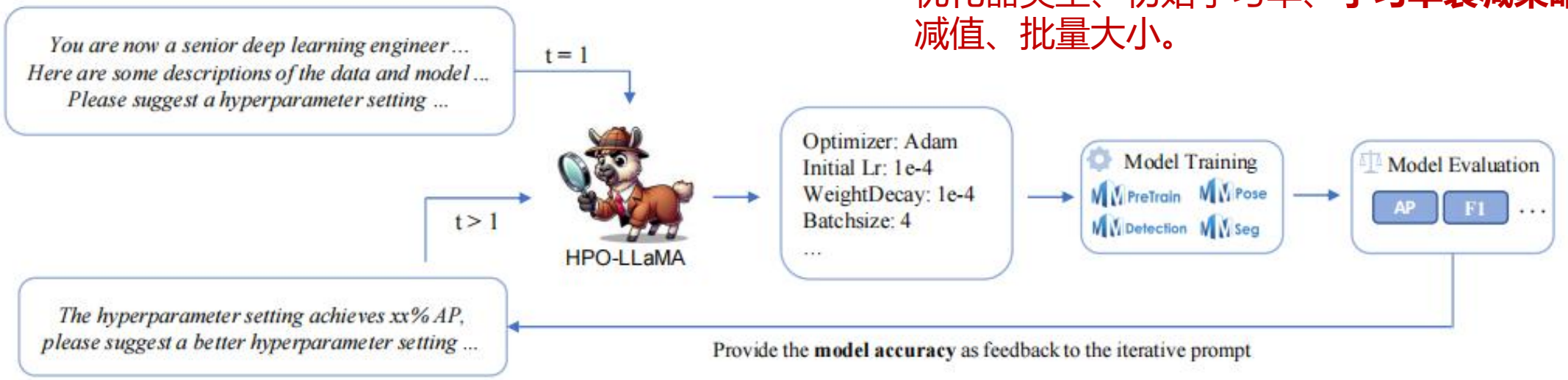
**优化器类型**、初始学习率、学习率衰减策略、权重衰减值、批量大小。

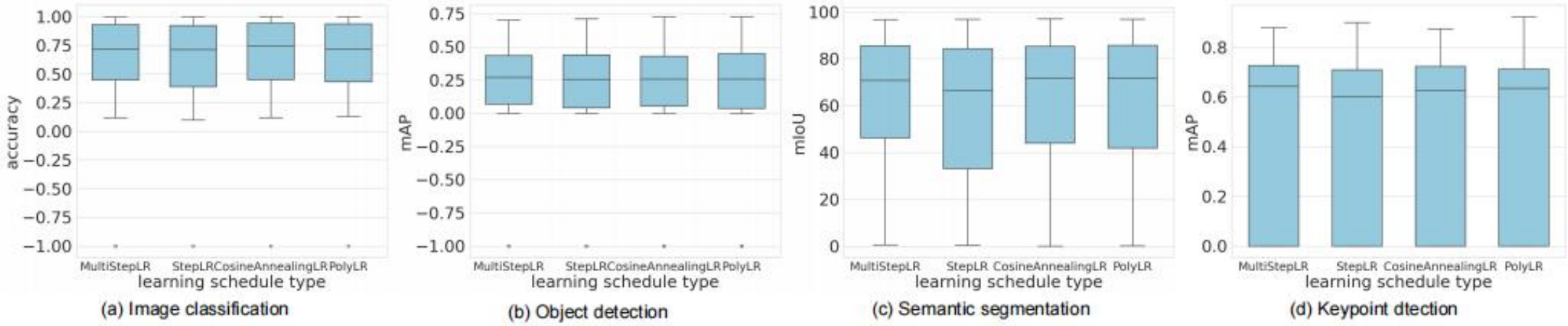


# HPO-LLaMA

优化器类型、初始学习率、**学习率衰减策略**、权重衰减值、批量大小。

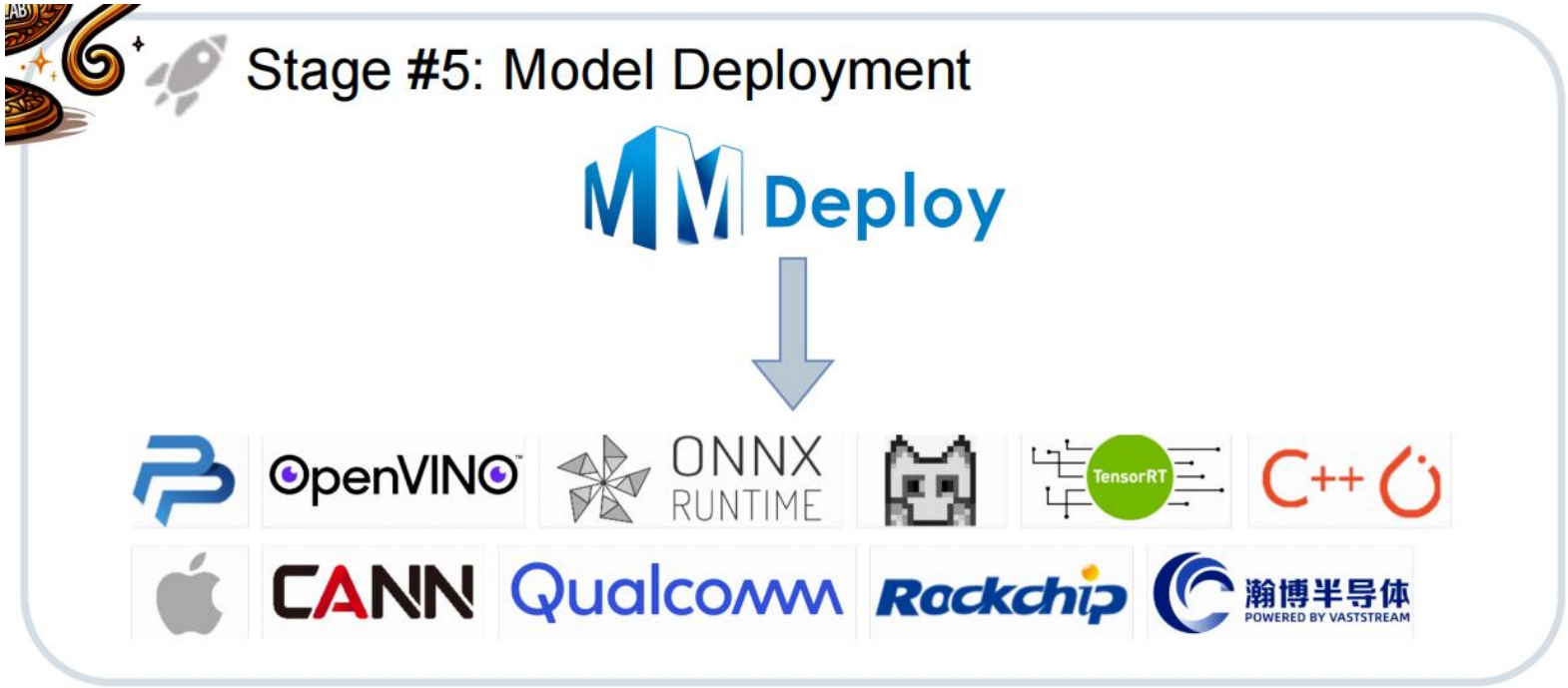



(a) Image classification (b) Object detection (c) Semantic segmentation (d) Keypoint dtection

# AutoMMLab

- 用户请求
- 请求理解
- 数据选择
- 模型选择
- 超参数优化和模型训练
- 模型部署



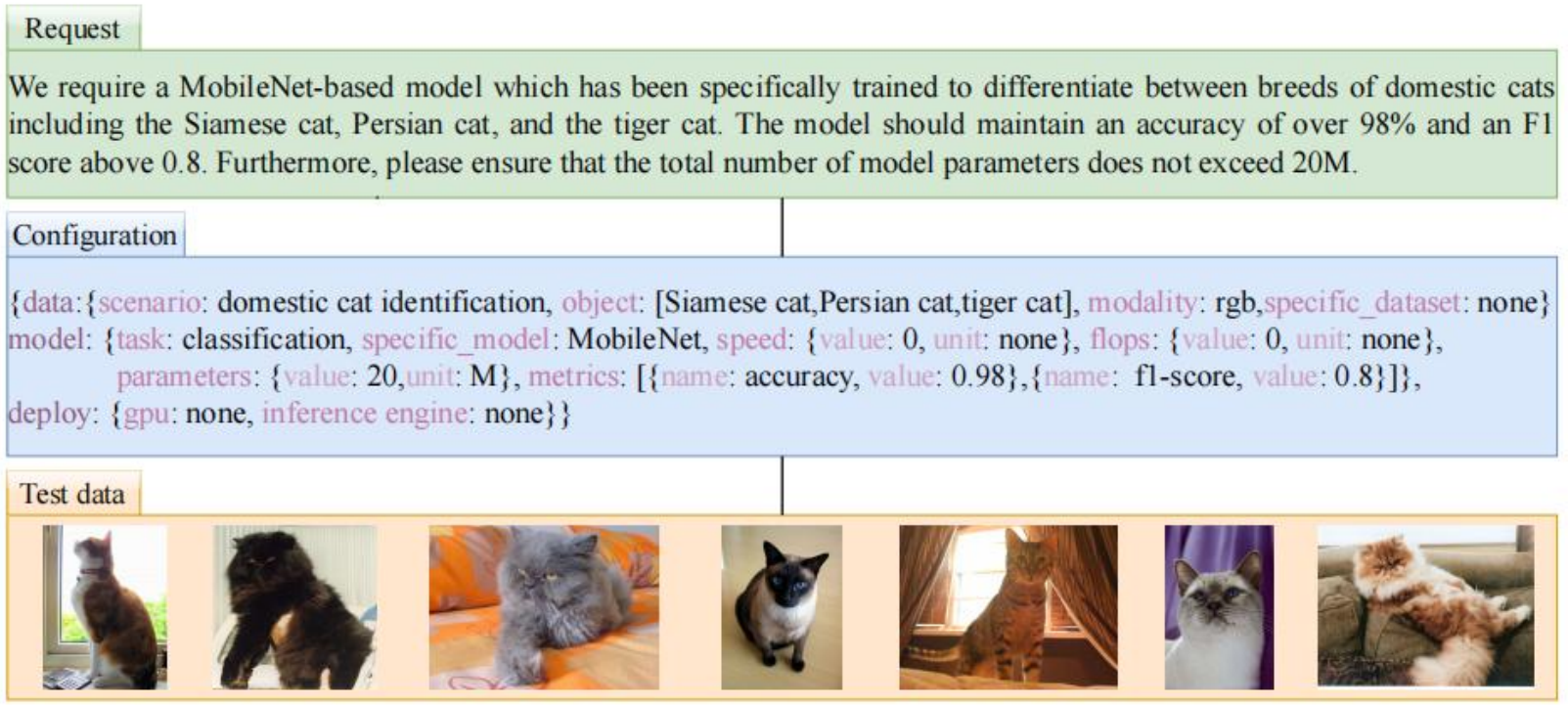

初始模型 → MMDeploy → 转换后模型

# LAMP基准

- 目前仍缺乏全面的评估标准来衡量这些模型在自动机器学习（AutoML）中的表现。

## 评估指标：

- 请求理解
- 超参数优化
- 端到端（模型质量）

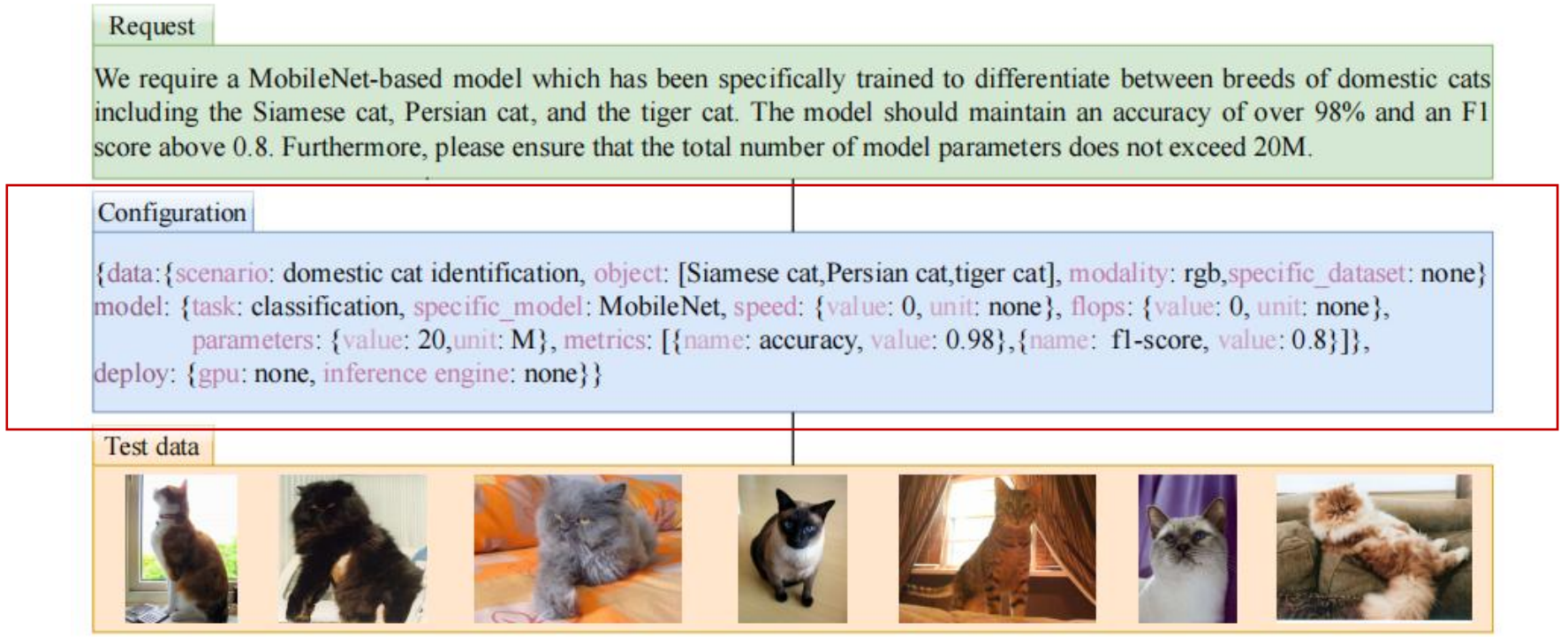


# LAMP基准

- 目前仍缺乏全面的评估标准来衡量这些模型在自动机器学习（AutoML）中的表现。

## 评估指标：

- 请求理解
- 超参数优化
- 端到端（模型质量）



# LAMP基准

- 目前仍缺乏全面的评估标准来衡量这些模型在自动机器学习（AutoML）中的表现。

评估指标：

- 请求理解
- 超参数优化
- 端到端（模型质量）

- 键级准确率：计算每个键值对的平均解析准确度；
  - 项目类型对
  - 列表类型对
- 请求级准确率：评估对整个请求的理解准确度。

# LAMP基准

- 目前仍缺乏全面的评估标准来衡量这些模型在自动机器学习（AutoML）中的表现。

评估指标：

- 请求理解
- 超参数优化
- 端到端（模型质量）

- 图像分类任务：采用最高准确率（top-1）；
- 目标检测任务：报告标准平均精度（mAP）；
- 语义分割任务：采用平均交并比（mIOU）；
- 关键点估计任务：则使用基于目标关键点相似性（OKS）的mAP。
- 同时计算四个任务的平均值与标准差，形成综合评估体系。

# LAMP基准

- 目前仍缺乏全面的评估标准来衡量这些模型在自动机器学习（AutoML）中的表现。

评估指标：

- 请求理解
- 超参数优化
- 端到端（模型质量）

  - F：表示完全失败，得分为0分，平台未能生成功能模型；
  - W：表示可运行模型，得1分，模型可运行但可能未完全满足所有用户需求（如准确率不足）；
  - P：表示完美模型，得2分，模型完全符合用户预期。
  - 评估包含四种任务类型，每项任务满分40分，总分160分。

# 提纲

# 实验设置

- 基线模型
  - 经过提示词指令微调的LLlama2-7B-Chat（Metat提出的微调的私有大模型），这些提示词包含指令说明及预期行为的示范示例（可选）。
  - GPT- 3.5-turbo （OpenAI提出的通用大模型）
  - GPT- 4（OpenAI提出的通用大模型）
  - PaLM2（谷歌提出的通用大模型）
  - 本文是基于LLlama2-7B微调的
- 基线算法
  - Random：随机采样
  - BayesianRF：基于随机森林的贝叶斯优化
  - BayesianGP：基于高斯过程的贝叶斯优化
- 主要实验结果
  - 提出的LAMP基准的所有指标

RU-LlaMA和HPO-LlaMA的微调：

- 8块NVIDIA Tesla A100 GPU上完成，每块显卡配备80GB显存。
- 采用AdamW优化器训练3个周期。初始学习率设为1e-4，每个周期后衰减0.2倍。文本最大长度设置为4096，批量大小设为4。在LoRA中将秩参数设为8，alpha值设为32。

AutoMMLab测试：

- 8块NVIDIA Tesla V100 GPU上进行，每块显存为32GB。

# 请求理解-评估

**Request**

We require a MobileNet-based model which has been specifically trained to differentiate between breeds of domestic cats including the Siamese cat, Persian cat, and the tiger cat. The model should maintain an accuracy of over 98% and an F1 score above 0.8. Furthermore, please ensure that the total number of model parameters does not exceed 20M.

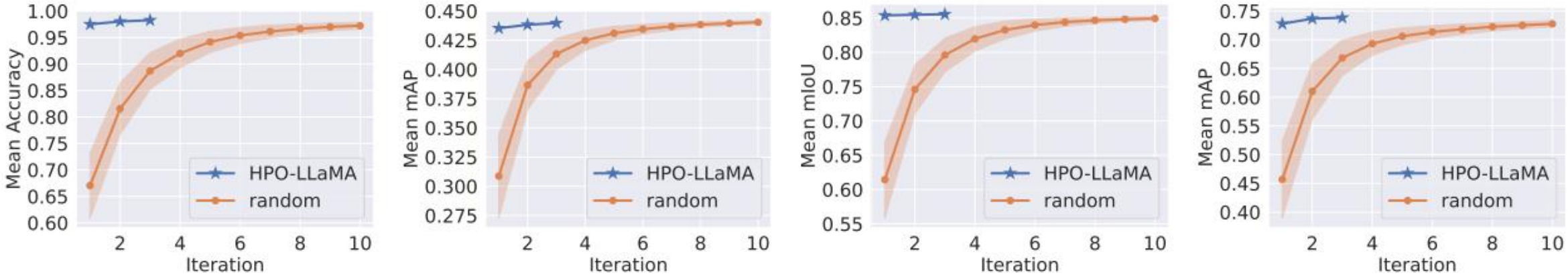
**Configuration**


{data:{scenario: domestic cat identification, object: [Siamese cat,Persian cat,tiger cat], modality: rgb,specific_dataset: none}
model: {task: classification, specific_model: MobileNet, speed: {value: 0, unit: none}, flops: {value: 0, unit: none},
parameters: {value: 20,unit: M}, metrics: [{name: accuracy, value: 0.98},{name: f1-score, value: 0.8}]},
deploy: {gpu: none, inference engine: none}}


| Model | Key-Level | | | Req-Level |
|---|---|---|---|---|
| | Item | List | Total | |
| LLaMA2-7B-Chat | 85.71 | 50.00 | 77.78 | 0 |
| PaLM2 | 96.79 | 88.13 | 94.86 | 63.75 |
| GPT-3.5-turbo | 96.43 | 95.63 | 96.25 | 72.50 |
| GPT-4 | 97.50 | 93.13 | 96.53 | 80.00 |
| RU-LLaMA | **98.57** | **96.88** | **98.20** | **86.25** |

**效果最优**

# 超参数优化-评估




(a) Image Classification (b) Object Detection (c) Semantic Segmentation (d) Keypoint Detection

| Model | #R | Cls. | Det. | Seg. | Kpt. |
|---|---|---|---|---|---|
| BayesianRF | 5 | 0.618±0.287 | 0.291±0.211 | 0.847±0.044 | 0.069±0.136 |
| BayesianGP | 5 | 0.761±0.264 | 0.280±0.208 | 0.848±0.041 | 0.081±0.150 |
| LLaMA2-7B | 1 | 0.839±0.213 | 0.128±0.164 | 0.291±0.409 | 0±0 |
| PaLM2 | 1 | 0.964±0.056 | 0.367±0.196 | 0.845±0.067 | 0.719±0.079 |
| GPT-3.5-turbo | 1 | 0.849±0.214 | 0.364±0.194 | 0.852±0.044 | 0.204±0.160 |
| GPT-4 | 1 | 0.861±0.188 | 0.434±0.147 | 0.803±0.194 | 0.096±0.158 |
| HPO-LLaMA | 1 | 0.975±0.028 | 0.435±0.148 | 0.854±0.042 | 0.728±0.051 |
| HPO-LLaMA | 3 | **0.983±0.020** | **0.440±0.150** | **0.856±0.043** | **0.738±0.053** |

**效果最优，并且在开始时就具备相对较好的结果**

# 端到端（模型质量）-评估

| RU | HPO | Cls. | Det. | Seg. | Kpt. | Total |
|---|---|---|---|---|---|---|
| LLaMA2-7B | LLaMA2-7B | 0 | 0 | 0 | 0 | 0 |
| PaLM2 | PaLM2 | 14 | 25 | 27 | 15 | 81 |
| GPT-3.5-turbo | GPT-3.5-turbo | 24 | 24 | 25 | 11 | 84 |
| GPT-4 | GPT-4 | 17 | 27 | 29 | 14 | 87 |
| RU-LLaMA | HPO-LLaMA | **31** | **31** | **32** | **18** | **112** |

**效果最优**

该评估采用从“0”到“2”的评分系统，其中“0”表示“完全失败”，“1”表示“可运行模型”，2表示“完美模型”。每个任务的满分是40分，总分满分是160分。

# 总结与思考

# 总结

- **框架**：提出一种创新的自动机器学习系统，通过利用语言模型（LLM）的强大功能，在自然语言指令引导下实现计算机视觉任务的全流程自动化开发。
- **技术**：提出了HPOLLaMA这一基于LLM的新型超参数优化方法。
- **领域**：推出了LAMP（语言指导自动化模型生产）基准测试，用于评估端到端提示生成模型的性能。
- **实验**：通过在LAMP平台上测试多种语言模型，验证了RU-LLaMA和HPO-LLaMA的优越性。

# 思考

## 1. 这个paper有什么问题，基于这个paper还能做什么?

- 当前使用的模型及其属性还是基于通用模型库和静态属性（比如Speed(fps): 174.42），但是模型运行在不同设备上的Speed(fps): 174.42只有真正测试后才能得知，因此仅通过静态属性选择的模型可能无法满足用户部署在真实设备的需求。

## 2. 这个paper提到的idea，能不能用在自己的方向/project上面?

- 我们可以基于这个架构，构建一套面向硬件感知的模型架构生成方案。实现一个面向硬件感知的模型全流程自动化开发。

## 3. 这个paper能不能泛化，需要较为熟悉这个小方向?

- 这篇论文主要面向CV领域，那我们可以结合huggingface实现面向NLP等领域的全自动化，并根据NLP领域与CV领域的差异性，比如评估方式之类的构建新的架构。

# 感谢各位老师和同学！
# 请大家提出宝贵意见！