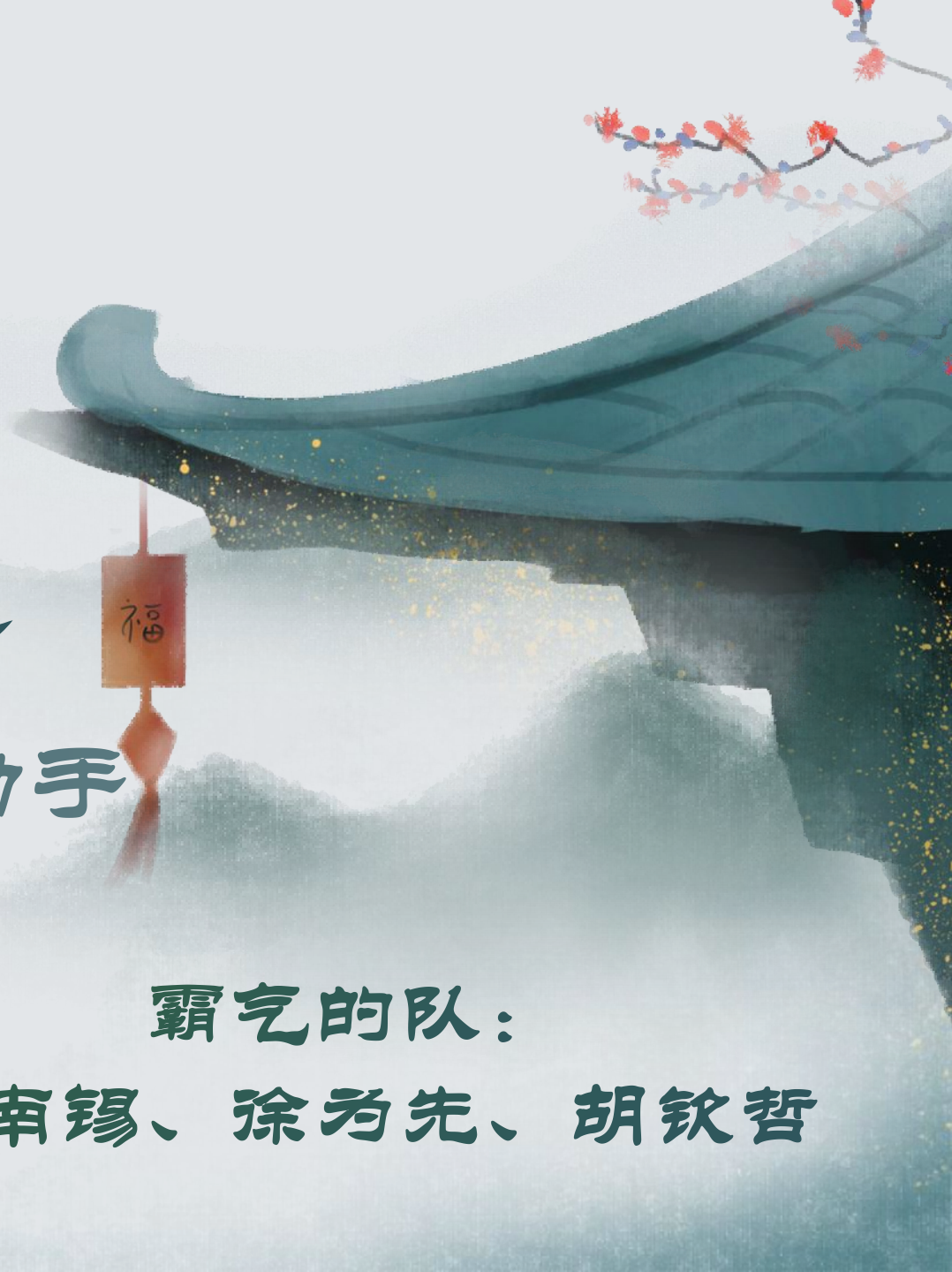
SEEKING THE CLASSICS

ALGORITHM COMPREHEND HERITAGE

问·道

您的私人文言文学习助手



霸气的队：

李南锡、徐为先、胡钦哲

壹 项目设计

贰 技术实现

叁 应用使用

肆 应用场景

伍 未来展望

陆 开发体会

项目设计

问·道是一款面向广大中小学生开发的私人文言文学习助手，旨在帮助中小学生更好掌握文言文知识。

利用BigDL-LLM轻量级开源大模型，收录易文言网、古诗文网等网站的文言词汇语料，通过对话问答的方式帮助用户更好掌握文言知识。

针对用户提出的问题，问·道会作出实时回答，提供精准文言翻译与重点字词的讲解，使得用户不再需要面对不可靠的翻译软件，同时能够更有针对性地进行文言文知识的学习，真正做到成为用户的私人学习助手，实现无痛学习文言文。

技术实现

*ChatGLM2-int4*模型压缩

BF16 到INT4

模型特点

采用4位量化技术，显著减少模型大小（由13GB到5.5GB），占用更少存储空间与计算资源。

应用场景

适用于资源受限的环境，如移动设备、嵌入式设备等。

性能优势

在保持较高性能(精度仅下降2%)的同时，降低对硬件资源的需求。

数据收集：使用爬虫从易文网、古诗词网等网站爬取数千条重点实词、虚词的所有释义与典型例句。

模型使用：先使用ChatGLM2-INT4模型来进行上下文对话，而后对齐两个模型的颗粒度，使用wenyanwen-chinese-translate-to-ancient模型进行文言文翻译。此外，我们还撷取重点字词进行讲解，以提高学习效果。

量化等级	编码 2048 长度的最小显存	生成 8192 长度的最小显存
FP16 / BF16	13.1 GB	12.8 GB
INT8	8.2 GB	8.1 GB
INT4	5.5 GB	5.1 GB

量化等级	Accuracy (MMLU)	Accuracy (C-Eval dev)
BF16	45.47	53.57
INT4	43.13	50.30

应用使用

```
from transformers import (
  EncoderDecoderModel,
  AutoTokenizer
)
PRETRAINED = "raynardj/wenyanwen-chinese-translate-to-ancient"
tokenizer = AutoTokenizer.from_pretrained(PRETRAINED)
model = EncoderDecoderModel.from_pretrained(PRETRAINED)
```

```
from transformers import AutoTokenizer, AutoModel
tokenizer = AutoTokenizer.from_pretrained("THUDM/chatglm3-6b",
trust_remote_code=True)
model = AutoModel.from_pretrained("THUDM/chatglm3-6b",
trust_remote_code=True).half().cuda()
model = model.eval()
response, history = model.chat(tokenizer, "你好", history=[])
print(response)
response, history = model.chat(tokenizer, "晚上睡不着应该怎么办",
history=history)
print(response)
```

应用场景

01

文言文阅读辅助

提供文言文阅读材料，辅助学生理解古代文献，提高文言文阅读能力。

02

文言文写作指导

提供文言文写作指导，帮助学生掌握文言文写作技巧，提升文言文写作能力。

03

文言文考试备考

提供文言文考试备考资料，帮助学生备考文言文考试，提高考试成绩。

未来展望

我们希望能够为问·道加入更多功能。

学习进度跟踪

通过用户行为分析，实时跟踪学习进度，为用户提供精准的学习建议。

小型终端部署

将软件部署到学习机等小型终端，让用户无需面对互联网的诱惑即可畅学文言文。

开发体会