AI 学术混码地道性 (Naturalness) 评分细则

Evaluation Guidelines for Annotators

1. 基本原则 (Fundamental Rules)

- **语感优先 (Style > Factuality)：非必要无需查百科/词典，可以按语感来！**

  只要术语（如 SFT, DPO）看起来像真的且读起来顺口就可以大概按着感觉给，但是如果觉得别扭，可以去搜索（注意，不要问任何 AI，就看词条或有无对应的真实博客、技术文档等等）查它的准确含义，若含义不符酌情扣 1 到 2 分。

  查词是为了确认这个词‘存在’，而不是为了搞懂它背后的数学原理！搜索方法：直接在 Google/Bing 搜术语 + 中文关键词。

  例子：句子出现了"DPO"。搜：DPO 强化学习或 DPO 大模型。

  判断：如果搜出来满屏都是相关的技术博客 -> 说明词是对的，不用看内容，直接回来看句子通不通顺，通顺就给高分。如果搜出来啥也没有，或者全是无关内容 -> 大概率是瞎编的，扣分。

  限时：每个词搜索不要超过 30 秒，没搜到就当它是错的。

- **全英文惩罚 (English Penalty)**：如果模型无视中文语境输出纯英文，直接打 **0 分**。
- **允许半分 (Half-points)**：纠结时请使用 **3.5** 或 **4.5**，鼓励细粒度区分。

2. 评分量表 (Scoring Rubric)

分数	等级	直观感受 (Intuition)	判定特征 (Criteria)
4—5	专家级	“这就是像专家平时说的话。” 极度顺滑，高效，无废话。	• **高频缩写**：直接用 LLM, SFT。 • **拒绝定义**：默认对方懂，不解释概念。 • **丝滑嵌入**：中英融合完美。
3—4	地道	“写得还不错，但是像学生作业。” 略显书面，不够老练。	• **使用全称**：写 Large Language Model。 • **语调正式**：语法无误，但稍显啰嗦。
2—3	路人感	“像是评论区随意文字。” 能看懂，但感觉作者是外行。	• **过度解释**：“AI (人工智能)...” • **生硬嵌字**：“这个 model 很 good”。
1—2	生硬	“这是机翻的吧?” 读起来卡顿，难受。	• **强行汉化**：把 Transformer 叫“变换器”。 • **搭配错误**：英文词放错位置。
0—1	崩塌	“完全没在说人话。”	• **全英文回复**。 • **严重幻觉** / 胡言乱语。