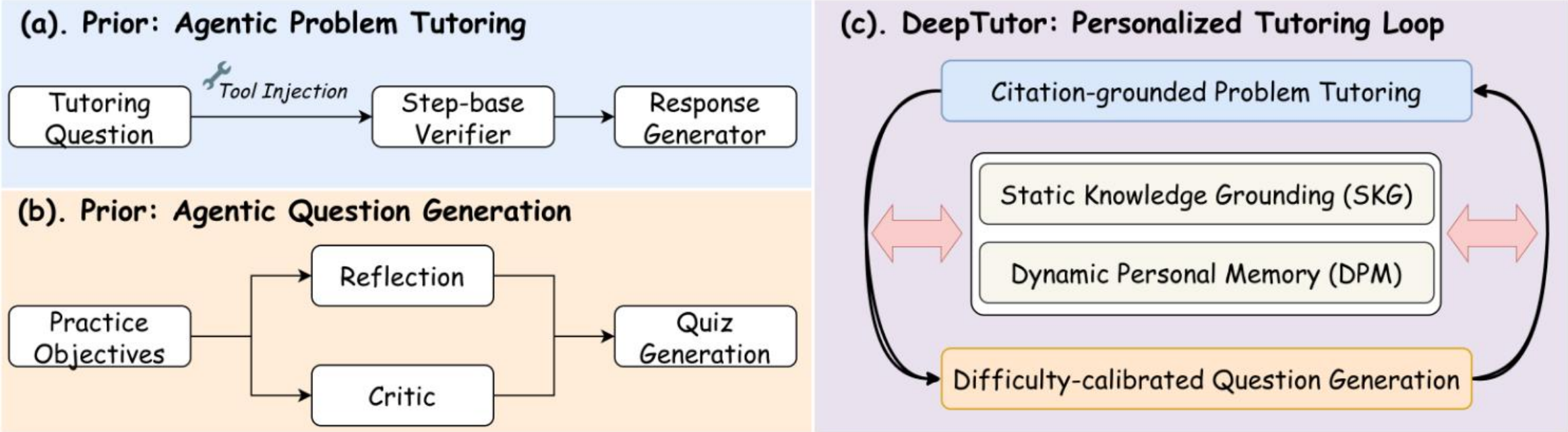


DeepTutor: Towards Agentic Personalized Tutoring

The University of Hong Kong 2026

Background



Problem Tutoring已从直接生成答案发展为：

- 评估开放式教学行为 (Pedagogical Behavior)
- 诊断学生推理错误 (Reasoning Error Diagnosis)
- 引导学生完成问题求解 (Guided Problem Solving)

局限： 诊断结果通常仅作用于当前辅导轮次。

Question Generation

- 利用规划 (Planning) 或自我优化 (Self-refinement) 生成练习题。
- 控制题目内容、难度和教学目标。

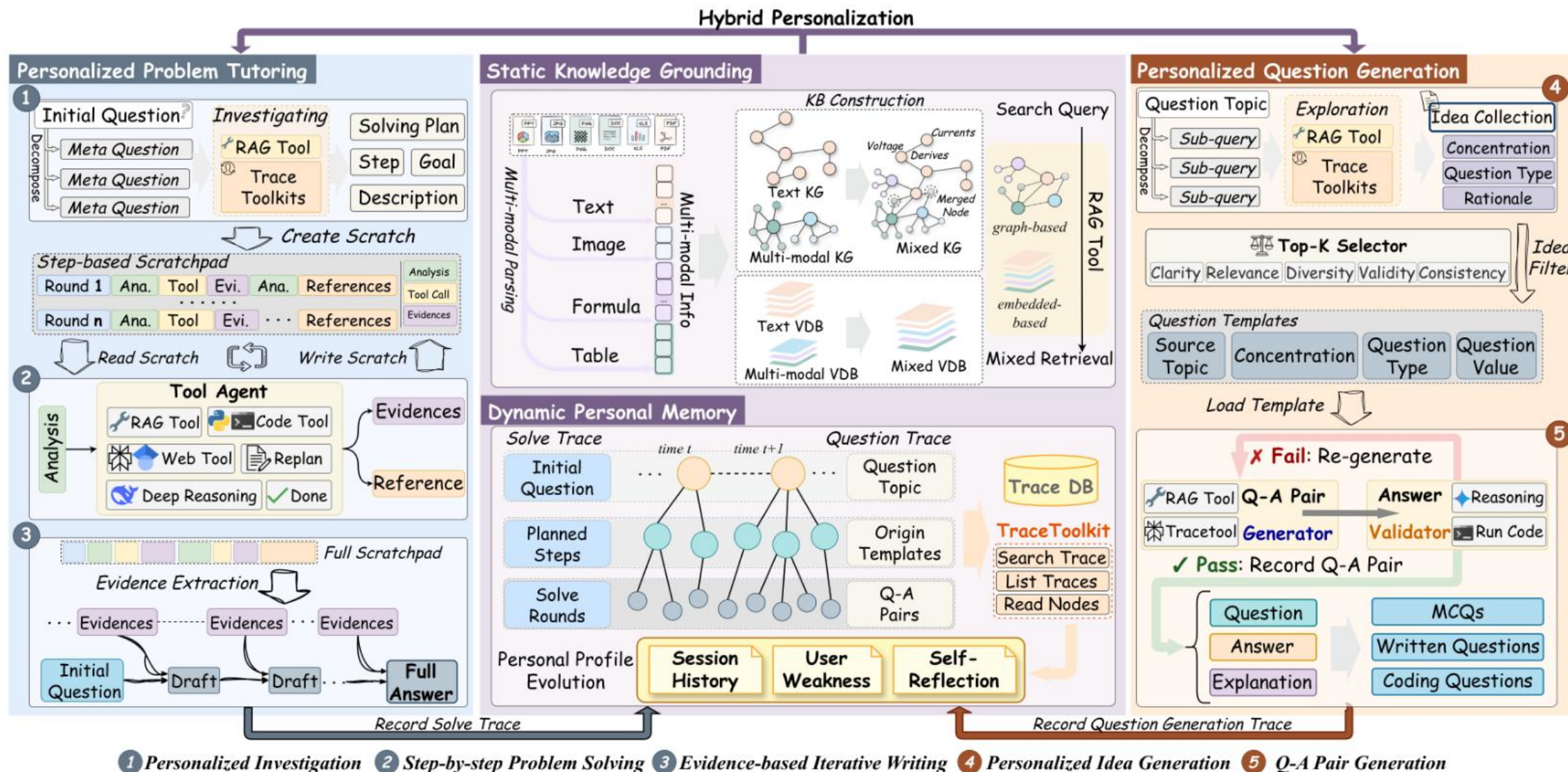
局限： 生成过程主要依赖预设目标，而非学生近期学习轨迹。

Common Limitation: Tutoring Trace 不关联下一道练习题, Practice Outcome 不更新后续教学解释。缺乏细粒度、持续演化的 Learner Model。

Hybrid Personalization Engine

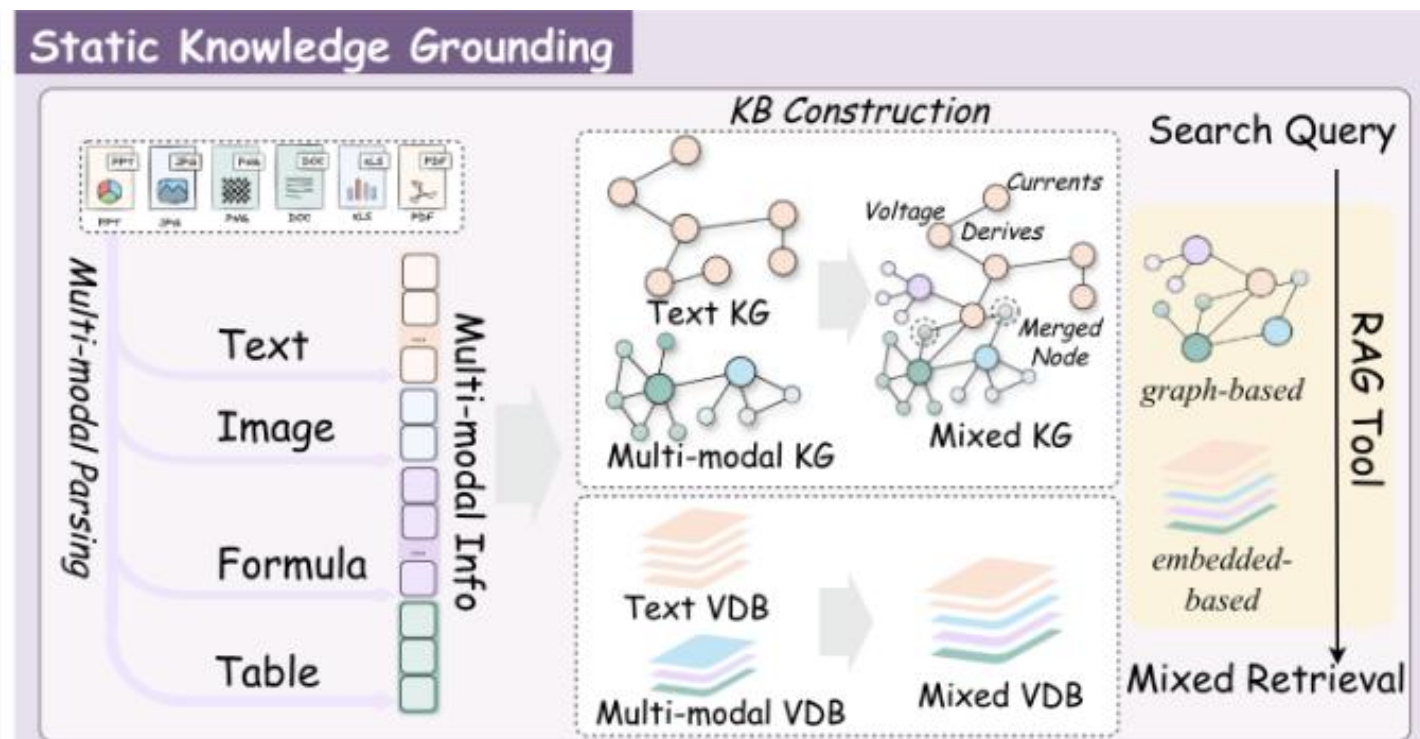


南京航空航天大学
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS



通过共享学习者画像与课程知识，同时动态更新个性化上下文，实现 Tutoring-Generation 的闭环协同。

Hybrid Personalization Engine



构建课程知识库（SKG）：

- 将课程文档拆分为原子知识单元（Atomic Content Units）
- 构建双索引结构：
 - Knowledge Graph (G)：组织知识结构关系
 - Embedding Index (B)：支持语义相似检索

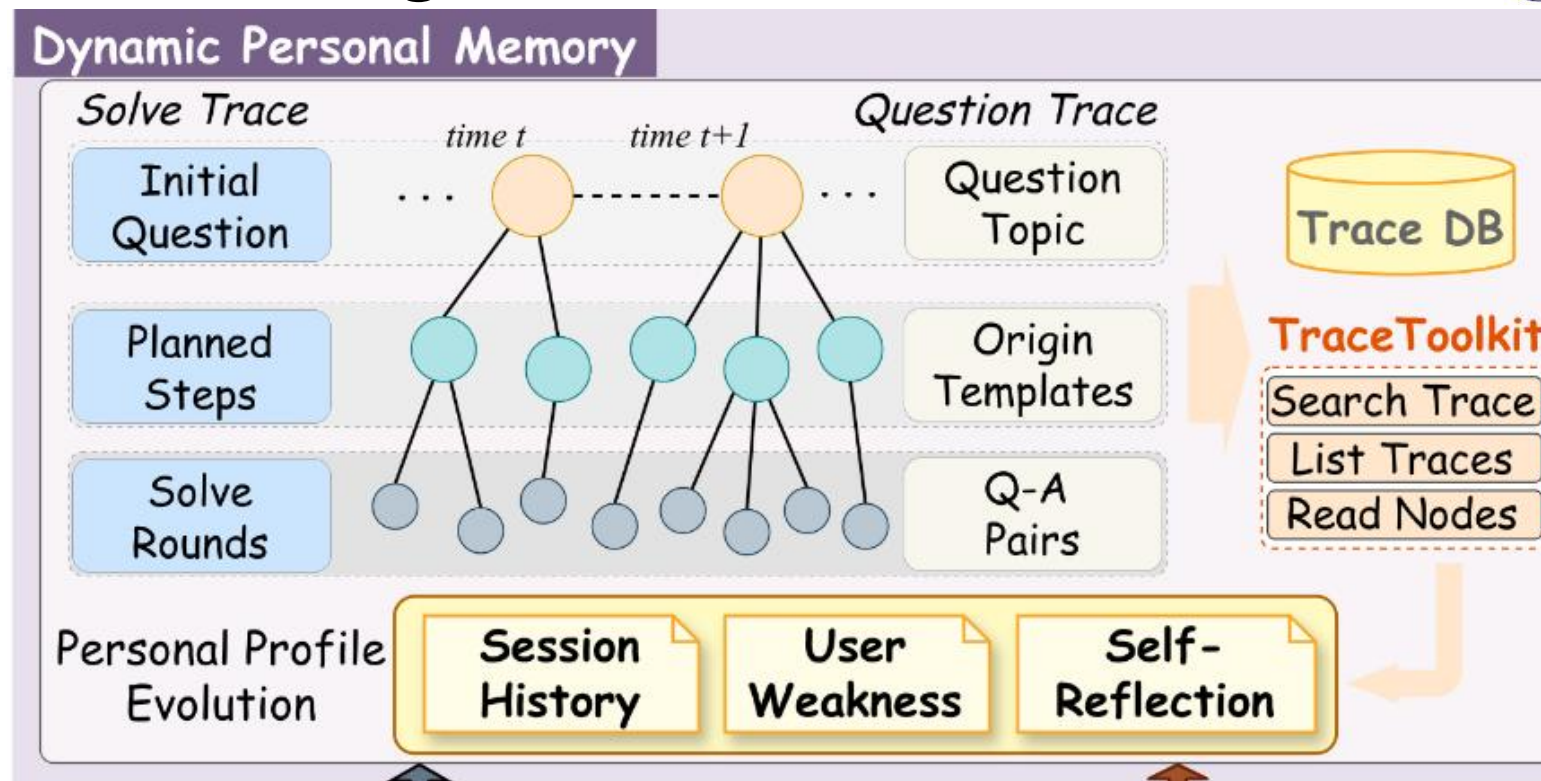
查询阶段：

- 图检索获取显式关联知识
- 向量检索获取语义相似内容
- 融合检索结果，形成课程知识上下文 C_{rag}

Hybrid Personalization Engine



南京航空航天大学
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS



Dynamic Personal Memory (DPM)

目标：记录学生如何学习

核心组成：

Trace Forest (F)

保存完整学习交互轨迹
态

支持历史轨迹检索

画像

Learner Profile (D)

总结学生学习状

持续更新个性化

Hybrid Personalization Engine



Trace Forest (F):

- 三级层次组织学习轨迹：
 - Level 1: 保存会话级输入与全局总结 (Session-level Input & Summary)
 - Level 2: 保存任务分解产生的中间规划单元 (Planning Units)
 - Level 3: 保存工具输出、证据及验证结果等细粒度执行记录 (Execution Records)
- TraceToolkit 提供三种访问方式：
 - SearchTrace: 语义检索历史轨迹
 - ListTraces: 按时间、任务、主题筛选轨迹
 - ReadNodes: 获取完整节点内容及祖先路径

Learner Profile (D):

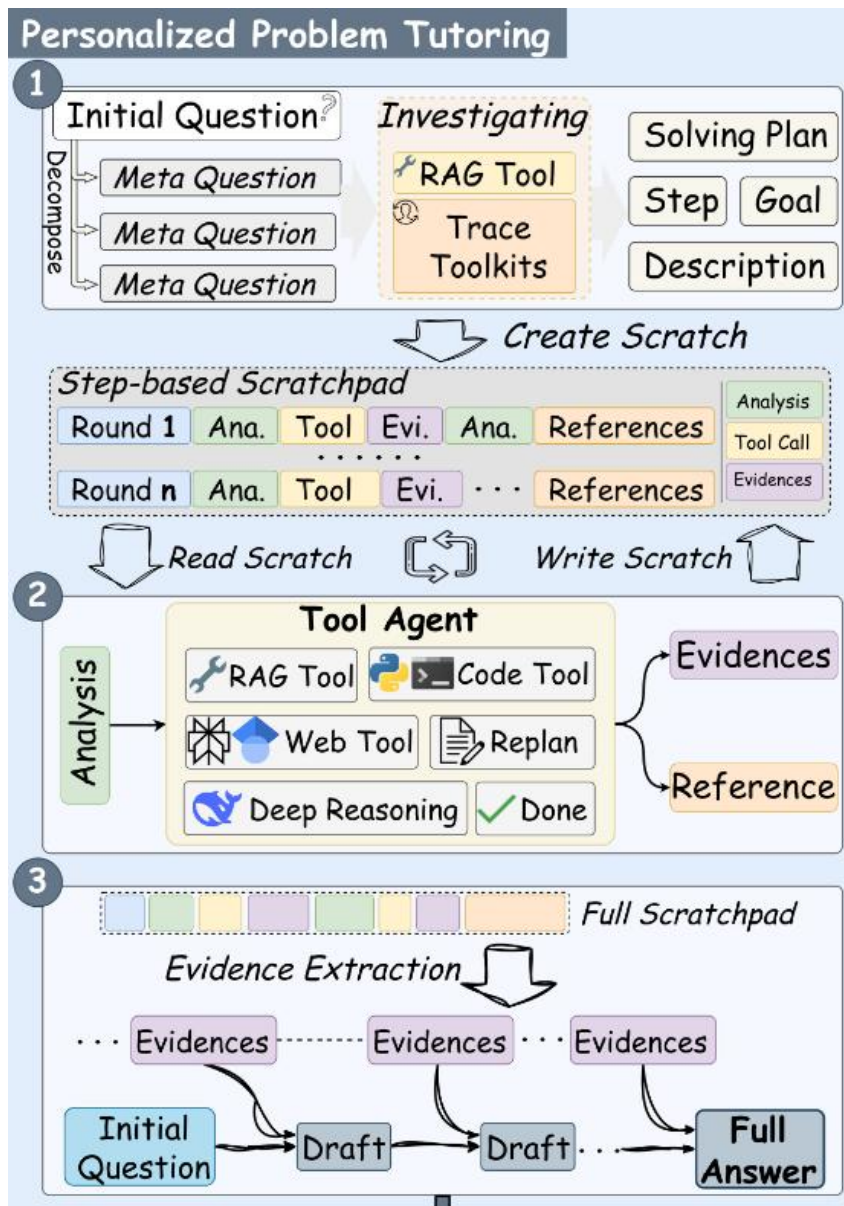
由 Memory Agent 基于 Trace Forest 持续更新:

- D_s: 学习历史 (Topics、Solving Paths、Performance Trends)
- D_w: 知识薄弱点 (Confusions、Wrong Patterns、Knowledge Gaps)
- D_r: 教学反思 (Pedagogical Self-reflections)

结合两个渠道构建个性化上下文 C_mem:

- 主动跟踪检索, 其中 TraceToolkit 呈现最相关的节点, 并按比例跨级别分配预算
- 特定于角色的画像摘录, 根据其功能将 D 的不同部分路由到不同的代理

Hybrid Personalization Engine



Personalized Problem Tutoring

将传统单轮推理拆分为三个阶段：Investigate → Solve → Write

Stage ① Personalized Investigation (个性化调查)

- 将问题分解为多个子目标
- 检索课程知识 C_{rag} 与学习者上下文 C_{mem}
- 制定针对学生知识薄弱点的教学计划

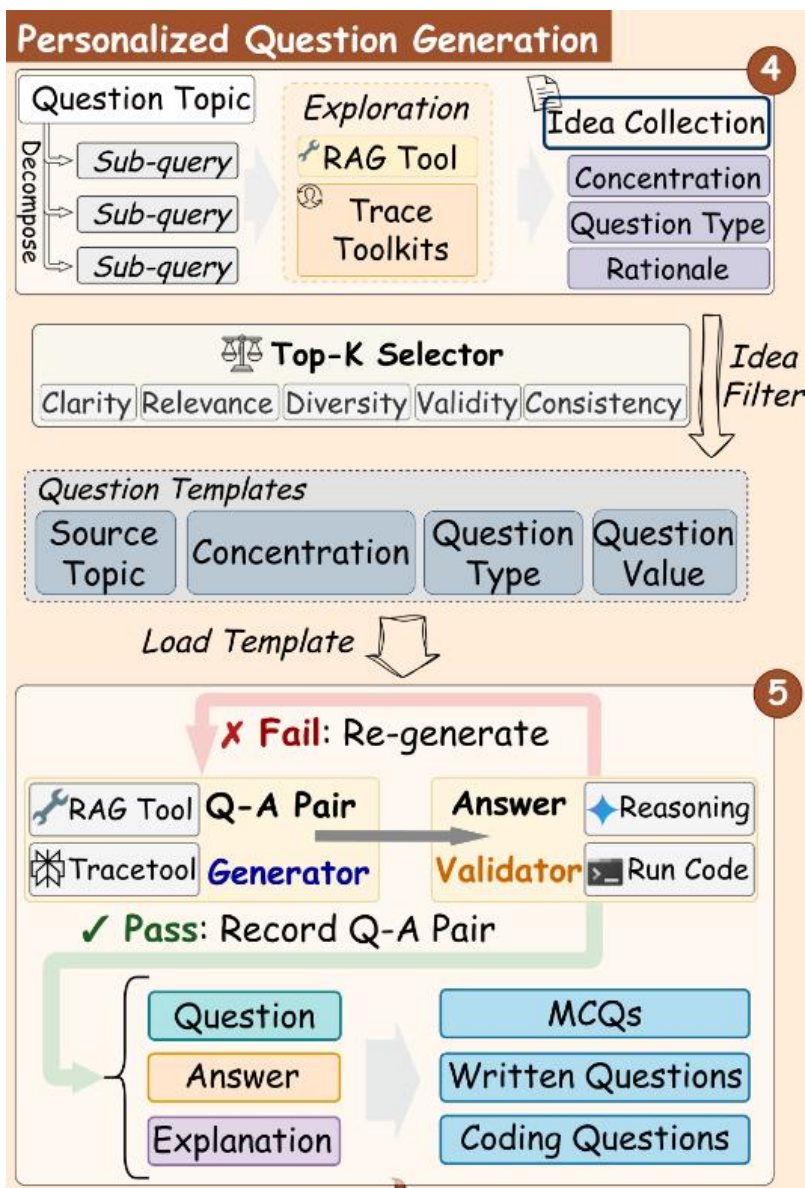
Stage ② Step-by-step Guided Solving (分步引导求解)

- 按子目标逐步完成推理
- 利用 Self-notes 与分层压缩上下文
- 支持 Adaptive Replanning 动态调整后续计划

Stage ③ Evidence-based Iterative Writing (循证迭代生成)

- 基于检索证据组织答案
- 根据 Learner Profile 调整讲解深度
- 输出带引用依据的个性化教学内容

Hybrid Personalization Engine



Personalized Question Generation

Stage ① Personalized Idea Generation:

- 结合 C_rag 与 C_mem 产生一系列候选创意，每个创意都指定一个目标概念、问题格式（MCQ、书面回答或编码）以及基于诊断差距的个性化基本原理
- 经过评估迭代修建筛选，生成结构化题目模板

Stage ② Critic-guided Generation:

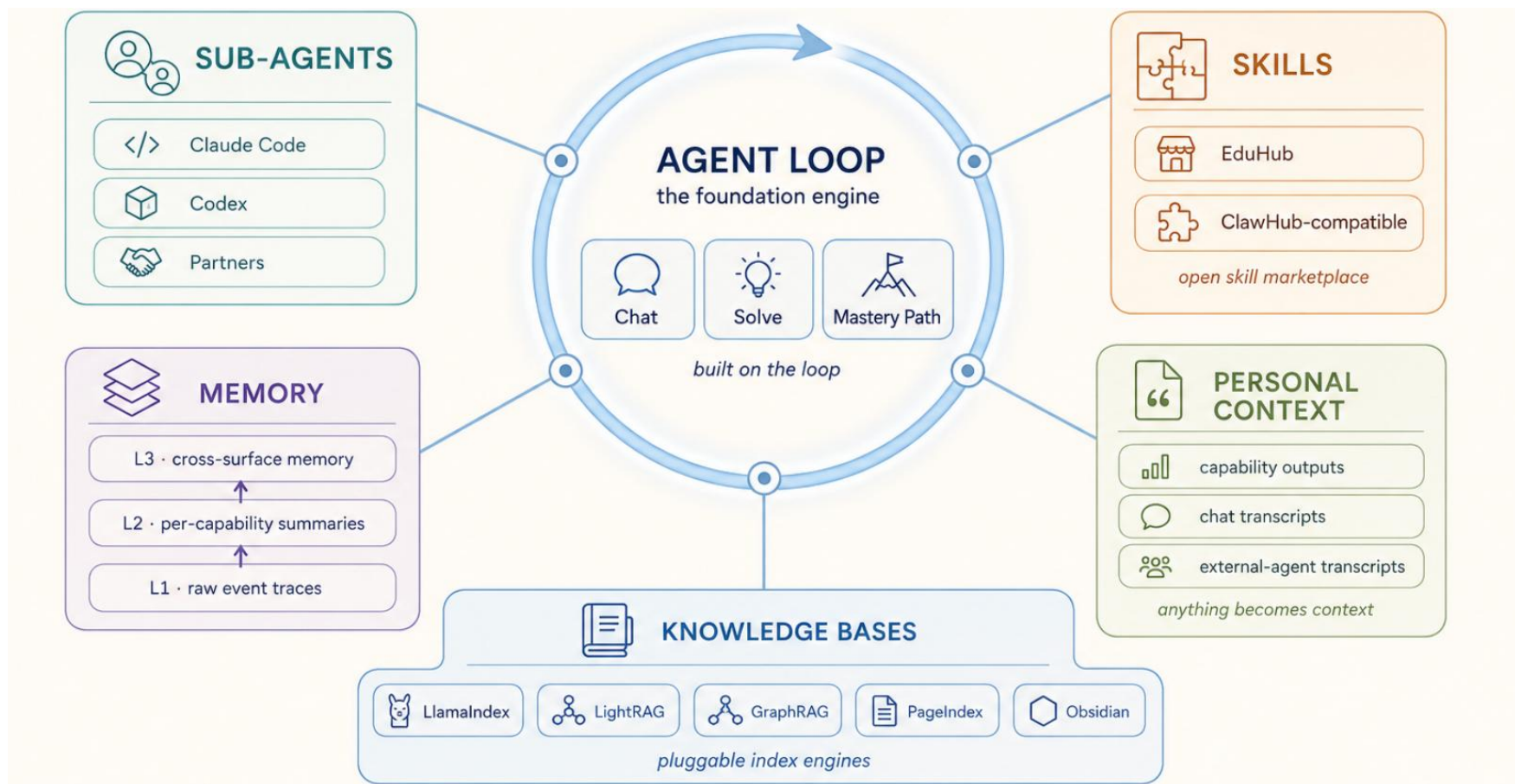
- 根据模板生成 题目-答案-解析 三元组
- 独立 Validator 验证：
 - 事实正确性（Factual Correctness）
 - 教学合理性（Pedagogical Soundness）

未通过验证则反馈并重新生成

Closed-loop Tutoring Cycle:

- 每轮交互更新 Trace Forest 与 Learner Profile
- Problem Tutoring ↔ Question Generation 双向协同
- 学生学习表现持续驱动后续教学与习题生成

Extending the Personalization Substrate



- Hybrid Personalization Engine 作为统一底座（Reusable Substrate）
- 各类学习能力共享 Trace Forest、Learner Profile 与课程知识
- 新功能无需重新构建 Learner Model，实现持续个性化学习

Extending the Personalization Substrate



Level	Instantiation	Relation to Core Tutoring	Reused Substrate
Capability runtime	Deep Research	First-class capability for open-ended knowledge synthesis, parallel to Deep Solve.	Query decomposition; KB/web/paper search; tool execution; report synthesis.
Capability runtime	Visualize / Math Animator	First-class representation capability for diagrams, charts, animations, and demos.	Dialogue context; visual-code generation and review; rendering loop.
Capability runtime	Subagent (My Agents)	Consults a live external agent, or imports its transcripts, without reimplementing its expertise.	Turn context; <code>consult_subagent</code> tool (bounded rounds); imported transcripts as context.
Learning surface	Mastery Path	Long-horizon practice plan that sequences and reschedules review across sessions.	Learner profile and weak concepts; scheduling; grading; question generation.
Learning surface	Co-Writer	Document-editing page with embedded AI writing actions, not another chat capability.	Markdown workspace; selected spans; optional KB/web grounding; rewrite/expand/shorten with

DeepTutor 将个性化教学扩展到完整学习流程：

- 运行时能力扩展（Capability Runtime）：
 - Deep Research
 - Visualize
 - Subagent Consultation
- 学习扩展（Learning Surfaces）：
 - Co-Writer
 - Mastery Path

所有能力共享同一个 Hybrid Personalization Engine

Extending the Personalization Substrate



13

🏠 主页

💞 伙伴

🧠 我的智能体

✍️ 智能写作

📖 书籍

📚 学习空间

最近

☆ DeepTutor源码拆解...

📄 论文PPT制作

💡 AI记忆模式解析

🧠 记忆

📖 知识中心

⚙️ 设置

v1.5.1

📖 🗨️

新对话



💬 聊天
灵活对话，可使用任意工具

✍️ 测验
自动校验的题目生成

🔬 研究
多智能体综合研究

📊 可视化
Generate charts, diagrams, interactive p...

🌟 更多能力
由对话引擎驱动的模式

你想学点什么?

🧠 解题
多步推理与问题求解

🎓 精通之路
掌握式辅导：硬门槛 + 间隔复习

💬 聊天 ^ +

📖 👤 🌈 🗨️ ⬆️

Extending the Personalization Substrate

Adaptive Learning Capabilities

- **Deep Research（知识综合）：**
 - 多阶段 Agent 完成问题分解 → 多源检索 → 证据综合 → 报告生成
 - 研究过程同步更新 Learner Profile
- **Subagent Consultation（外部专家协作）：**
 - 调用外部 Agent（如 Claude Code、Codex）
 - 将返回结果作为可验证证据参与教学，而非直接替代 Tutor
- **Visualize（知识表示）：**
 - 根据知识特点自动生成图表、动画等可视化内容
 - 将表示方式作为教学策略，降低学习认知负担
- **Co-Writer（知识表达）：**
 - 基于 Markdown 持续写作与 AI 辅助编辑，帮助查漏补缺
 - 文档修改记录可沉淀为后期学习证据
- **Mastery Path（长期练习）：**
 - 根据 Learner Profile 制定长期学习计划
 - 动态安排练习、评估结果并持续调整学习路径

Extending the Personalization Substrate



南京航空航天大学
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

Interactive Books through Substrate Reuse

学习上下文构建 (Input Snapshot)

整合学习目标、历史对话、知识库、笔记等多源信息，生成**教材提案**，确定范围、受众与教学目标

01

教材组织 (Curricular Spine)

构建章节结构与概念图
显式组织知识依赖关系
与学习路径

02

交互教材生成 (Block-level Compilation)

按模块生成文本、图表、测验、动画等教学单元
整合 Research、Visualization、Question Generation 等能力

03

持续学习 (Persistent Interaction)

支持教材编辑、章节补充与测验更新
学习记录持续更新
Learner Profile，实现长期个性化学习

04

Extending the Personalization Substrate



Proactive Multi-channel Tutoring 从被动响应到主动辅导

Persistent Partner:



- 每个 Partner 拥有独立工作空间、Persona 与长期记忆
- 共享 DeepTutor 能力层，实现数学导师、研究助手等专用角色

Skill-based Capability Reuse:



- Skills 定义高层教学流程，而非重新实现 Tutor
- 支持解释、解题、研究、可视化、文档编辑等能力复用

Heartbeat & Multi-channel Interaction:

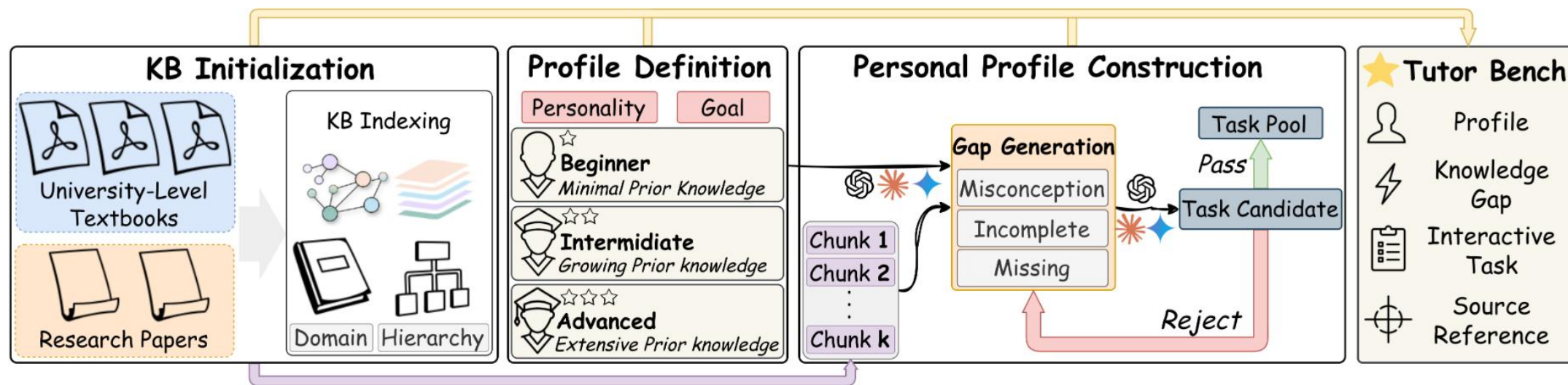


- 根据学习状态主动提醒、复习与跟进
- 支持 Slack、Discord、Telegram 等多平台统一交互
- 跨设备共享同一 Learner Profile 与学习历史

Core Idea:



通过 Persistent Partner + Skills + Heartbeat，实现长期、多渠道、主动式个性化辅导。



Benchmark Motivation:

- 现有教育 Benchmark 多以教师或模型为中心，缺乏真实学习者建模
- TutorBench 将学习者画像、知识薄弱点与教学任务统一建模，实现个性化辅导评测

Construction Pipeline:

- 在5大学科（涵盖人文、科学、工程、商业和前沿研究）上构建知识库（30 个知识库）
- 为每个知识库构建初级、中级、高级三类学习者画像
- 针对每个画像生成三类知识缺口：
 - Misconceptions（错误概念）
 - Incomplete Understanding（理解不完整）
 - Missing Knowledge（知识缺失）

结合知识缺口生成交互式教学任务，并经过人工审核保证质量

Dataset Statistics:

30 个知识库，覆盖 5 个学科领域

90 个学习者画像（ 3×30 ）

270 个教学任务（每个画像对应 3 个任务）

每个任务均关联真实知识页面与知识缺口作为评测依据

First-Person Interactive Evaluation

实验设置:

- 从 5 大互补维度评估 DeepTutor: 第一人称交互评测、跨领域泛化、人类偏好、消融实验与 Agent 泛化能力
- Interactive Book 与 Partners 等长期能力留作未来研究
- 定量实验主要聚焦闭环个性化辅导核心

First-person Interactive Evaluation:

- 提出第一人称交互评测协议 (TutorBench)
- Student Simulator (Gemini-3-Flash) 根据学习者画像、知识缺口和任务开展多轮辅导
- 对辅导过程进行个性化评分 (Claude Sonnet 4.6 serves as the judge), 而非传统单轮问答评测

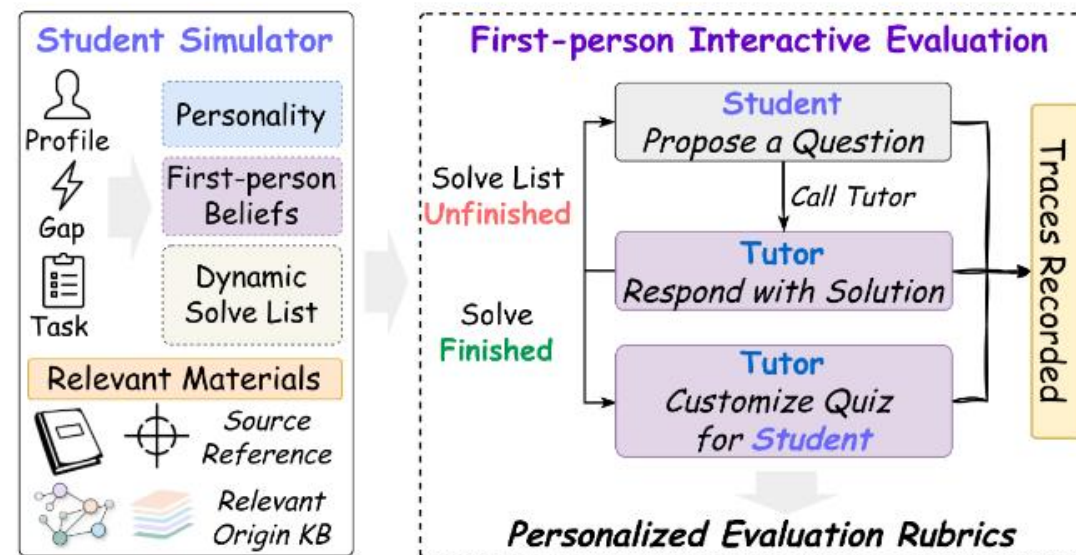


Figure 5 | TutorBench first-person interactive evaluation protocol.

First-Person Interactive Evaluation



Evaluation Metrics:

Tutor (辅导质量) * Source Faithfulness (事实一致性) * Personalization (个性化程度) * Applicability (可操作性) * Vividness (表达丰富性) * Logical Depth (推理深度)	Practice (练习质量) * Fitness (难度匹配) * Groundedness (知识依据) * Diversity (题目多样性) * Answer Quality (答案质量) * Cross Concept (跨概念综合)
--	--

Baselines & Results:

Baseline:

- Naive Tutor
- CoT Tutor
- Self-Refine Tutor
- ReAct Tutor

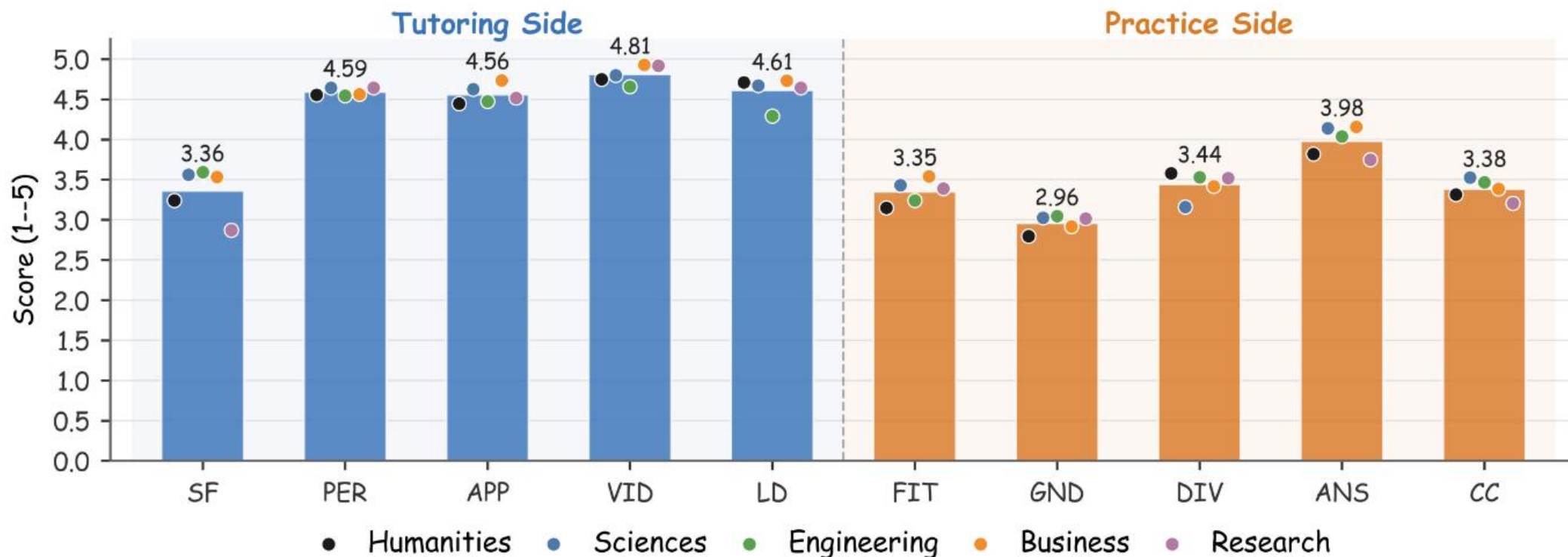
Results:

- DeepTutor 整体质量提升 10.76%
- Tutor 指标在 Personalization、Vividness、Logical Depth 提升最明显
- Practice 指标在 Groundedness、Diversity、Cross Concept 表现最佳

DeepTutor: Towards Agentic Personalized Tutoring

System	Tutoring Quality						Practice Quality						Overall	
	SF	PER	APP	VID	LD	Avg	FIT	GND	DIV	ANS	CC	Avg	OQ	Δ%
Naive Tutor	3.22	4.24	4.44	3.89	3.99	3.96	<u>3.20</u>	2.46	2.83	3.87	<u>3.12</u>	<u>3.10</u>	3.53	—
CoT Tutor	3.34	4.22	4.47	3.82	4.01	3.97	3.18	<u>2.50</u>	2.79	3.81	3.04	3.06	3.52	-0.28%
Self-Refine Tutor	3.28	4.28	4.61	<u>3.94</u>	<u>4.14</u>	<u>4.05</u>	3.17	2.49	<u>2.88</u>	3.88	2.97	3.08	<u>3.57</u>	+1.13%
ReAct Tutor	<u>3.35</u>	<u>4.37</u>	4.40	3.70	3.96	3.96	3.16	2.47	2.75	<u>3.94</u>	3.07	3.08	3.52	-0.28%
DEEPTUTOR	3.36	4.59	<u>4.56</u>	4.81	4.61	4.39	3.35	2.96	3.44	3.98	3.38	3.42	3.91	+10.76%

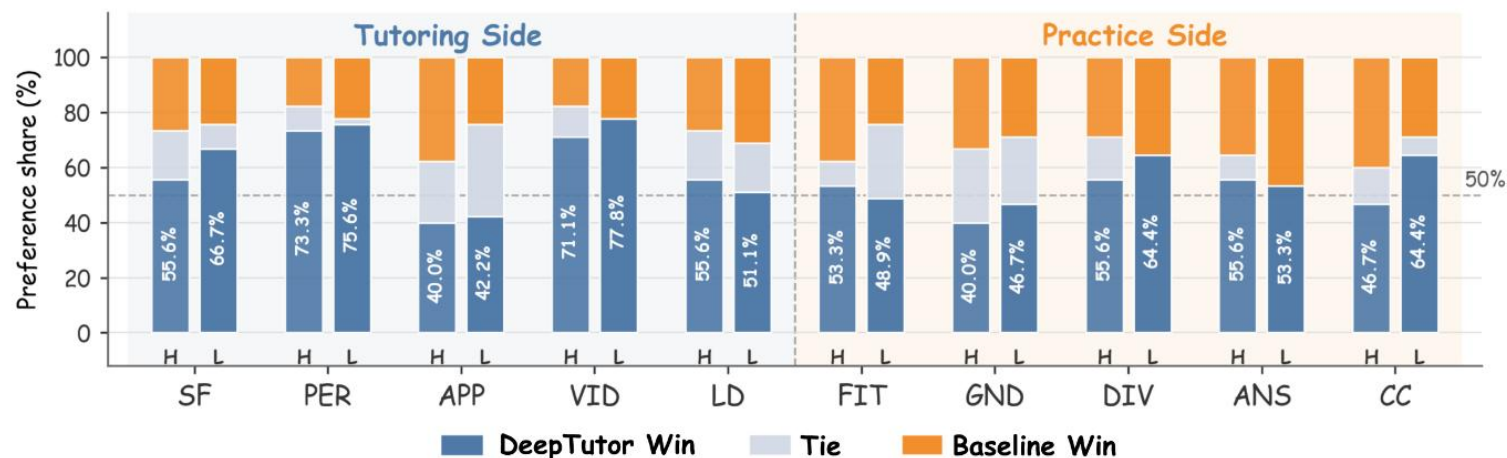
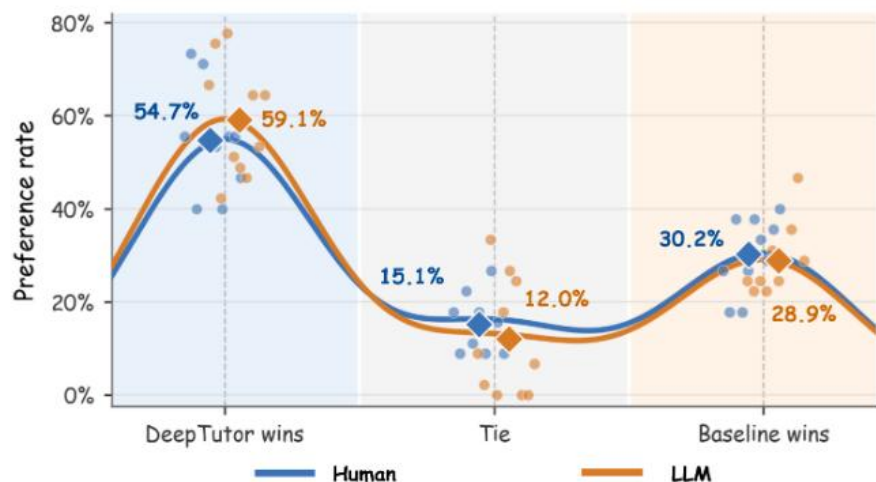
Cross-Domain Generalization



跨领域泛化能力:

- 在 5 个学科领域中均保持稳定性能，具有良好的跨领域泛化能力。
- 不同领域总体质量仅波动 0.16 分，各项指标平均波动 0.36 分。
- Source Faithfulness 波动最大，对不同学科的敏感性更高。
- 教学能力相关指标整体表现稳定。
- 练习生成中 Answer Quality 表现最佳，Groundedness 仍是主要瓶颈。

Human Evaluation of Preference Alignment



Blind Pairwise Evaluation:

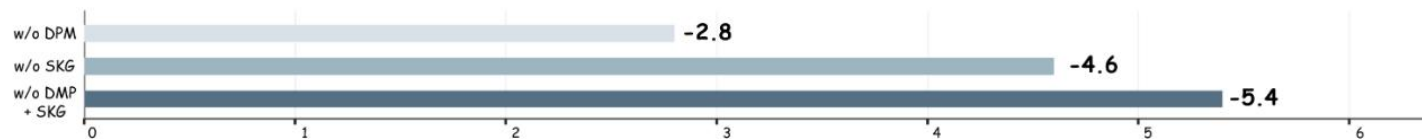
- 在 TutorBench 中抽取 45 个会话（5 个领域，每个领域 9 个）
- 人工评审对匿名的 DeepTutor 与 Naive Tutor 输出进行两两比较
- 两个系统共享相同的学习者画像、任务、课程知识、学生模拟器及练习生成请求
- 若两位评审意见不一致，则记为 Tie

Results:

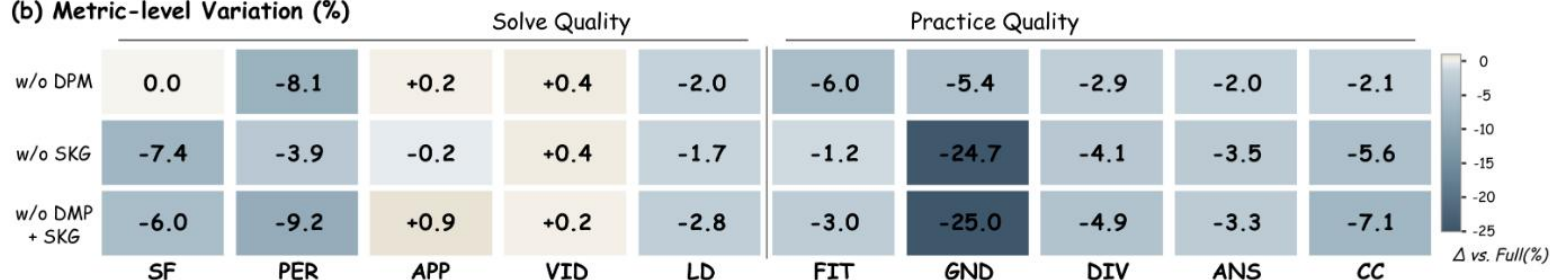
- 人工评审与 LLM Judge 均更偏好 DeepTutor
- 在 Personalization、Vividness、Diversity 上优势最明显
- 人工与 LLM 的偏好高度一致（Pearson $r=0.82$, Spearman $\rho=0.83$ ）
- 说明 LLM Judge 能较好反映人工评审偏好，而非简单偏向 DeepTutor

Experiments

(a) Overall Quality (OQ) Drop (%)



(b) Metric-level Variation (%)



Ablation Study:

- 分别移除 SKG、DPM 及两者共同移除
- 去除 SKG 后，GND、SF、CC 下降最明显
- 去除 DPM 后，PER、FIT 下降最明显
- 同时移除两者性能下降最大，说明 SKG 与 DPM 相互补充

Pass@1	STEM Reasoning			General Agent (GAIA)				Long Ctx	Avg Δ
	HLE*	GPQA-D**	LiveBench***	L1	L2	L3	Overall†	AA-LCR	
Gemini-3-Flash	19.40	81.31	70.00	43.40	40.70	15.38	37.58	63.00	+29.24%
+DEEPTUTOR	30.80	84.85	96.00	56.60	50.00	23.08	47.88	74.67	
Sonnet-4.5	8.40	72.22	64.33	37.74	30.23	7.69	29.09	53.33	+32.03%
+DEEPTUTOR	14.60	73.23	82.00	50.94	48.84	23.08	45.45	54.00	
Qwen-3.5-Plus	16.80	88.38	69.00	44.23	35.71	13.64	33.94	66.00	+25.69%
+DEEPTUTOR	24.20	87.88	93.00	50.94	53.49	32.00	49.09	69.67	
GPT-5-Mini	16.46	80.81	71.00	32.08	30.23	7.69	27.27	68.67	+28.51%
+DEEPTUTOR	21.20	80.30	93.00	54.72	52.33	26.92	49.09	71.00	
Minimax-M2.5	14.00	82.83	59.30	35.29	22.78	12.00	23.64	66.00	+31.50%
+DEEPTUTOR	19.40	83.33	73.00	52.94	44.30	32.00	42.42	76.40	

Extended Generalization:

- 去除个性化模块，仅保留 Investigate-Solve-Write 推理流程
- 在 HLE、GPQA、LiveBench、GAIA、AA-LCR 五个公开 Benchmark 上评测
- 五类 Backbone 平均性能提升 25.69%~32.03%
- 说明三阶段推理流程具有良好的通用 Agent 问题求解能力

Thanks