面向军事装备领域的可解释知识推理评测任务

推理是认知能力的重要组成，知识推理已成为军事智能问答、情报分析、判断决策等应用的核心技术。现有知识推理方法多侧重答案准确性，在推理过程可解释性方面缺乏支撑，难以满足军事领域高可信、可追溯的应用需求。为此，任务组织方在 2020-2025 年连续 6 年组织测评任务的基础上，特设面向军事装备领域的可解释知识推理评测任务。任务基于公开数据中收集整理的军事装备领域文本，构建了高精度知识图谱，并协调领域专家人工标注了可解释知识推理样例数据集，包括复杂多跳推理问题、问题答案以及由多个关联三元组组成的推理证据链。要求参赛队伍针对给定的自然语言问题，依托从文本证据中抽取得到的知识图谱执行知识推理，并输出最终答案以及推理证据链，以体现推理过程的可解释性。本次任务旨在推进军事领域可解释认知推理技术研究，构建军事领域专业化数据体系，同时遴选优秀团队，共建“以图促智”的军事垂直领域智能服务生态。

1、任务定义

面向军事装备领域的可解释知识推理评测任务要求参赛队伍针对给定的自然语言问题，依托从文本证据中抽取得到的知识图谱执行多跳推理，并输出最终答案以及由多个关联三元组组成的推理证据链，以体现推理过程的可解释性。

1.1 输入

（1）非结构化军事文本；（2）从文本中抽取的知识图谱；（3）自然语言问题。

1.2 输出

（1）最终答案；（2）由三元组构成的支持答案生成的推理路径。

2、数据集描述

样例数据包括若干组“军事装备领域文本-知识图谱-自然语言问题组”，以每个问题对应的答案和推理链。

2.1 输入数据：

军事装备领域文本、知识图谱和自然语言问题，包含一个文本文件 document.json、多个知识图谱“KG_01.json、KG_02.json...”、一个查询文件 qa.json。

数据类型	内容	作用
文本	描述、新闻、事件片段	用于构造知识图谱
知识图谱	实体、关系、时间事实	结构推理
查询	问题（包含约束条件）	任务输入

数据由以下部分组成：

（1）文本

非结构化军事装备领域知识文档

（2）知识图谱

从文本证据中抽取得到的知识图谱，每个文本抽取出一个知识图谱，形式为三元组：（头实体，关系，尾实体/时间、地点属性值）。

（3）查询

自然语言推理问题，每个文本&知识图谱对应一组推理问题，每组问题包含不同推理类型/不同推理难度等级，推理类型（推理难度等级）划分如下：

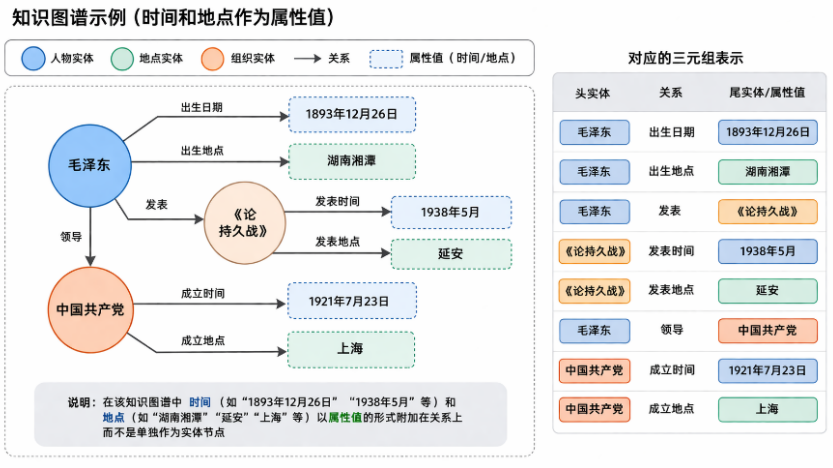
- L1：事实查询
- L2：标准多跳推理
- L3：多约束复合推理

数据采用统一 JSON 格式发布，每个文件包含的字段如下：

document.json：

字段名	描述	样例
doc_id	文本的唯一 id	doc_01
content	文本内容	此次部署第 13 陆战队远征队将搭舰随行。\n\n 地面作战单元为陆战 4 团 2 营，航空作战单元为....

KG_01.json：（每个文本抽取出一个知识图谱，存储文件命名格式为“KG_”+doc_id+“.json”）

字段名	描述	样例
triple_id	三元组唯一 id	triple_01
sub	三元组主体	
sub_type	三元组主体类别	
relation	关系	
obj	三元组客体	
obj_type	三元组客体类别	

qa.json：

字段名	描述	样例
query_id	查询问题 ID	doc_01_001
doc_id	问题对应的文本和知识图谱编号	doc_01
question	查询的自然语言问题	“在第二次世界大战期间 指挥过诺曼底登陆且隶属于同盟国的军事人物是谁？”
difficulty	难度等级（L1、L2、L3）	"L3"

2.2 输出数据

可解释知识推理的输出为：基于上述文本和知识图谱回答自然语言问题，并输出答案和推理路径，对应文件 qa_answer.json。

输入数据中每一个问题可能有一个或多个可支持答案的推理路径。qa_answer.json 文件包含基于上述输入数据回答的答案和所有可支持答案的推理路径，推理路径以三元组形式表示，包括一条最标准、最短、最核心的路径 gold_reasoning_paths_main，和若干条其他同样可支持答案的替代推理路径 gold_reasoning_paths_alt，只要命中上述任一推理路径即算正确。

qa_answer.json 文件中包含的字段如下：

字段名	描述	样例
query_id	查询问题 ID	doc_01_001
doc_id	问题对应的文本和知识图谱编号	doc_01
question	查询的自然语言问题	“在第二次世界大战期间 指挥过诺曼底登陆且隶属于同盟国的军事人物是谁？”
answer_type	答案类型，单值答案（单个实体/数值/布尔值）或者集合型答案（实体集合）	"人物"
answers	答案	["艾森豪威尔"],
gold_reasoning_paths_main	主要推理路径（得出答案所需三元组数量最少）	[["诺曼底登陆", "指挥官", "艾森豪威尔"], ["艾森豪威尔", "隶属于", "同盟国"], ["诺曼底登陆", "发生时间", "1944 年 6 月 6 日"], ["1944 年 6 月 6 日", "处于时期内", "第二次世界大战"]]
gold_reasoning_paths_alt：	可替代推理路径，可能有一条或者多条，也可能没有	[["诺曼底登陆", "执行方", "盟军远征部队"], ["盟军远征部队", "指挥官", "艾森豪威尔"], ["盟军远征部队", "隶属于", "同盟国"],

字段名	描述	样例
		["诺曼底登陆", "发生时间", "1944 年 6 月 6 日"], ["第二次世界大战", "开始时间", "1939 年 9 月 1 日"], ["第二次世界大战", "结束时间", "1945 年 9 月 2 日"], ["1944 年 6 月 6 日", "晚于", "1939 年 9 月 1 日"], ["1944 年 6 月 6 日", "早于", "1945 年 9 月 2 日"]]

constraints	约束条件(除去基于问题抽取的关键实体外，应关注**时间/地点**等属性值)	{
		"time": "第二次世界大战期间",
		"camp": "同盟国"
		}

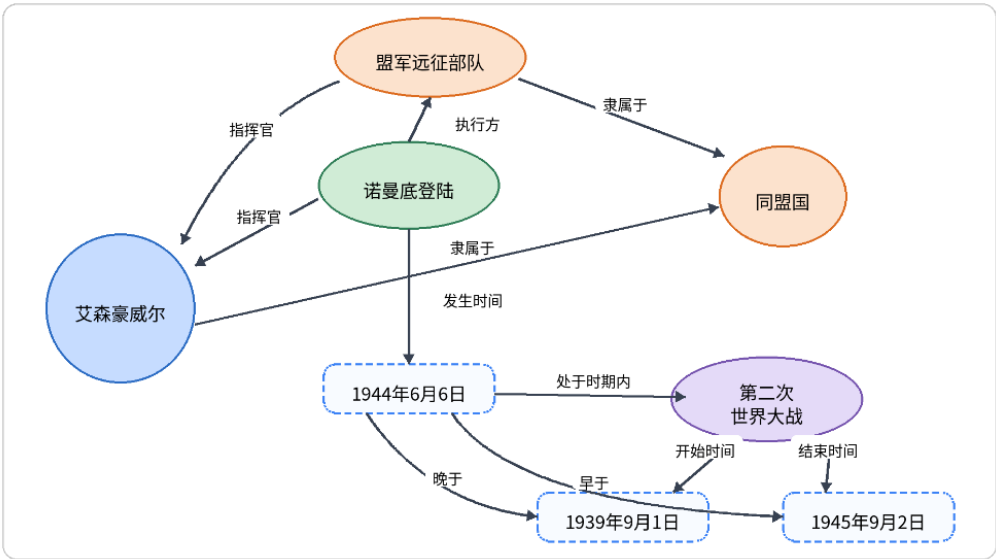
difficulty	难度等级（L1、L2、L3）	"L3"

示例：

qa_answer.json


```
{
  "query_id": "doc_01_001",
  "question": "在第二次世界大战期间 指挥过诺曼底登陆且隶属于同盟国的军事人物是谁",
  "answer_type": "人物",
  "answers": ["艾森豪威尔"],
  "gold_reasoning_paths_main": [
    ["诺曼底登陆", "指挥官", "艾森豪威尔"],
    ["艾森豪威尔", "隶属于", "同盟国"],
    ["诺曼底登陆", "发生时间", "1944 年 6 月 6 日"],
    ["1944 年 6 月 6 日", "处于时期内", "第二次世界大战"]
  ],
  "gold_reasoning_paths_alt": [
    ["诺曼底登陆", "执行方", "盟军远征部队"],
    ["盟军远征部队", "指挥官", "艾森豪威尔"],
    ["盟军远征部队", "隶属于", "同盟国"],
    ["诺曼底登陆", "发生时间", "1944 年 6 月 6 日"],
    ["第二次世界大战", "开始时间", "1939 年 9 月 1 日"],
    ["第二次世界大战", "结束时间", "1945 年 9 月 2 日"],
    ["1944 年 6 月 6 日", "晚于", "1939 年 9 月 1 日"],
    ["1944 年 6 月 6 日", "早于", "1945 年 9 月 2 日"]
  ],
  "constraints": {
    "time": "第二次世界大战期间",
    "camp": "同盟国"
  },
```

"difficulty": "L3"
}



3、评价指标

本次可解释知识推理任务的评测包含以下 4 个要素：答案正确性 AnswerScore 和证据质量 EvidenceScore，推理有效性 ReasoningScore，约束满足度 ConstraintScore。任务最终得分为每个问题得分总和，对每个问题，系统综合得分定义为：

$$Score=\alpha*AnswerScore+\beta*EvidenceScore+\gamma*ReasoningScore+\delta*ConstraintScore$$

针对不同难度等级采用差异化权重：

L1：事实查询，更关注答案正确性，α=0.65, β=0.2, γ=0.1, δ=0.05。

$$Score_{L1}=0.65*AnswerScore+0.2*EvidenceScore+0.1*ReasoningScore+0.05*ConstraintScore$$

L2：标准多跳推理，重点是答案 + 路径质量平衡，α=0.45, β=03, γ=0.15, δ=0.1。

$$Score_{L2}=0.45*AnswerScore+0.3*EvidenceScore+0.15*ReasoningScore+0.1*ConstraintScore$$

L3：多约束复合推理，更关注约束满足与推理有效性，α=0.3, β=0.25, γ=0.25, δ=0.2。

$$Score_{L3}=0.3*AnswerScore+0.25*EvidenceScore+0.25*ReasoningScore+0.2*ConstraintScore$$

指标	内涵
Score	每个问题综合指标
AnswerScore	答案正确性，反映“答对”的基本能力
EvidenceScore	证据质量，反映“能否用证据支撑答案”
ReasoningScore	推理有效性，衡量解释链是否真正支撑答案，并具有内部一致性
ConstraintScore	约束满足度，衡量解释是否完整覆盖并正确绑定问题中的关键约束条件

指标	内涵
EM	Exact Match，问题是否完全答对
F1	准确和完整的综合平衡
P	Precision，输出结果是否干净、少误报
R	Recall，是否找全所有正确答案
TripleMatch	证据三元组的匹配质量，通过计算 F1 值判断找到的证据是否既准确又完整
P_t	Triple Precision 给出的证据三元组是否准确
R_t	Triple Recall，是否找全了证据三元组
Parsimony	证据简洁性 / 最小充分证据
Constrain F1	输出的约束是否覆盖所有约束条件，约束集合匹配 F1
P_c	Constrain Precision 给出的约束是否准确
R_c	Constrain Recall，是否找全了所有约束

3.1 答案正确性

答案正确性 AnswerScore，反映“答对”的基本能力。

(1) 单值答案

单实体答案/单数值答案/单布尔答案/格式明确的标准化答案，采用规范化精确匹配。EM（Exact Match）衡量是否给出了严格正确的最终答案，输出与标准答案完全一致则记 1，否则记 0。

(2) 集合型答案：

对于多实体集合型答案，采用规范化集合 F1。F1：Precision（精确率）和 Recall（召回率）的调和平均，在“答得准”和“答得全”之间的综合平衡能力：

$$F1=\frac{2PR}{P+R}，\ P=\frac{|Pred\cap Gold|}{|Pred|}，\ R=\frac{|Pred\cap Gold|}{|Gold|}$$

$$AnswerScore=\begin{cases}EM, & \text{答案类型为单实体答案/单数值答案/单布尔答案/格式明确的标准化答案}\\ F1, & \text{答案类型为集合型答案}\end{cases}$$

3.2 证据质量

证据质量 *EvidenceScore* 用于输出三元组与 gold 推理路径的重合情况。

$$EvidenceScore = 0.6 * TripleMatch + 0.4 * Parsimony$$

(1) *TripleMatch*：证据三元组匹配质量

衡量对提交的三元组集合$Pred_t$与标准路径集合$Gold_t$：

Triple Precision（证据精确率）：输出的证据三元组路径中，有多少真正属于标准路径：

$$P_t=\frac{|Pred_t\cap Gold_t|}{|Pred_t|}$$

Triple Recall（证据召回率）：标准路径中，有多少证据三元组被找到了，衡量是否找全关键证据：

$$R_e=\frac{|Pred_t\cap Gold_t|}{|Gold_t|}$$

TripleMatch F1：综合衡量证据的“准确”和“完整”：

$$TripleMatch=F1(P_t,R_t)=\frac{2P_tR_t}{P_t+R_t}$$

(2) *Parsimony*：证据简洁性

衡量证据简洁性，即是否为最小充分证据。同样支持答案，若输出冗余三元组和只输出核心三元组得分不同，引入冗余惩罚：

$$Parsimony = min\left(1,\frac{|Gold_{t,min}|}{|Pred_t|}\right)$$

其中：$Gold_{t,min}$是主路径 gold_reasoning_paths_main 的三元组数量，$Pred_t$是提交证据的三元组数量。

3.3 推理有效性

推理有效性 *ReasoningScore* 用于衡量解释链是否真正支撑答案，并具有内部一致性。仅仅命中若干三元组还不够，还要看是否组成从关键实体到答案的有效链，通过图路径匹配判断。

$$ReasoningScore = 0.4 \cdot \text{NodeCoverage} + 0.4 \cdot \text{EdgeOrder} + 0.2 \cdot \text{HopMatch}$$

其中$NodeCoverage$用于衡量问题关键实体是否命中；$EdgeOrder$用于衡量答案节点是否被证据链正确连接以及关系序列一致性，即关系顺序是否符合 gold；HopMatch用于衡量标准多跳链中的中间桥接节点是否出现。

3.4 约束满足度

约束满足度 *ConstraintScore* 用于衡量输出的约束是否覆盖了问题中的所有关键约束条件。通过计算约束集合匹配 F1。

$$ConstraintScore=Constrain\ F1=\frac{2P_cR_c}{P_c+R_c},\ P_c=\frac{|Pred_c\cap Gold_c|}{|Pred_c|},\ R_c=\frac{|Pred_c\cap Gold_c|}{|Gold_c|}$$

4、任务提交及评分

4.1 实现方案要求

本次评测，要求参赛队调用开源大模型来完成可解释知识推理，一方面需要构建有效的 prompt，一方面需要考虑多个大模型的联合使用。

模型选择：开源大模型（便于测试以及复现）。

实现方式：可以只调用一个大模型，也可以同时使用多个大模型进行整合。

4.2 最终提交材料

组织方将于赛程规定时间节点依托指定平台发布标准化测试数据集，各参赛队需在限定内（具体时间以组委会公告为准，届时将提前通知）提交代码及测试结果。要求代码可复现，测试结果需严格遵循输出格式示例，封装为指定格式的 JSON 文件，并命名为"result.json"。

测试结果提交完成后，各参赛队需在规定时间节点内提交所有参赛材料，包括：

(1) 方法描述文档

(2) 方法的实现代码，要求代码可复现

4.3 依托平台

本次评测的任务提交，将依托红山开源平台（https://www.osredm.com/competition/zstp2026/）开展，参赛队在该平台完成注册、数据下载、结果提交等工作。

4.4 评估方式

组织方将于赛程规定时间节点提供样例数据集，用于参赛者自行验证。遴选军事装备领域的若干组“文本知识图谱-自然语言问题组”数据作为测试数据集，用于最终评估模型在军事垂直领域的应用效果。进入最终测试环节后，参赛队需在限定时间内，针对测试数据集给定的自然语言问题，依托从文本证据中抽取得到的知识图谱执行多跳推理，并按格式输出最终答案以及由多个关联三元组组成的推理证据链，而后提交测试结果及方案代码。结果提交后，平台将调用预设评估脚本进行多维度指标验证，评分结果即时展示在队伍控制台。赛后，组织方将结合参赛队提交的说明文档对代码进行复现，验证提交结果的真实性。

(1) 样例数据发布阶段

组委会将于赛程规定时间节点依托指定平台发布样例数据集，参赛队可自行用于验证。

(2) 测试数据发布阶段

组委会将于赛程规定时间节点通过指定平台发布标准化测试数据集，各参赛队需在限定内（具体时间以组委会公告为准，届时将提前通知）完成以下操作：

- 执行自动化流程生成测试结果；
- 严格遵循输出格式示例，将输出结果封装为指定格式的 JSON 文件。

(3) 结果提交与自动评估阶段

- 文件规范：输出文件必须按"result.json"格式命名（区分大小写），同时需要上传方法的实现代码，要求代码可复现
- 提交方式：在指定平台上传系统进行提交（每个 TeamID 限 5 次提交，恶意刷分将判定结果无效）
- 评估机制：系统将实时调用预设评估脚本进行指标验证，在线展示评分结果。
- 时效要求：以截止时间前最高一次成绩作为最终成绩（逾期提交通道自动关闭）

(4) 结果真实性检验阶段

结果提交截止后，组织方将结合参赛队提交的说明文档对代码进行复现，验证提交结果的真实性。待验证符合后，确认参赛队得分，从而确定最后名次。

5、报名方式

参赛队需首先在红山开源平台完成注册（https://www.osredm.com/competition/zstp2026），并在“参赛报名”处填写相关信息，鉴于本次评测专家标注样例数据集的价值，参赛队需提交队长所在单位证明（相关要求见报名页面）。参赛队通过审批环节后，可在红山开源平台完成样例数据集下载与结果提交。

6、奖励设置

CCKS2026 组委会为本次任务提供 3 万元奖金（税前），奖励 3 支获奖团队，奖金设置如下：

第一名：12,000 元（税前）

第二名：10,000 元（税前）

第三名： 8,000 元（税前）

7、时间安排

时间安排初定如下，后续如果有调整，以 CCKS2026 发布信息为准。

- 报名时间：4 月 17 日—7 月 3 日
- 样例数据集发布：6 月 1 日
- 结果提交：7 月 3 日
- 结果及测试数据集公示：7 月 17 日
- CCKS 会议日期(评测报告及颁奖)：8 月 21 日—23 日

8、组织者信息

任务组织者：
张　静，军事科学院系统工程研究院

任务联系人邮箱：
（可通过该邮箱联系沟通）
xwyu18@163.com