arXiv:2301.10107v1 [cs.AI] 24 Jan 2023

故事塑造：通过故事教会智能体类人行为

Xiangyu Peng*，**Christopher Cui***，**Wei Zhou**，**Renee Jia**，**Mark Riedl**佐治亚理工学院{xpeng62, ccui46, wzhou322, rjia35}@gatech.edu, riedl@cc.gatech.edu

摘要

在某些情境下，强化学习智能体的奖励设计可能十分困难：我们不仅希望智能体在世界中实现某种效果，还关心它如何实现这一效果。例如，我们可能希望智能体遵循对常识的缄默理解，使其行为符合安全方面的偏好，或在交互式游戏中扮演特定角色。讲故事是一种传达隐性程序性知识的方式。我们提出一种名为故事塑造的技术：强化学习智能体从一则示例故事中推断出完成任务所需的隐性知识，并通过内在奖励机制，奖励自身采取行动，使当前环境符合推断出的故事世界。具体而言，故事塑造根据观测推断世界状态的知识图谱表示，并从示例故事中推断另一知识图谱。智能体推断出的世界状态图与故事世界图之间的相似度决定了内在奖励。我们在需要常识推理的文本游戏中进行了实验，并借此塑造智能体作为虚拟游戏角色的行为。



图 1：文本游戏《9:05》的节选，其中原始强化学习智能体执行的动作以 表示，而采用**故事塑造**的强化学习智能体执行的动作以 表示。采取动作“向西走”，这只会最大化期望奖励，却忽略了环境中的常识知识。选择做更像人类会做的事——吃一个 Pop-Tart——这是基于一段以自然语言撰写的故事；该故事可以由我们的系统提供，也可以由系统自动创建。

1 引言

强化学习（RL）是一类技术，智能体通过与环境反复交互，学习如何执行序列任务。当智能体在环境中达到某些状态、执行某些动作，或使世界在特定状态对之间发生转换时，就会获得奖励。因此，强化学习是一个学习如何最大化预期未来奖励的过程。奖励信号的设计——即何时给予奖励（或惩罚），以及奖励的多少——决定了智能体应采取何种最优行为。

在某些情况下，奖励设计尤其困难：我们不仅希望智能体在世界中实现某种效果，还关心如何实现这种效果。例如，我们可能希望智能体在执行任务时遵守社会规范。这样一来，人类与人工智能驱动系统协作会更安全，因为这些系统符合我们的预期和偏好 [Frazier 等，2020年；Nahian 等，2021年；Ammanabrolu 等， 2022]。我们可能希望智能体在执行过程中展现常识知识 [Dambekodi 等， 2020]。在电子游戏环境中，我们可能希望人工智能驱动智能体扮演具有不同兴趣或做事方式的各类角色 [Urbanek 等，2019年；Ammanabrolu等，2021]。

我们介绍一种用于指定智能体行为偏好的新技术，**故事塑造**。在故事塑造中，

*贡献相同

奖励设计者通过提供示例故事来指定如何执行任务。故事是在人与人之间传递隐性程序性知识的高效方式。讲故事是一种交流方式，其中的细节会被抽象省略，因为假定接收者拥有共同的知识基础，能够在需要时重建细节以理解内容 [Hedlund 等人， 2002]。

故事塑造是一个过程，在此过程中，智能体从示例故事中反向推断隐式状态信息，并构建丰富的内在的 奖励信号，引导自身采取行动，使其环境更接近故事中的隐含世界。具体而言，智能体从故事中推断出一个知识图谱 ，其中包含$\langle subject, relation, object\rangle$三元组，分别对应从故事中推断出的每一种关系。智能体还从观测中提取关系，以构建一个概括运行环境的知识图谱。智能体会根据当前世界状态表示与从示例故事中得出的期望表示之间的相似程度来奖励自己。

故事塑造与从示范中学习（LfD）有关。在从示范中学习中，智能体会获得一组轨迹，这些轨迹通常由人类在同一环境中执行。学习智能体必须重建人类示范者所遵循的潜在策略。通常假定示范是完整的（没有遗漏步骤），且在与智能体所处环境相同或非常相似的环境中完成。相比之下，故事塑造不假定故事完整或可执行，也不假定故事描述的是同一个环境。相反，故事可能提及关键物体、地点或活动，但省略细节。此外，由于故事并未描述智能体所处的确切环境，物体和地点可能有所不同或缺失，动作也可能无法执行。

我们在文本游戏中对故事塑造进行实验，其中观测和动作完全以文本形式呈现。文本游戏属于部分可观测环境，具有庞大的状态和动作空间，并且往往包含需要长程因果依赖的谜题 [Hausknecht 等，2020a]。研究还表明，文本游戏可以迁移到视觉领域和现实世界领域 [Wang 等，2022年；Shridhar 等，2021]。我们在三个游戏平台上开展实验，这些平台要么 (a) 需要常识，要么 (b) 能够展示我们的智能体可以以更像人类的方式进行角色扮演。此外，我们还表明，经过故事塑造的智能体能够根据不同的角色偏好调整行为。

2 背景与相关工作

文本游戏是一种回合制游戏，玩家必须阅读由人类撰写的自然语言（通常为英语）描述，了解局部环境，并用简短的文本描述动作作出回应。文本游戏可以定义为一个部分可观察的马尔可夫决策过程： $\langle S, P, A, O, \Omega\ R, \gamma\rangle$，分别表示环境状态集合、状态间的条件转移概率、用于组成文本指令的词汇表或词语、观测、观测条件概率、奖励函数和折扣因子。转移概率和观测概率P 以及 Ω 通常对智能体而言是未知的。观测是文本序列，动作由词汇表中的一个至五个词元组成。强化学习智能体尝试学习一种策略 $\pi(o) \rightarrow a$，以最大化未来的期望奖励。

用于文本游戏的知识图谱。知识图谱是一组 $\langle subject, relation, object\rangle$ 元组。基于知识图谱的强化学习智能体已被证明在文本游戏中达到了最先进水平 [Ammanabrolu 和 Riedl，2018年；Ammanabrolu 等，2020a；Ammanabrolu 和 Hausknecht，2020年；Ammanabrolu 等，2020b；Xu 等，2020年；Peng 等，2021a]。这些智能体从文本观察中推断物体和关系，并将该知识图谱用作世界状态的长期记忆，以此应对部分可观测性。我们的工作基于 KG-A2C [Am- manabrolu 和 Hausknecht， 2020] 智能体架构；该架构使用 ALBERT [Lan 等， 2019] 语言模型从文本观察中推断物体和关系，并利用图注意力网络生成动作序列。KG-A2C 使用知识图谱表示世界状态并筛选动作，而我们的故事塑造方法还利用知识图谱计算稠密奖励信号，这表明基于知识图谱的强化学习仍有尚未发掘的潜力。

内在奖励。 内在奖励为探索提供定性指导 [Schmidhuber，1991年；Oudeyer 等，2007年；Barto，2013] 并促使智能体在没有环境直接反馈的情况下表现出特定行为。这些奖励可以有多种形式，例如比较智能体的预测与现实 [Stadie 等，2015；Burda等，2018年；Kim 等， 2019], 或衡量其在自生成目标上的表现 [Vezhnevets 等，2017年；Levy 等，2018年；Nachum等，2018年；Nair 等，2018年；Pong 等， 2019]。Ammanabrolu 等 [2020b] 通过向知识图谱添加节点和边，为文本游戏智能体提供内在奖励。我们的方法在内在层面激励智能体探索那些能够添加示例故事所预期节点和边的状态。与此相关，Harrison 等 [2016] 提出的从故事中学习 技术利用故事指导强化学习智能体。然而，该方法需要数十个示例故事，并将每个事件视为模块化分层策略中的一个目标；而我们只需要一个故事，并生成一个统一策略。

3 故事塑造

故事塑造 使强化学习智能体能够从关于某项任务的示例故事中学习隐性知识，并对那些使运行环境更符合推断出的故事世界的动作给予自身奖励。我们的方法从给定的示例故事开始（也可以自动生成），智能体会将其转换为一个知识图谱，称为故事知识图谱(§3.1)。在强化学习训练过程中，智能体探索游戏世界，并构建一个内部状态知识图谱，称为世界知识图谱 (§3.2)。随后，根据世界知识图谱与故事知识图谱之间的相似度计算内在奖励，并据此更新智能体(§3.3)。该方法的整体概览如图2所示。

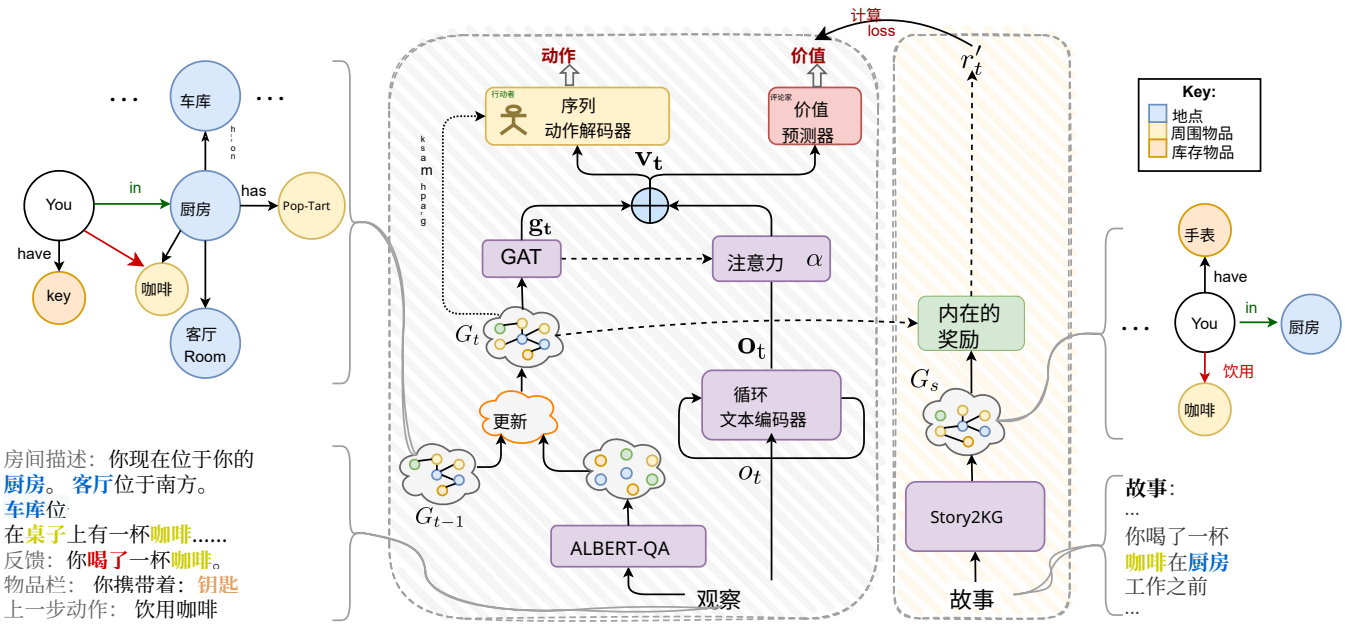

图2：步骤 t中知识图谱提取和采用故事塑造的强化学习智能体架构。

游戏：9:05
首次观察：
我早上醒来。南边有一间浴室，东边有一扇门通往客厅。茶几上放着一部电话、一个钱包和几把钥匙。电话响了。

目标：去工作。

关于日常生活的人工创作故事：
早晨醒来后，我以*Pop-Tart* 早餐开启新的一天，然后淋浴，再开始工作。

关于角色（衣着讲究的男士）的人工创作故事：
醒来后，我先通过淋浴 和使用马桶来打理个人卫生。之后，我换上合适的衣服，再吃早餐。然后我离开家，开始一天的工作。

ChatGPT 自动生成的故事：
我想，我会去南方的车库，穿好衣服，并检查我的钱包和钥匙，确保我带齐了当天所需的一切。离开家去工作前，我也可以喝一杯咖啡。

表1：由人类或 ChatGPT 编写的示例故事。带下划线的词语表示游戏中可以执行的动作和实体。带波浪形下划线的词语表示游戏中无效或不允许的动作或实体。

3.1 示例故事

提供一个自然语言故事作为示例，展示智能体应执行的行为。表1展示了一些关于“早上去上班”的可能故事示例。我们假设，故事由某个人提供，用于高层次地描述智能体应尝试实现的目标，或说明任务应如何完成。故事也可以由自动故事生成系统生成 [Peng等，2021b；Goldfarb-Tarrant 等，2020] 也可以通过提示工程生成——使用 ChatGPT[OpenAI，2022]等大语言模型来描述某件事的典型做法（有关 ChatGPT 提示工程的更多详情，请参阅附录 C）。

与表1中的最后一个示例类似，故事可能会省略有关环境的细节。讲述故事的人或实体，无论是人类还是大语言模型，都可能不了解智能体运行环境的确切参数。例如，在游戏“9:05”中，“咖啡”并非一个实体，但它可以出现在示例中。这是使用故事示例的一项有意优势；故事可以被视为简洁的描述，聚焦于显著细节，并假设接收者与讲述者共享常识，因此能够重建更细致的细节。

然而，示例故事的目的并非用于演示，而是生成内在奖励信号，引导强化学习智能体采取行动，使运行环境与故事中描绘的隐含世界相一致，从而“塑造”智能体的行为。在智能体开始训练之前，示例故事会被转换为一个知识图谱——称为故事知识图谱——其形式为 RDF 三元组 [Klyne，2004] of$\langle subject, relation, object\rangle$。故事知识图谱是对故事中提及的实体的显式、持久记忆。该知识图谱包含直接从示例故事文本中提取的实体和关系，以及可根据故事中的事件推断出的其他世界细节。

为构建故事知识图谱，我们在 VerbAtlas 上训练一个语义角色标注（SRL）模型 [Gildea 和 Jurafsky，2002]，[Di Fabio等人，2019] 以自动识别并标注故事的论元结构。VerbAtlas 是一种语言资源，根据句子中的动词及其用法提供语义标注

。动词是故事的重要组成部分，因为它们表达会改变故事世界状态的动作。例如，“我在厨房喝咖啡”首先由 VerbAtlas 语义角色标注处理，得到“{’verbatlas’: ’饮用’, ’description’: ’[ARG0: 你] [drink.01: 喝] [ARG1: 咖啡] [ARGM-LOC: 在厨房]}”。随后，将主题和属性作为实体，并将 VerbAtlas框架作为边，例如 ⟨你、饮用、咖啡⟩。我们还将位置和时间数据纳入知识图谱（即 ⟨你、在、厨房⟩）。训练详情见附录A.1。将动词的 VerbAtlas 框架名称纳入关系，可满足三项需求。首先，当常识效应未知时，它可充当占位符。例如，咖啡处于已被喝掉的状态。其次，它记录世界中发生过的动作——这是一个发生过这些动作的世界。第三，有些动作并不会产生关系的正向效应，例如，摧毁某物会移除物体和关系，因为该物体不复存在。

故事知识图谱记录示例故事结束时成立的正向关系，其中包含事件记录，作为隐含效应的占位内容。这种方法的一个重要特点是，它有意不记录事件的顺序。智能体将通过试错来学习是否存在顺序约束。根据环境条件，故事中的事件可能可以按不同顺序执行，也可能必须按不同顺序执行。某些事件可能完全缺失（例如，故事没有明确说明要从卧室前往厨房）。此外，有些事件可能无法实现（例如，厨房里没有咖啡）。因此，故事知识图谱提供了智能体应该遇到的关键元素，而智能体遇到并完成的元素越多，获得的奖励也越多。

3.2 强化学习智能体

我们考虑标准的强化学习设定：智能体在若干离散时间步内与文本游戏环境交互。文本环境中最先进的强化学习方法使用知识图谱，作为世界状态的外部持久记忆 [Ammanabrolu 和 Riedl, 2018年; Ammanabrolu 和 Hausknecht, 2020年; Ammanabrolu 等人, 2020b]。随着智能体探索游戏世界，一个称为世界知识图谱的知识图谱会被构建起来，并用作状态表示。强化学习智能体通过优势演员评论家算法（A2C）进行训练 [Mnih 等人,2016] 以维持一个策略 $\pi(a_t \mid s_t;\theta)$，并估计价值函数 $V(s_t;\theta_v)$。强化学习智能体以某种方式最大化游戏中的长期期望奖励，并使用相同的 n步回报组合，同时更新策略和价值函数。

在文本游戏中，动作是由词元组成的字符串。我们不会向游戏环境查询可执行动作——也就是那些能够保证产生效果的动作。不过，我们确实通过使用基于动词的模板来简化动作空间 [Ammanabrolu 和 Hausknecht, 2020]。模板由可互换的动词短语（VP）组成，后面还可选择跟上介词短语（VP PP），例如（$[drink/eat]$）和（$[apply/ask/put]$ $[on/about/down]$ ），其中 [.] 内的动词和介词都是别名。智能体生成动作时，会先对模板进行采样，然后从游戏的词汇表中采样词语来填充空白处。

随着智能体探索游戏世界，我们构建一个内部的世界知识图谱状态表示。这个知识图谱不同于故事知识图谱。按照 Ammanabrolu 等人 [2020], 的做法，我们将构建知识图谱的过程视为一项问答任务。我们在 Jeri-choQA [Hausknecht 等人，2020b] 数据集上对 ALBERT 模型 [Lan 等人， 2019] 进行微调。该数据集专为文本游戏中的问答而设计。这使模型能够回答“我带着什么？”和“我在哪里？”之类的问题。我们将答案用作当前步骤的候选顶点集合 V_t，并将问题用作关系集合 R_t。然后，我们将 V_t 和 R_t 与上一步的游戏知识图谱 G_{t-1} 结合起来，从而将游戏知识图谱更新为 G_t。更多详情请见附录A.2。

用于构建世界知识图谱 的 ALBERT-QA 技术已被证明能够提升强化学习智能体的表现，因为它是针对文本游戏的特点进行训练的 [Ammanabrolu 和 Hausknecht, 2020]。另一方面，故事知识图谱是使用 VerbAtlas语义角色标注模型构建的，因为示例故事来自不同的文本分布——预计其中关于世界细节的描述不那么冗长，而是有意留下更多隐含信息。此外，故事知识图谱只需提取少量关键细节，就能引导智能体采取特定行动并前往环境中的特定位置。

综合来看，在训练的每一步，游戏环境都会返回一个总得分 R_t 和一条自然语言观测 o_t，其中包含 $\left(o_{t_{\text{desc}}},\ o_{t_{\text{game}}},\ o_{t_{\text{inv}}},\ a_{t-1}\right)$，分别对应房间描述、游戏反馈、物品栏和上一步行动。图2左侧展示了一个示例。如第3.2节所述，世界知识图谱 G_t 也会在时间步 t 更新。 $o_t \in \mathbb{R}^{d_o \times c}$ 的每个组成部分都经过基于GRU的编码器处理，以获得$\mathbf{o}_t$；世界知识图谱 G_t则通过图注意力网络（GATs） [Veli$\check{\text{c}}$kovi$\acute{\text{c}}$ 等, 2017] 进行处理，然后经过一个线性层，得到图表示$\mathbf{g}_\mathbf{t} \in \mathbb{R}^{d_o}$； c 表示 $\boldsymbol{o}_\mathrm{t}$的组成部分数量。随后，我们计算 α $\boldsymbol{o}_\mathrm{t}$ 与$\mathbf{g_t}$之间的注意力，

$$\boldsymbol{\alpha} = \operatorname{softmax}\left(\boldsymbol{W}_\mathrm{l}\boldsymbol{h} + \boldsymbol{b}_\mathrm{l}\right) \tag{1}$$

$$\boldsymbol{h} = \tanh\left(\boldsymbol{W}_\mathrm{o}\boldsymbol{o}_\mathrm{t} \oplus \left(\boldsymbol{W}_\mathrm{g}\mathbf{g}_\mathrm{t} + \boldsymbol{b}_\mathrm{g}\right)\right) \tag{2}$$

其中， $\oplus$ 表示矩阵与向量相加，$\odot$ 表示点积。 $\boldsymbol{W}_\mathrm{l} \in \mathbb{R}^{d_o \times d_o}$、 $\boldsymbol{W}_\mathrm{g} \in \mathbb{R}^{d_o \times d_o}$、$\boldsymbol{W}_\mathrm{o} \in \mathbb{R}^{d_o \times d_o}$ 是权重， $\boldsymbol{b}_\mathrm{l} \in \mathbb{R}^{d_o}$、 $\boldsymbol{b}_\mathrm{o} \in \mathbb{R}^{d_o}$ 是偏置。最后，整体表示向量$\mathbf{v_t}$ 通过$\mathbf{v_t} = \mathbf{g_t} + \sum_i^c \boldsymbol{\alpha}_i \odot \mathbf{o}_{\mathrm{t},i}$进行更新

3.3 奖励智能体

得到上述整体表示$\mathbf{v_t}$ 后，我们将两种内在奖励纳入强化学习智能体的训练中，以激励其采取更贴近故事隐含世界的行动。

知识图谱内在奖励由比较智能体的世界知识图谱与故事知识图谱之间的相似度来确定。这

反映了智能体的动作在多大程度上使现实世界与故事所设定的世界保持一致。当新的边加入世界知识图谱时，$\mathbf{G}_t$，我们会验证相应的三元组是否已存在于故事知识图谱中。每发现一个新的三元组，都会获得正向内在奖励 $r_t^{\mathrm{s}} = n \times \rho > 0$，其中 n 是相同三元组的数量。示例见图2（知识图谱中的红色边）。例如，智能体执行了一个与故事的隐含世界相符的动作，例如“饮用咖啡”，此时会向世界知识图谱添加一条新边，⟨You, DRINK, coffee⟩。该三元组被判定为与故事知识图谱中的一个三元组完全相同，因此智能体在此步骤获得知识图谱内在奖励$r_t^{\mathrm{s}} = 1 \times \rho$。边⟨You, in, kitchen⟩（知识图谱中的绿色边）不计入此处，因为我们只考虑在当前游戏回合中添加的新边。

受 Ammanabrolu 等人 [2020b], 的启发，我们还通过定义一种探索内在奖励 $r_t^{\mathrm{e}} = \Delta(\mathbf{G}^{\text{global}} - \mathbf{G}_t)$，鼓励智能体探索更多位置和场景，其中$\mathbf{G}_{\text{global}} = \bigcup_{i=1}^{t-1} \mathbf{G}_i$ 是智能体曾经纳入其知识图谱的所有边的集合。当智能体学到有关世界的更多信息时，其世界知识图谱的规模也会扩大，从而提高获得奖励和成功的可能性。

在时间步 t 获得的总体内在奖励为：

$$r'_t = r_t + \alpha \times r_t^{\mathrm{s}} + \beta \times r_t^{\mathrm{e}} \tag{3}$$

其中，α 和 β 是缩放因子；r_t 是游戏得分；r'_t是时间步中提供给智能体的奖励。其余训练方法与Ammanabrolu等人 [2020b]的工作相同。

4 实验

我们分三个阶段进行了实验：
1. 我们训练智能体玩文本游戏，在这些游戏中，智能体必须成功完成一些规范性日常惯例。我们向智能体提供关于这些任务的不同示例故事。在这组实验中，我们评估故事塑造能否促进常识与社会规范知识的表达（第4.1节）。2. 我们训练智能体玩开放式角色扮演游戏，并使用不同的示例任务塑造不同的角色。在这组实验中，我们评估故事塑造能否以人类可以辨识的方式塑造智能体行为。这也表明，故事塑造能够学习套路知识，即特定于不同叙事惯例的知识（第 4.2 节）。3. 我们训练智能体玩这样的游戏：示例故事与运行环境存在差异，要么提及了不存在的物体，要么提及了无法执行的动作。这组实验评估故事塑造能否适应故事提供者与智能体环境之间的不同假设。这些差异可能由多种原因造成，其中一种是故事由另一个系统（如 ChatGPT [OpenAI, 2022] ）生成（第4.3节）。

基线 我们分别实现了下列强化学习智能体的有故事塑造和无故事塑造版本：

- **Q*BERT** [Ammanabrolu等人，2020b], 一种面向 Jericho 游戏和 TextWorld 设计的先进文本游戏强化学习智能体。它通过使用 ALBERT-QA 回答问题来构建其世界知识图谱。

- **KG-A2C** [Ammanabrolu 和 Hausknecht， 2020], 该方法使用斯坦福的开放信息抽取（Ope- nIE） [Angeli 等人， 2015] 来构建其世界知识图谱，因为 ALBERT-QA 是在 JerichoQA 数据集上微调的。KG-A2C 用于 LIGHT 实验。

4.1 常识与社会规范知识

我们首先试图了解故事塑造是否能增强强化学习智能体对常识和社会规范知识的表达。

游戏。我们在两个游戏平台上实现了三款日常生活类文本游戏：

- *9:05*：一款要求智能体顺利完成上班前日常准备流程的游戏，在 Jericho 中实现 [Hausknecht 等人， 2020b], 这是一个用于与文本游戏交互的框架，充当连接学习型智能体与互动小说游戏的接口。

- 购物：一款要求智能体成功购买衣服的游戏，基于 TextWorld 开发 [Cˆot´e 等人， 2018], 这是一款开源、可扩展的引擎，既能生成文本游戏，也能模拟文本游戏。

- *See Doctor*：一款智能体会生病并必须寻求医疗救治的游戏，也由 TextWorld 开发。

9:05 的详情见附录 B.1。购物 和 *See Doctor*的详情见附录 B.2。

训练。 我们为每款游戏训练两个智能体。*Q*BERT-S*使用故事塑造，其故事由人类撰写。基线模型 *Q*BERT* 与之相同，但不采用故事塑造。我们在 20 个随机种子下运行测试游戏，以评估这两个训练后的智能体。训练详情见附录 A.3。

自动评估。 我们通过以下方式自动评估智能体对常识和社会规范知识的表达：

- **胜率**：训练后的智能体在 20 个随机种子下进行测试游戏时的胜率。
- **平均步数**：每个智能体赢得游戏所需的平均步数。超过 50 步后，游戏将自动结束。

- **平均常识得分**：一次测试试验中累积的内在奖励总量；得分越高，表示智能体采取了更多体现常识和社会规范知识的动作。关于该测试游戏环境的详细信息，请参见附录 A.4。
- **平均游戏得分**：测试游戏中每个智能体的平均得分，反映了智能体在多大程度上接近获胜条件，而不考虑如何到达最远点。

Game	智能体	胜利 %	平均步数	平均游戏得分	最大游戏得分	平均常识得分	最大社会得分	比率
9:05	Q*BERT-S	100	16.30	**3.90**	4	5.00		5
	Q*BERT	100	7.25	0.40		5.00		
购物	Q*BERT-S	100	12.35	**3.70**	4	5.00		5
	Q*BERT	100	6.30	0.90		5.00		
医生	Q*BERT-S	95	19.15	**6.70**	8	4.75		5
	Q*BERT	95	14.30	0.70		4.75		

表2：在相同的游戏环境下训练各系统，对 Q*BERT-S 与基线 Q*BERT 进行 20 次独立运行的自动评估结果。

Game	常识／社会规范 塑造 %	基础 %	平局 %	理解能力 塑造 %	基础 %	平局 %
9:05	**63.63***	9.09	27.27	36.36	**45.45**	18.18
购物	**66.67****	8.33	25.00	**41.67**	33.33	25.00
医生	**53.85**	23.08	23.08	**46.15**	38.46	15.38

表3：人工评估结果，显示偏好故事塑造的 Q*BERT-S 而非基线 Q*BERT，或认为两个系统相当的参与者百分比。每个系统均在相同的游戏环境下训练。* 表示人工评估结果在 $p < 0.05$ 置信水平下具有显著性；** 表示在 $p < 0.01$ 置信水平下，使用威尔科克森符号检验对胜负配对进行检验所得的结果具有显著性。

结果如表2所示。启用和未启用故事塑造的智能体，其胜率和平均游戏得分完全相同。然而，我们的智能体的“平均常识得分”显著高于基线智能体，表明故事塑造智能体在展现常识和社会规范知识方面能力更强。我们的智能体的“平均步数”也更高，这表明它在赢得游戏前会采取更多动作，进一步说明它并未寻求尽可能短的轨迹。

人工评估。 我们通过 Cloud Research 研究平台和亚马逊机械土耳其招募了 30 名参与者 [Lit- man 等， 2017]。我们筛选出通常不熟悉文本游戏的参与者。每位参与者先阅读一个随机选定游戏的获胜目标，然后阅读一对游戏文字记录，分别由 Q*BERT-S 和 Q*BERT 完成。每份记录都包含游戏观察内容及相应动作。随后，我们向参与者提供以下指标，并请他们选择在该指标上更偏好哪份游戏文字记录：

- 这一系列动作在动作选择上体现了更多常识思维（以及社会规范知识）。
- 考虑到你对目标的了解，这一系列动作让你明白智能体为何会采取这些动作。

参与者必须为每次比较中的选择提供详细说明，至少包含 50 个字符的自由文本。我们人工审查了回答，并剔除了未提供详细说明的参与者的数据。每对游戏轨迹至少由 10 位参与者进行评估。我们的研究已获机构审查委员会批准，并向参与者支付了相当于每小时 15 美元的报酬。人工评估问卷示例见附录 D.1

表3展示了各系统生成的故事在每项指标上受到偏好的比例。在相同的游戏环境中，Q*BERT-S 在“常识与社会规范意识”这一维度上的表现显著优于 Q*BERT。我们可以得出结论，故事塑造显著促进了训练后的强化学习智能体表达常识与社会规范知识。在“理解度”这一维度上，我们预期经故事塑造的智能体至少不会比基线更难理解，这表明更长的轨迹并非随机生成。Q*BERT-S 的结果与 Q*BERT 相当。偏好 Q*BERT 的参与者提到，他们觉得较短的游戏路径更容易理解。如表2所示，Q*BERT 完成游戏所需的步数更少，因此人类参与者更容易跟进。由此可见，我们的系统在提升智能体表达常识与社会规范知识的同时，也保留了可理解性。

4.2 角色理解

我们评估故事塑造能否以人类可辨识的方式塑造智能体的行为。我们在 LIGHT [Urbanek 等人,2019] 环境中开发了一款角色扮演游戏。LIGHT 是一个大规模奇幻文本冒险游戏研究平台，用于训练既能交谈又能行动的智能体，使其能够与其他模型或人类互动。我们为四种角色设定提供了不同的示例任务：小偷、懒汉、冒险者和暴徒。我们提供了一个 LIGHT 世界，其中包含丰富的地点和物品，供所有角色设定使用或忽略。关于该游戏的详细信息见附录 B.3。在这些实验中，我们比较了两个采用故事塑造的 KG-A2C 智能体，它们使用了不同的人类撰写的故事。

我们另外招募了 29 名参与者。参与者将阅读一个获胜目标，并获知智能体的角色设定（小偷、冒险者、暴徒、懒汉），随后阅读两个采用故事塑造的 KG-A2C 版本所进行的游戏对话记录。其中一个版本使用与给定角色设定相关的故事进行训练，另一个则使用从其余角色设定中随机选取的故事进行训练。参与者需要指出两个智能体中哪一个具有给定的角色设定。每场游戏至少由 10 名参与者进行评估。关于这项人工评估调查的更多详情见附录 D.2。

表 4 展示了参与者选择智能体针对给定角色生成的游戏文字记录（采用示例故事）而非随机选取的示例故事的百分比。该百分比反映了故事塑造在塑造智能体行为方面的有效性，使其行为能够被人类识别。结果表明，采用故事塑造的强化学习智能体能够理解适用于不同角色的示例任务，并生成符合给定角色的动作。此外，所有采用故事塑造的 KG-A2C 版本的胜率均为 100%，无论使用的是哪个示例故事。这证明了故事塑造的能力，能够

Game	已知角色	参与者选择 正确率	错误率	无法判断率	Win Rate
LIGHT	小偷	**58.06***	19.35	22.58	100
	冒险者	**62.50***	15.62	21.88	100
	Thug	**72.73****	12.12	15.15	100
	Bum	**64.71***	23.53	11.76	100

表 4：当向人类参与者提供游戏目标和角色时，偏好采用故事塑造并使用给定角色特定故事的智能体、采用故事塑造但使用其他角色故事的智能体，或认为两个系统表现相同的参与者所占百分比。“胜率”指训练后的智能体在 20 个随机种子下，采用故事塑造并使用给定角色特定故事时，在测试游戏中的胜率。用于表示显著性的符号（* 和 **）与表3中的相同。

游戏角色设定		常识/社会 GPT%	基础%	领带%	理解能力 GPT%	基础%	领带%
9:05	-	**71.43***	14.29	14.29	**57.14**	42.86	0.00
商店。	-	**62.50***	25.00	12.50	**37.50**	25.00	**37.50**
LIGHT	小偷	**57.14**	28.57	14.29	**42.86**	28.57	28.57
	Adv	**83.33****	16.67	0.00	**50.00**	16.67	33.33
	Thug	**64.71***	11.76	23.53	**52.94**	35.29	11.76

表5：偏好采用故事塑造、使用 ChatGPT 生成故事的智能体而非基线的参与者所占百分比，或认为这些系统无法区分的参与者所占百分比。符号（* 和 **）用于表示显著性，规则与表3相同。

以人类能够识别的方式塑造智能体行为，同时不损失任何性能。

4.3 鲁棒性

我们的最后一项实验考察了故事塑造在示例故事提供者与智能体环境之间的假设存在差异时的鲁棒性。我们不再使用人类根据对环境的理解设计的故事，而是采用 ChatGPT 自动生成的故事。值得注意的是，ChatGPT 对特定游戏世界并不熟悉，因此这些示例故事与智能体的游戏环境存在显著差异，例如提及不存在的物体，或提供不可用的动作。

我们向 ChatGPT 提供第一个游戏状态的描述，以及可选的角色个性，以生成引导故事；随后可将其转换为知识图谱三元组格式，用于引导智能体。接着，我们按照第 4.1 节所述的训练方法进行复现。生成的描述及相应三元组的更多示例见附录C。我们采用相同的评估流程，并在一个众包平台招募了 34 名参与者，让他们回答与第 4.1 节相同的问题。每对游戏轨迹至少由 10 名参与者进行评估。

表5展示了各系统生成的故事在各项指标上的偏好百分比。较高的偏好百分比表明，即使使用自动生成的故事，故事塑造也能引导强化学习智能体表达更多常

游戏：购物

第一次观察：我站在购物中心前面。南边有一家咖啡馆，而东边有一条路通往购物中心。

目标：购买衣服。

来自 ChatGPT 的示例故事：T o buy cloth是的，你进入了h购物中心，然后前往那家出售您感兴趣的服装的商店。许多购物中心都设有目录列表。位于入口附近，可以帮助您找到特定的商店。找到商店后，您可以浏览并试穿服装，然后在收银台付款。

Q*BERT-Story Shaping 生成的动作序列示例：

位置：街道	动作：向东走
位置：购物中心	动作：向北走
位置：商店	动作：查看衣服；试穿衣服
	动作：把钱交给员工
	动作：购买衣服；拿取衣服

Q*BERT 的示例动作序列：

位置：街道	行动：向东走
位置：购物中心	行动：向北走
位置：商店	行动：把钱给员工
	行动：购买衣服；拿衣服

表6：游戏购物中的动作序列示例。带下划线的词表示在游戏环境中可以执行的动作。带波浪线下划线的词表示在游戏中无效或不允许执行的动作。

识和社会规范知识，尽管自动生成故事的知识图谱中有些信息无法获取。例如，在表6中，ChatGPT 生成的示例故事包含一个游戏中不存在的实体——“目录列表”。尽管存在这种不匹配，我们的技术仍具有所需的灵活性，使智能体能够利用示例故事中的其他元素，并通过试错补足其余部分，从而成功完成任务。在所有测试游戏环境中，智能体的游戏完成率均为100%。

5 结论

故事塑造是应对奖励设计挑战的一种直接方法：我们不仅希望在智能体完成任务时给予奖励，还希望根据它如何完成任务来给予奖励。这可能意味着使智能体的行为符合人类偏好和期望，教会智能体常识推理和社会规范，或在游戏中塑造角色人格。

我们的方法允许用户提供一个高层示例故事，智能体可从中自动提取有关重要物体、地点和动作的知识。随后，当遇到这些物体和地点，或执行这些动作时，智能体便会给予自身奖励。由于故事是高层抽象，故事塑造能够补足缺失的细节，并且在故事无法按原样精确实现的情况下依然稳健。我们已经证明，故事塑造对智能体行为具有显著且可由人观察到的影响，同时不会损害任务完成。

参考文献

[Ammanabrolu 和 Hausknecht， 2020] Prithviraj Ammanabrolu 和 Matthew Hausknecht。面向自然语言动作空间的图约束强化学习。*arXiv* 预印本 *arXiv:2001.08837*，2020年。

[Ammanabrolu和Riedl， 2018] Prithviraj Ammanabrolu和Mark O Riedl。使用基于图的深度强化学习进行文字冒险游戏。*arXiv*预印本 *arXiv:1812.01628*，2018年。

[Ammanabrolu等，2020a] PrithvirajAmmanabrolu、Wes- ley Cheung、Dan Tu、William Broniec和Mark Riedl。让故事栩栩如生：生成互动小说世界。载于人工智能促进协会人工智能与互动数字娱乐会议论文集，第16卷，第3–9页，2020年。

[Ammanabrolu等，2020b] Prithviraj Ammanabrolu、Ethan Tien、Matthew Hausknecht和Mark O Riedl。如何避免被格鲁怪吃掉：文本世界的结构化探索策略。*arXiv*预印本 *arXiv:2006.07409*，2020年。

[Ammanabrolu等， 2021] Prithviraj Ammanabrolu、Jack Urbanek、Margaret Li、Arthur Szlam、Tim Rockt¨aschel 和 Jason Weston。如何激励你的巨龙：教会目标驱动智能体在奇幻世界中说话和行动。载于*2021*年北美计算语言学协会分会年会论文集：人类语言技术，第807–833页，线上举行，2021年6月。计算语言学协会。

[Ammanabrolu等， 2022] Prithviraj Ammanabrolu、Liwei Jiang、Maarten Sap、Hannaneh Hajishirzi 和 Yejin Choi。让交互式叙事与社会规范和价值观保持一致。*arXiv* 预印本 *arXiv:2205.01975*，2022年。

[Angeli等， 2015] Gabor Angeli、Melvin Jose Johnson Premkumar 和 Christopher D Manning。利用语言结构进行开放域信息抽取。收录于第*53*届计算语言学协会年会暨第7届自然语言处理国际联合会议论文集（第*1*卷：长篇论文），第344–354页，2015年。

[Barto， 2013] Andrew G Barto。内在动机与强化学习。收录于 自然与人工系统中的内在动机学习，第17–47页。施普林格，2013年。

[Burda等， 2018] Yuri Burda、Harri Edwards、Deepak Pathak、AmosStorkey、TrevorDarrell 和 Alexei A Efros。好奇心驱动学习的大规模研究。*arXiv* 预印本 *arXiv:1808.04355*，2018年。

[Cˆot´e等， 2018] Marc-Alexandre Cˆot´e、Akos K´ad´ar、Xingdi Yuan、Ben Kybartas、Tavian Barnes、Emery Fine、James Moore、Matthew Hausknecht、Layla El Asri、Mah- moud Adada 等。Textworld：一种用于文本游戏的学习环境。载于计算机游戏研讨会，第 41–75 页。施普林格，2018年。

[Dambekodi等， 2020] Sahith Dambekodi、Spencer Fra- zier、Prithviraj Ammanabrolu 和 Mark O Riedl。运用常识玩文本游戏。载于*NeurIPS Wordplay* 研讨会论文集，2020年。[Devlin 等， 2019] Jacob Devlin、Ming-Wei Chang、Ken- ton Lee 和 Kristina Toutanova。BERT：用于语言理解的深层双向 Transformer 预训练。载于计算语言学协会北美分会 *2019*年年会：人类语言技术论文集，第 *1* 卷（长短论文），第 4171–4186 页，2019年。

[Di Fabio 等, 2019] Andrea Di Fabio、Simone Conia 和 Roberto Navigli。VerbAtlas：一种新型大规模动词语义资源及其在语义角色标注中的应用。载于《*2019*年自然语言处理经验方法会议暨第九届国际自然语言处理联合会议（*EMNLP-IJCNLP*）论文集》，第 627–637 页，2019年。

[Frazier 等, 2020] Spencer Frazier、Md Sultan Al Nahian、Mark O. Riedl 和 Brent Harrison。从故事中学习规范：价值观对齐的智能体的一种先验。载于《*AAAI/ACM* 人工智能、伦理与社会会议论文集》，2020年。

[Gildea 和 Jurafsky, 2002] Daniel Gildea 和 Daniel Jurafsky。语义角色的自动标注。计算语言学，28(3)：245–288，2002年。

[Goldfarb-Tarrant 等， 2020] Seraphina Goldfarb-Tarrant、Tuhin Chakrabarty、Ralph Weischedel 和 NanyunPeng。采用亚里士多德式重评分进行神经故事生成的内容规划。*arXiv* 预印本 *arXiv:2009.09870*，2020年。

[Harrison 和 Riedl， 2016] Brent Harrison 和 Mark O Riedl。迈向从故事中学习：一种互动式机器学习方法。发表于第三十届人工智能促进协会人工智能会议研讨会，2016年。

[Hausknecht等，2020a] Matthew Hausknecht、Prithviraj Ammanabrolu、Marc-Alexandre Cˆot´e 和 Xingdi Yuan。互动小说游戏：一场巨型冒险。发表于第三十四届人工智能促进协会人工智能会议（人工智能促进协会），2020年。

[Hausknecht等，2020b] Matthew Hausknecht、Prithviraj Ammanabrolu、Marc-Alexandre Cˆot´e 和 Xingdi Yuan。互动小说游戏：一场巨型冒险。发表于人工智能促进协会人工智能会议论文集，第34卷，第7903–7910页，2020年。

[Hedlund 等人, 2002] J. Hedlund、J. Antonakis 和 R.J. Sternberg。缄默知识与实践智力：理解经验教训（*ARI*研究报告*2003-04*）。美国陆军行为与社会科学研究所，华盛顿哥伦比亚特区，2002年。

[Kim等， 2019] Youngjin Kim、Wontae Nam、Hyunwoo Kim、Ji-Hoon Kim 和 Gunhee Kim。《好奇心瓶颈：通过提炼任务特定新颖性进行探索》。载于国际机器学习大会，第 3379–3388 页。PMLR，2019年。

[Klyne, 2004] Graham Klyne。《资源描述框架（资源描述框架）：概念与抽象语法》。*http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/*，2004年。

[Lan等人, 2019] Zhenzhong Lan、Mingda Chen、Sebastian Goodman、Kevin Gimpel、Piyush Sharma 和 Radu Sori-cut。《ALBERT：用于语言表征自监督学习的轻量级 BERT》。*arXiv* 预印本 *arXiv:1909.11942*，2019年。[Levy 等人, 2018] Andrew Levy、Robert Platt 和 Kate Saenko。《利用事后经验进行分层强化学习》。*arXiv* 预印本 *arXiv:1805.08180*，2018年。[Litman 等人, 2017] Leib Litman、Jonathan Robinson 和 Tzvi Abberbock。Turkprime.com：面向行为科学的多功能众包数据采集平台。行为研究方法，49(2):433–442，2017年。[Mnih 等人, 2016] Volodymyr Mnih、Adria Puigdomenech Badia、Mehdi Mirza、Alex Graves、Timothy Lillicrap、Tim Harley、David Silver 和 Koray Kavukcuoglu。《深度强化学习的异步方法》。发表于国际机器学习会议，页码 1928–1937。PMLR，2016年。

[Nachum 等人, 2018] Ofir Nachum、Shixiang Shane Gu、Honglak Lee 和 Sergey Levine。高数据效率的分层强化学习。神经信息处理系统进展，31，2018年。

[Nahian 等人, 2021] Md Sultan Al Nahian、Spencer Frazier、BrentHarrison 和 Mark O. Riedl。使用规范性先验训练价值对齐的强化学习智能体。*arXiv:2104.09469*，2021。

[Nair等人, 2018] Ashvin V Nair、Vitchyr Pong、Murtaza Dalal、Shikhar Bahl、Steven Lin 和 Sergey Levine。利用想象目标进行视觉强化学习。神经信息处理系统进展，31，2018年。

[OpenAI, 2022] OpenAI。ChatGPT：一款大规模开放域聊天机器人。https://openai.com/blog/chatgpt/，2022年。

[Oudeyer 等人, 2007] Pierre-Yves Oudeyer、Frdric Kaplan 和 Verena V Hafner。用于自主心智发展的内在动机系统。*IEEE* 进化计算汇刊，11(2):265–286，2007年。

[Palmer等人, 2005] Martha Palmer、Daniel Gildea 和 Paul Kingsbury。命题库：一个语义角色标注语料库。计算语言学，31(1):71–106，2005年。

[Peng 等人，2021a] Xiangyu Peng、Mark O Riedl 和 Prithviraj Ammanabrolu。自然语言中的内在可解释强化学习。*arXiv* 预印本 *arXiv:2112.08907*，2021年。

[Peng 等人，2021b] Xiangyu Peng、Kaige Xie、Amal Alabdulkarim、Harshith Kayam、Samihan Dani 和 Mark O Riedl。利用读者模型引导神经故事生成。*arXiv* 预印本 *arXiv:2112.08596*，2021年。

[Pong 等人，2019] Vitchyr H Pong、Murtaza Dalal、Steven Lin、Ashvin Nair、Shikhar Bahl 和 Sergey Levine。偏斜拟合：覆盖状态的自监督强化学习。*arXiv*预印本 *arXiv:1903.03698*，2019年。

[Rajpurkar 等人, 2018] Pranav Rajpurkar、Robin Jia 和 Percy Liang。《知道自己不知道什么：SQuAD 中无法回答的问题》。载于《第56届计算语言学协会年会论文集》（第2卷：短文），第784–789页，澳大利亚墨尔本，2018年7月。计算语言学协会。[Schmidhuber, 1991] J¨urgen Schmidhuber。《在模型构建神经控制器中实现好奇心和无聊的一种可能性》。载于《自适应行为模拟国际会议论文集：从动物到人工生命体》，第222–227页，1991年。[Shi 和 Lin, 2019] Peng Shi 和 Jimmy Lin。《用于关系抽取和语义角色标注的简易 BERT 模型》。arXiv 预印本 arXiv:1904.05255，2019年。[Shridhar 等人, 2021] Mohit Shridhar、Xingdi Yuan、Marc-Alexandre Cˆot´e、Yonatan Bisk、Adam Trischler 和 Matthew Hausknecht。《ALFWorld：为交互式学习对齐文本与具身环境》。载于《国际学习表征会议（ICLR）论文集》，2021年。[Stadie 等人, 2015] Bradly C Stadie、Sergey Levine 和 Pieter Abbeel。《利用深度预测模型激励强化学习中的探索》。arXiv 预印本 arXiv:1507.00814，2015年。[Urbanek 等人, 2019] Jack Urbanek、Angela Fan、Siddharth Karamcheti、Saachi Jain、Samuel Humeau、Emily Dinan、Tim Rockt¨aschel、Douwe Kiela、Arthur Szlam 和 Jason Weston。《在奇幻文本冒险游戏中学习对话与行动》。arXiv 预印本 arXiv:1903.03094，2019年。[Veli˘ckovi´c 等人, 2017] Petar Veli˘ckovi´c、Guillem Cucurull、Arantxa Casanova、Adriana Romero、Pietro Lio 和 Yoshua Bengio。《图注意力网络》。arXiv 预印本 arXiv:1710.10903，2017年。[Vezhnevets 等人, 2017] Alexander Sasha Vezhnevets、Si- mon Osindero、Tom Schaul、Nicolas Heess、Max Jader- berg、David Silver 和 Koray Kavukcuoglu。《用于分层强化学习的封建网络》。载于《国际机器学习会议》，第3540–3549页。PMLR，2017年。[Wang 等人, 2022] Ruoyao Wang、Peter Jansen、Marc- Alexandre Cˆot´e 和 Prithviraj Ammanabrolu。《科学世界：你的智能体是否比五年级学生更聪明？》，2022年。[Xu 等人, 2020] Yunqiu Xu、Meng Fang、Ling Chen、Yali Du、Joey Tianyi Zhou 和 Chengqi Zhang。《通过堆叠式分层注意力，在文本游戏中进行深度强化学习》。载于《神经信息处理系统进展》第33卷，2020年。

A 实现细节

A.1 使用 VerbAtlas 进行语义角色标注

语义角色标注模型会自动识别故事中的论元结构并为其添加标签。例如，它会提取‘verbatlas’：‘EXIST LIVE’、‘args words’：{‘Theme’：‘Jenny’，

“属性”：‘Georgia’}来自“*Jenny livedin Georgia*”。故事中的动词将以 VerbAtlas 框架表示。例如，“live”表示为“EXIST LIVE”。

我们使用经过微调的 Transformer 模型进行语义角色标注（SRL）。该模型是一个 BERT [Devlin 等人模型，在线性分类层上使用 Ontonotes 5.0 数据集训练，用于预测 PropBank[Palmer 等人的 SRL。该模型由 Shi [2019], 提出，目前是英语 SRL 领域的最先进模型。我们使用一个开源实现[1]，该实现基于官方的 AllenNLP BERT-SRL 模型[2]。训练时使用以下超参数：

- 批次大小：32
- 输入嵌入的 Dropout：0.1
- 学习率： $5e^{-5}$
- 优化器：Adam
- 总轮数：15

接着，我们使用从 Propbank 框架到 VerbAtlas [Di Fabio 等人， 2019] 类别的映射，返回正确对应的 VerbAtlas 类别，而不是 Propbank 的 [Palmer等人， 2005]。我们可以直接将 VerbAtlas 类别映射到 PropBank 框架，因为每个 VerbAtlas 类别都只有一个 PropBank 框架。这使我们能够使用 VerbAtlas 中丰富的内容，同时沿用最初用于预测 PropBank 的同一模型。

A.2 知识图谱表示问答模型

基于 ALBERT 的问答网络 [Lan 等人， 2019] 使用了原论文中列出的超参数。这些超参数已被证明在 SQuAD 2.0 [Rajpurkar 等人， 2018] 数据集上效果良好。我们没有对超参数进行任何额外调优。

参数	价值
批次大小	8
学习率	3e-5
最大序列长度	512
文档步幅	128
预热步数	814
最大步数	8144
梯度累积步数	24

A.3 使用 StoryShaping 的强化学习智能体

图2所示内容的更多细节。序列动作解码器由两个相互连接的门控循环单元组成，如 Ammanabrolu [2020] 中所示。第一个门控循环单元解码动作模板，第二个则解码可填入模板的物体。这些物体受到图掩码的约束，即解码器只能选择知识图谱中已存在的实体。

[1]https://github.com/Riccorl/transformer-srl [2] https://demo.allennlp.org/semantic-role-labeling

与 Ammanabrolu [2020], 相同，损失由模板损失、对象损失、价值损失、行动者损失和熵损失组成。给定特定状态和当前网络参数后，模板损失作用于解码器。同样，对象损失作用于解码器，其计算方式是对所有对象解码步骤的交叉熵损失求和。针对有效动作的熵损失旨在防止智能体过早收敛到某一条轨迹。以下超参数取自原论文，已知它们在文本游戏中效果良好。

参数	价值
折扣因子	0.9
熵系数	0.03
价值系数	9
模板系数	3
对象系数	9

A.4 人类相似度测试游戏

为了自动评估受训强化学习智能体对常识与社会规范知识的表达，我们为每款文字游戏都开发了一种测试游戏模式。在训练期间，玩家只有赢得游戏才能获得得分。然而，在评估强化学习智能体时，我们会在测试游戏中测试基线和我们的智能体；该测试游戏的地图和动作与训练游戏相同，但计分系统不同。具体而言，测试游戏环境不仅会在玩家赢得游戏时给予得分，还会在智能体达到特定状态时给予得分。例如，在游戏 9:05 中，智能体在厨房吃掉一个 Pop-Tart 时，测试游戏环境会给予 2 分；而在训练游戏中，这一动作不会获得任何分数。我们的测试游戏模式让我们能够自动评估受训强化学习智能体是否具备类似人类的特质。我们认为，在这种游戏模式中得分越高，就表明受训强化学习智能体对常识与社会规范知识的表达程度越高。

B 游戏设计

B.1 Jericho 游戏

9:05 是一款由 Adam Cadre 创作的文字冒险游戏，面向休闲玩家，包括刚接触此类游戏的玩家。这是一款短小简单的游戏，采用分支叙事，讲述一名男子醒来后接到电话，得知自己要去上班。由于游戏中有许多玩家可以选择执行的常识性动作，因此很适合用来测试我们的智能体。在游戏中，玩家可以离开家直接去工作，从而立即赢得游戏；也可以选择花些时间做好准备，比如换衣服、淋浴和吃吐司，然后再开始一天的生活。这些动作都是可选的，不会影响游戏结果。

示例故事 我们在第4.1节中使用的内容：“早上醒来后，我以 Pop-Tart 早餐开启一天，然后洗个淋浴，再开始工作。”游戏会以游戏得分：5结束。你可以在图4中看到这个游戏的布局

我们设计了一款游戏，并提供可选的*CS*分数，用于评估智能体的类人程度。设计的 CS 分数如下：

- 得分 2：使用了淋浴。
- 得分 2：吃了 Pop-Tart。

B.2 TextWorld

我们在 TextWorld [Cˆot´e 等人的 2018] 文字游戏引擎中设计了两款游戏。

购物是一款目标为在购物中心购买服装的游戏。玩家从街道出发，可以选择途中在各家咖啡馆和餐馆停留。玩家还可能遇到能提供获取优惠券相关信息的非玩家角色。玩家可以选择多条路线前往服装店，也可以选择参与可选活动，例如购买咖啡或使用优惠券。游戏布局如图3所示。

只有玩家完成游戏后，才能获得游戏得分。当玩家的物品栏中有“衣服”时，游戏结束，游戏得分为：5。

示例故事（用于第4.1节的实验）如下：“为了省钱，我需要获得一张优惠券。试穿衣服后，我会把它们买下来。”

我们设计了一款带有可选*CS*分数 的游戏，用于评估故事塑造是否有助于表达常识和社会规范知识。为第4.1节设计的CS分数如下，

- 得分 2：使用了优惠券。
- 得分 2：试穿了衣服。

See Doctor 是一款目标为获得药物的游戏。玩家从家出发，手里有一杯热水和钱。玩家可以煮热汤、买热咖啡、去医院看医生，或去药店买药。玩家有多条路线可选，

获得药物，他们还可以执行各种可选动作，例如买一杯咖啡或饮用热水。

只有玩家完成游戏后才能获得游戏得分。当玩家的物品栏中有“药物”时，

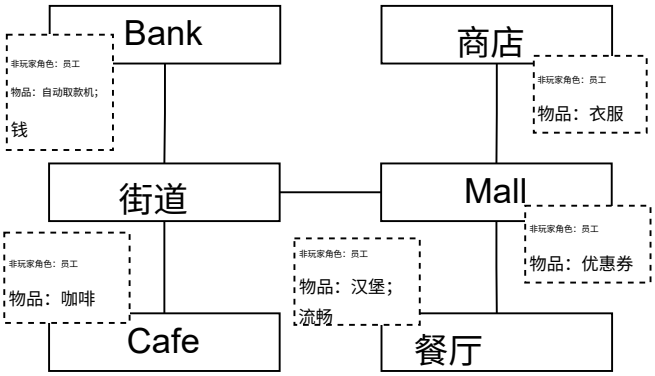

图3：“购物”游戏的布局。

实验用示例故事（第4.1节）如下：“我感冒了，喝了热水，但洗完澡后还是没有好转。我去了医院看医生，并拿到处方去买药。”

我们设计了一款带有可选*CS*分数的游戏，用于评估故事塑造是否有助于表达常识和社会规范知识。设计的*CS*分数如下：

- 得分2：饮用了水。
- 得分 2：淋浴已使用。
- 得分 2：已看到医生。
- 得分 2：处方位于物品栏中。

B.3 LIGHT

LIGHT [Urbanek 等人， 2019] 是一个用于训练能够对话和采取行动的智能体的文字冒险游戏检索平台。在我们的实验中，我们开发了一张大型地图作为沙盒。由于游戏中玩家可执行的动作种类繁多，我们的智能体可以在其中得到有效测试。这些动作包括与物体和非玩家角色互动，以及在地图中移动。所有动作均为可选项；智能体到达指定目标地点后，游戏即告结束。

Urbanek [2019] 使用生成模型和检索模型，使非玩家角色能够与玩家进行对话。不过，我们并未使用这些模型，而是将智能体与非玩家角色的互动限制为动作，因为添加对话会带来不必要的复杂性，而我们的实验并不需要这种复杂性。

我们设计了一款名为“黄金”的角色扮演游戏，玩家可以扮演不同的角色设定、收集物品，最终“前往草地”。玩家从 Simple Town 出发，可以选择前往城镇中的多个地点，例如城镇富人区、布道厅或军械库，并在这些地点收集不同的物品。玩家可以选择多条路线前往草地，也可以选择参与可选活动，例如参观军械库，或在布道厅拾取旧祈祷书。游戏布局如图 5 所示。

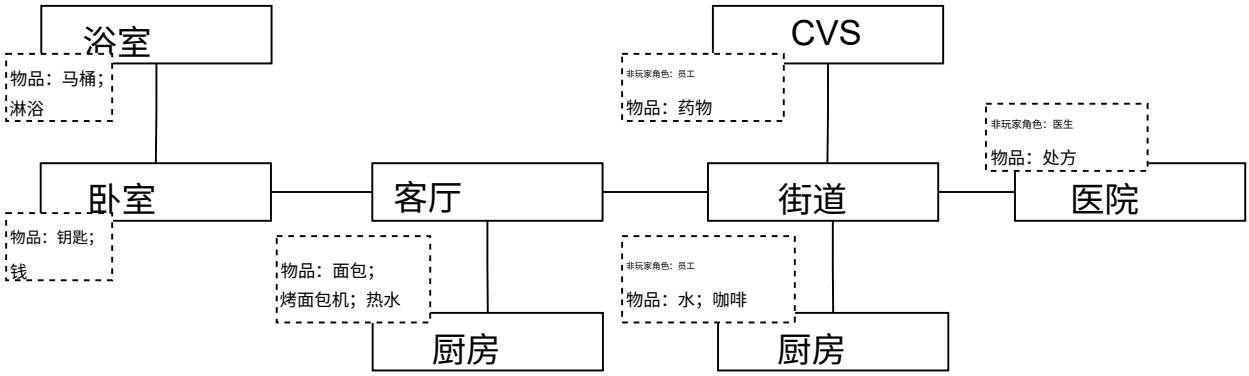

图 4：“看医生”游戏的布局。

只有玩家完成游戏后才能获得游戏得分。当玩家进入 草地时，游戏结束，游戏得分为：5。

此外，我们还证明，只要其他引导叙事能够在环境中进行，我们的系统也能够泛化到这些叙事。为此，我们提供了第四种引导叙事，它有一个不同但仍可在游戏中实现的目标。

示例故事：冒险者角色：“我是一个勇敢的冒险者。我知道有一座地牢，里面有珍贵的宝藏。我前往军械库，拿起一把剑、一面盾牌、盔甲和一张弓来保护自己。装备妥当后，我前往地牢。我拿走了黄金、珠宝、金杯和一个金高脚杯。然后我经由草地离开。”

我们设计了一个带有可选*CS*分数 的角色，用于评估故事塑造是否有助于在奇幻环境中表达与角色相符的规范。为第 4.2 节设计的 CS分数如下：

- 得分 10：获得剑。
- 得分 10：获得盔甲。
- 得分 10：获得盾牌。
- 得分 10：获得弓。
- 得分 10：获得黄金。
- 得分 10：获得珠宝。
- 得分 10：获得金杯。
- 得分10：获得金高脚杯。

示例故事 小偷角色：“我是一个胆小的小偷。我前往城镇富人区寻找贵重物品。我进入山坡庄园，拿走了那里的金条。我悄悄前往布道厅，拿走了一小袋黄金。然后我穿过草地离开。”

我们设计了一个带有可选*CS*分数 的角色，用于评估故事塑造是否有助于在奇幻环境中表达与角色相符的规范。第 4.2 节的 CS分数设计如下，

- 得分 10：获得金条。
- 得分10：获得一小袋黄金。

示例故事角色懒汉：“我是个懒汉。我希望尽可能少做事，就能拿到一些硬币然后离开。我立即离开Simple Town。离开前，我只在城镇广场停留，领取捐款，然后穿过草地离开。”

我们设计了一个带有可选*CS*分数 的角色，用于评估故事塑造是否有助于在奇幻环境中表达与角色相符的规范。第 4.2 节的 CS分数设计如下，

- 得分 10：获得捐款。

示例故事，角色暴徒：“我是一个暴徒。我想向欺骗我的钟表匠报仇。我在布道厅里找到了钟表匠。我打了钟表匠，他倒在地上死了。然后我从草地离开。”我们设计了一个带有可选*CS*分数 的角色，用于评估故事塑造是否能促进在奇幻环境中表达与角色相符的规范。为第 4.2 节设计的 CS分数如下：

- 得分5：钟表匠被击中。

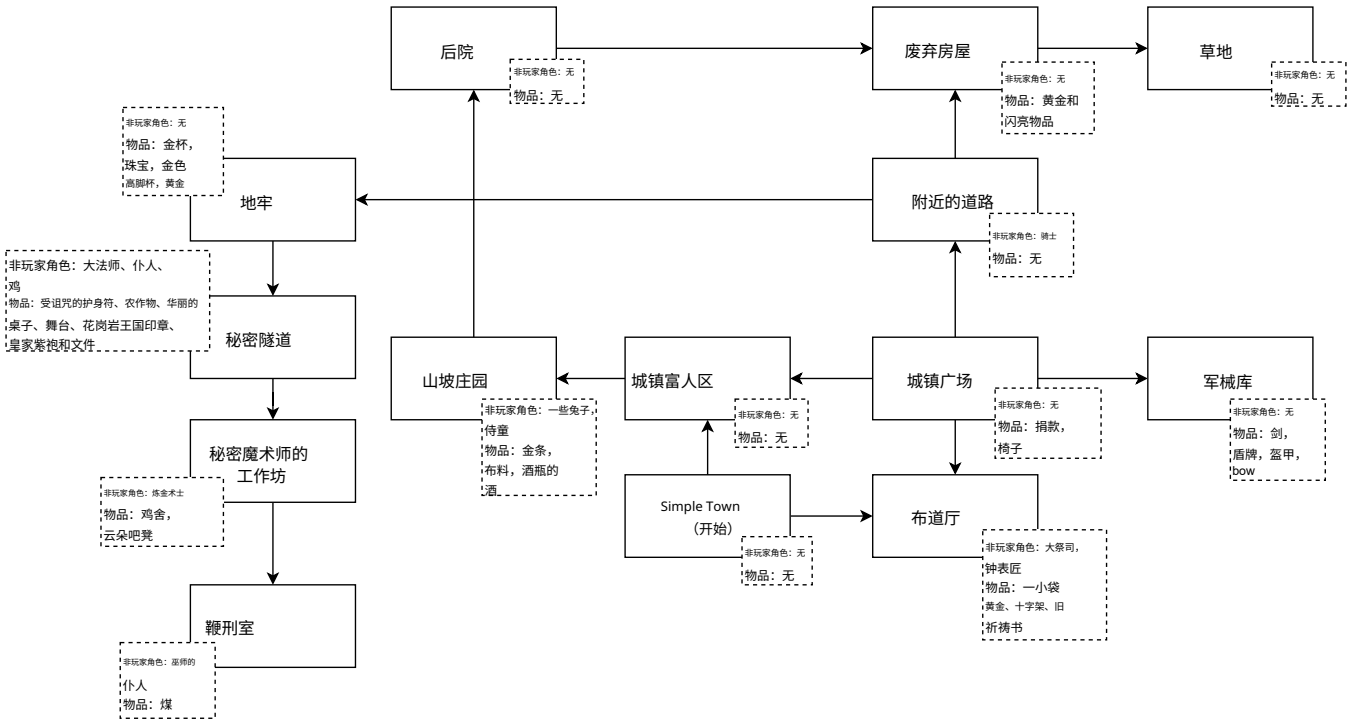

图 5：测试中使用的 LIGHT 地图布局

C ChatGPT

ChatGPT 是一种最先进的语言生成模型，由 OpenAI 开发。它基于 GPT（生成式预训练 Transformer）架构，使用机器学习生成类似人类的文本。ChatGPT 使用大量对话文本数据进行训练，能够在各种情境下针对提示词生成回应，因此可用于聊天机器人、语言翻译和文本摘要等自然语言处理任务，是一种强大的工具。

我们发现，只要提示得当，ChatGPT 就能生成 Story Shaping 所需的示例故事，因为故事可以省略环境细节。本节介绍了我们用于让 ChatGPT 自动生成故事的提示。

C.1 Jericho游戏：9:05

我们使用提示“[观察]。为了 [目标]，我会做什么？”来提示 ChatGPT。在游戏“9:05”中，我们使用的提示是：“我早上醒来。南边有一间浴室，东边有一扇门通往客厅。茶几上有一部电话、一个钱包和几把钥匙。电话响了。上班前我会做些什么？”ChatGPT 给出的 3 个结果如下：

- “我可能会去南边的浴室洗个澡，穿好衣服，并检查钱包和钥匙，确保带齐当天所需的一切。离开家去上班前，我也可能会喝一杯咖啡。”
- “去工作之前，你很可能会接电话，看看是谁打来的。你也可能会检查钱包和钥匙，确保带齐了当天所需的一切。离家前，你也可能会去一趟浴室。”
- 去上班之前，我会先接听电话，看看是谁打来的。接着，我会去浴室梳洗一番并刷牙。之后，我会考虑做些锻炼，以健康的方式开启新的一天。最后，我会拿上钥匙，出门开始新的一天。

我们使用 verbAtlas框架将这些生成的故事转换为知识图谱。然后，使用随机选取的示例故事，通过故事塑造来训练智能体。

C.2 TextWorld游戏：购物

我们使用与附录C.1相同的提示：“[观察]。为了 [GOAL]，我会做什么”。在“购物”游戏中，我们使用以下提示：“我在购物中心前面。北边有一家银行，南边有一家咖啡馆。在购买衣服之前，我会做什么？”ChatGPT的输出如下：

- “为了在购物中心买衣服，你可以进入购物中心，前往出售你感兴趣的服装的商店。许多购物中心的入口附近都设有目录列表，可以帮助你找到想去的商店。找到商店后，你可以浏览并试穿衣服，然后在收银台付款。”
- “为了购买衣服，我会先去北方的银行，查看银行账户里是否有足够的钱。如果账户资金充足，我就会

前往购物中心。进入购物中心后，我会逛逛不同的服装店，看看有没有让我感兴趣的促销活动或优惠活动。我也会留意是否有我正在寻找的特定衣服。逛过各家店后，我会决定要买哪些衣服。离开购物中心前，我可能会在南方的咖啡馆休息一下，喝杯咖啡或吃点零食。最后，我会买下衣服，然后带着新衣服回家。”

• 为了购买衣服，我会先前往北方的银行取一些现金；如果现金不足，也可以查询账户余额。之后，我会前往正前方的购物中心并进入其中。接着，我会在购物中心里四处寻找服装店。如果有需要，我可能会向购物中心的员工问路。找到服装店后，我会浏览不同的选项，并试穿任何吸引我的衣服。最后，我会购买心仪的衣服，然后前往南方的咖啡馆休息，享用一杯咖啡。

然后，我们将生成的故事转换为知识图谱，并对强化学习智能体应用故事塑造技术。

C.3 LIGHT

由于 LIGHT 环境比 Jericho 或 TextWorld 更复杂，我们使用一系列针对 LIGHT 环境的提示词，为 ChatGPT 提供有关该环境的背景信息以及最终目标。此外，由于 ChatGPT 不允许谈论暴力，因此无法为“暴徒”这一角色生成输出。基于这个原因，我们未将其纳入其中。ChatGPT 的访问日期为 2023 年 1 月 7 日。

为了提供有关环境的背景信息，我们使用以下提示——“我将描述一个游戏。请使用这段描述回答我的下一个问题。在 LIGHT 的世界里，有一座 Simple Town。东边有一座布道厅。布道厅里有一位「大祭司」、一位「钟表匠」、一「小袋黄金」、一个「十字架」和一些「旧祈祷书」。城镇富人区在北边。城镇富人区里有一座山坡庄园。山坡庄园里有一些「兔子」、一位「侍童」、「金条」、「布料」和「酒瓶」。布道厅北边是城镇广场。城镇广场里有「捐款」和「椅子」。城镇广场东边是军械库。军械库里有一把「剑」、一面「盾牌」、「盔甲」和一把「弓」。城镇广场北边有一条附近的道路。那里有一位「骑士」。附近道路西边有一座地牢。里面有「黄金」、「珠宝」、「金杯」和一个「金高脚杯」。地牢更深处有一间秘密魔法师工坊。秘密魔法师工坊里有一位「大法师」、「仆人」、「鸡」、「华丽的桌子」和「花岗岩王国印章」。穿过秘密魔法师工坊后面就是一间鞭刑室。里面有「煤」和「魔法师的仆人」。附近道路北边有一座废弃房屋。里面有「黄金和闪亮物品」。废弃房屋东边是一片草地。游戏在你到达这里时结束。所有相关的非玩家角色和物体都用引号标出。”

ChatGPT 的输出如下，

• “LIGHT 游戏的目标是什么？”

为了提供目标，我们使用以下提示：“简要描述我可以做些什么来获得黄金。”ChatGPT 的输出如下：

• “要在游戏中获得黄金，你可以前往布道厅寻找那一小袋黄金，前往城镇富人区的山坡庄园寻找金条，前往地牢寻找黄金、珠宝、金杯和一只金高脚杯，还可以前往废弃房屋寻找黄金和闪亮物品。你还可以尝试通过与布道厅里的钟表匠、附近道路上的骑士，以及秘密魔法师工坊里的大法师互动来找到黄金。”

为了提供这些角色设定，我们使用了以下提示词，

• “我是一名勇敢的冒险者。我没有武器，也没有携带任何东西。简要说明哪种方法最适合我。明确说出我会与哪些物品互动，以及我会在哪里结束游戏。”

• “我是一个胆小又贪婪的小偷。我没有武器，也没有携带任何东西。简要说明哪种方法最适合我。明确说出我会与哪些物品互动，以及我会在哪里结束游戏。”

• “我是个懒汉，手无寸铁，也身无分文。请简要说明哪种方法最适合我，并明确说出我会与哪些物品互动，以及我会在哪里结束游戏。”

ChatGPT 的输出如下：

• “作为一名勇敢的冒险者，你手无寸铁，也身无分文，最好前往军械库，获取剑、盾牌、盔甲和弓来保护自己。然后，你可以前往地牢，搜寻黄金、珠宝、金杯和金高脚杯，以积累财富。拿到这些物品后，你可以前往草地结束游戏。”

• “作为一个胆小又贪婪的小偷，在游戏中获取黄金的最佳方法很可能是前往布道厅，找到那一小袋黄金，然后前往城镇富人区，在山坡庄园里找到金条。你也可以试着与布道厅里的钟表匠互动来获取黄金，不过在这种情况下，需要与哪些物品互动才能获得黄金并未说明。”

要完成游戏，你应该避开地牢和秘密魔法师工坊，因为那里可能有危险，而且从这些区域偷窃可能会造成后果。相反，你应该直接前往废弃房屋以东的草地，因为游戏会在那里结束。”

• “对于一个懒惰且手无寸铁的玩家来说，获取黄金最合适的方法似乎是前往废弃房屋，那里有黄金和闪亮物品，无需任何特殊装备或费力就能轻松拿走。要到达废弃房屋，玩家可以从 Simple Town 出发，向北前往城镇广场。从那里继续向北走到附近的道路，然后向西转，就能到达废弃房屋。

或者，玩家也可以尝试在游戏中为非玩家角色完成任务来获取黄金。这可能需要付出一些努力，但也有可能获得更丰厚的奖励。比如，玩家可以和附近道路上的骑士交谈，看看他是否有需要完成的任务，并以此换取黄金。不过，描述中并未说明玩家如何与非玩家角色互动，也没有提到他们可能会提供什么任务。”

此外，我们还展示了 ChatGPT 能够为 LIGHT 从目标知识图谱中生成三元组。我们再次通过提示词提供有关环境的背景信息：“我将描述一个游戏。请根据这段描述回答我的下一个问题。在 LIGHT 的世界里，有一座 Simple Town。东边有一座布道厅。布道厅里有一位‘大祭司’、一位‘钟表匠’、‘一小袋黄金’、一个‘十字架’和‘旧祈祷书’。城镇富人区位于北方。城镇富人区里有一座山坡庄园。山坡庄园里有一些‘兔子’、一位‘侍童’、‘金条’、‘布料’和‘酒瓶’。布道厅以北是城镇广场。城镇广场上有‘捐款’和‘椅子’。城镇广场以东是军械库。军械库里有一把‘剑’、一面‘盾牌’、‘盔甲’和一张‘弓’。城镇广场以北有一条附近的道路。那里有一位‘骑士’。附近的道路以西是地牢。里面有‘黄金’、‘珠宝’、‘金杯’和一个‘金高脚杯’。地牢更深处有一间秘密魔法师工坊。秘密魔法师工坊里有一位‘大法师’、‘仆人’、‘鸡’、‘华丽的桌子’和‘花岗岩王国印章’。穿过秘密魔法师工坊后面就是一间鞭刑室。里面有‘煤’和‘巫师的仆人’。附近的道路以北有一座废弃房屋。里面有‘黄金和闪亮物品’。废弃房屋以东是一片草地。到达这里时游戏结束。所有相关的非玩家角色和物体都用引号括起来。”

ChatGPT 的输出如下，

- “你描述的游戏名为 ”LIGHT”，故事发生在一个名为 Simple Town 的城镇。城镇中有多个地点，包括布道厅、富人区、城镇广场、军械库、附近的道路、地牢、秘密魔法师工坊、鞭刑室和废弃房屋。每个地点都有各种非玩家角色（NPC）和物体，例如“大祭司”、“钟表匠”、“一小袋黄金”、“十字架”、“旧祈祷书”、“兔子”、“侍童”、“金条”、“布料”、“酒瓶”、“捐款”、“椅子”、“剑”、“盾牌”、“盔甲”、“弓”、“黄金”、“珠宝”、“金杯”、“金高脚杯”、“大法师”、“仆人”、“鸡”、“华丽的桌子”以及“花岗岩王国印章”。玩家到达草地后，游戏便结束。”

为了让 ChatGPT 获得以知识图谱形式表示的世界知识，我们使用以下提示：“将这个游戏转换为一个由逗号分隔的三元组组成的知识图谱，格式为 (”宾语”, ”在”, ”位置”), (”非玩家角色”, ”是”, ”存活”), (”非玩家角色2”, ”是”, ”死亡”) 或 (”你”, ”拥有”, ”宾语”)。假设所有非玩家角色初始时都处于”存活”状态。将此表示为 (”非玩家角色”, ”是”, ”存活)。”

ChatGPT 的输出如下：

- ”(”大祭司”, ”在”, ”布道厅”), (”钟表匠”, ”在”, ”布道厅”), (” 一小袋黄金 ”, ” 在 ”, ” 布道厅 ”), (” 十字架 ”, ” 在 ”, ” 布道厅”), (”旧祈祷书”, ”在”, ”布道厅”), (”大祭司”, ”是”,” 活着的”),(”钟表匠”, ”是”,”活着的”), (”兔子”, ”在”, ”山坡庄园”), (”侍童”, ”在”, ”山坡庄园”), (”金条”, ”在”, ”山坡庄园”), (”布料”, ”在”, ”山坡庄园”), (”酒瓶”, ”在”, ”山坡庄园”), (”侍童”, ”是”, ”活着的”), (”捐款”, ”在”, ”城镇广场”), (”椅子”, ”在”, ”城镇广场”), (”剑”, ”在”, ”军械库”), (”盾牌”, ”在”, ”军械库”), (”盔甲”, ”在”, ”军械库”), (”弓”, ” 在 ”, ” 军械库 ”), (” 黄金 ”, ” 在 ”, ” 地牢 ”), (” 珠宝 ”, ” 在”, ”地牢”), (”金杯”, ”在”, ”地牢”), (”金高脚杯”, ”在”, ” 地牢”), (”大法师”, ”在”, ”秘密魔法师工坊”), (”仆人”, ” 在”, ”秘密魔法师工坊”), (”鸡”, ”在”, ”秘密魔法师工坊”), (” 装饰华丽的桌子 ”, ” 在 ”, ” 秘密魔法师工坊 ”), (” 花岗岩王国印章”, ”在”, ”秘密魔法师工坊”), (”煤”, ”在”, ” 鞭刑室”), (”巫师的仆人”, ”在”, ”鞭刑室”), (”黄金和闪亮物品”, ”在”, ”废弃房屋”), (”大法师”, ”是”, ”活着的”), (” 仆人”, ”是”, ”活着的”), (”巫师的仆人”,”是”,”活着的”),(” 骑士”,”是”, ”活着的”)”

为了提示 ChatGPT 在目标知识图谱中生成三元组，我们使用以下提示词，

- ”这是一个发生在相同设定中的故事。只有当指定的宾语或非玩家角色转变为另一种状态时，三元组才会发生变化。如果发生了某个动作，就将其及其效果添加到知识图谱中。对于状态，使用（”你”，关系，宾语）。对于动作，使用（”我”，动作，宾语）。描述动作时，尽量使用最简单的词。游戏结束时，使用（”你”，”在”，位置）说明玩家所在的位置

将动作转换为最简单的形式。例如，使用”获取”而不是”抓取”。
不要包含动作是”前往”的三元组。
不要包含潜在动作。
在这个故事之后，哪些三元组会被添加到知识图谱中?

“作为一名勇敢的冒险者，你手无寸铁、一无所有，最好先去军械库领取剑、盾牌、盔甲和弓来保护自己。然后，你可以前往地牢，搜寻黄金、珠宝、金杯和金高脚杯，积累一些财富。获得这些物品后，你就可以前往草地，结束游戏。”

- ”这是一个发生在同一设定中的故事。只有当指定的宾语或非玩家角色转变为另一种状态时，三元组才会改变。如果发生了动作，就将其及其效果添加到知识图谱中。对于状态，使用（”你”，关系，宾语）。对于动作，使用（”我”，动作，宾语）。描述动作时尽量使用最简单的词。游戏结束时，用（”你”，”在”，位置）说明玩家所在的位置。

将动作转换为最简单的形式。例如，使用”拿”而不是”抓取”。
不要包含动作是“前往”的三元组。
请勿包含潜在动作。
这个故事之后，哪些三元组会被添加到知识图谱中?

“作为一个胆小又贪婪的小偷，在游戏中获取黄金的最佳方法可能是前往布道厅找到一小袋黄金，然后前往城镇富人区，在山坡庄园里找到金条。你也有可能通过与布道厅里的钟表匠互动来获得黄金，不过没有说明在这种情况下你需要用哪些物品与他互动才能获得黄金。
要完成游戏，你应该避开地牢和秘密魔法师工坊，因为这些区域可能有危险，而且在那里偷窃可能会带来后果。相反，你应该直接前往废弃房屋东边的草地，因为游戏会在那里结束。””

• ”这是一个发生在相同背景下的故事。只有当指定的宾语或非玩家角色转变到另一种状态时，才会更改三元组。如果发生了动作，就将其连同产生的效果一起添加到知识图谱中。对于状态，使用（”你”，关系，宾语）。对于动作，使用（”我”，动作，宾语）。描述动作时，尽可能使用最简单的词。游戏结束时，使用（”你”，”在”，位置）说明玩家所在的位置。

将动作转换为最简单的形式。例如，使用“获取”而不是“抓取”。
不要包含动作是“前往”的三元组。
不要包含潜在动作。
读完这个故事后，哪些三元组会被添加到知识图谱中?

“对于一个懒惰且手无寸铁的玩家来说，获取黄金最合适的方法似乎是前往废弃房屋，那里有黄金和闪亮物品，可以轻松拿走，无需任何特殊装备或付出任何努力。要到达废弃房屋，玩家可以从 Simple Town 出发，向北前往城镇广场。从那里，他们可以继续向北前往附近的道路，然后向西转，就能到达废弃房屋。

或者，玩家也可以尝试通过为游戏中的非玩家角色完成任务或委托来获取黄金。这可能需要付出一些努力，但有望获得更多奖励。例如，玩家可以和附近道路上的骑士交谈，看看他是否有需要完成的任务，并以此换取黄金。不过，描述中并未说明玩家如何与非玩家角色互动，也没有说明他们可能会提供什么任务。””

ChatGPT 的输出如下，

• “(”剑”, ”位于”, ”军械库”), (”盾牌”,”位于”, ”军械库”), (”盔甲”, ” 位于 ”, ” 军械库 ”), (” 弓 ”, ” 位于 ”, ” 军械库 ”), (” 黄金 ”, ” 位于”, ”地牢”), (”珠宝”, ”位于”, ”Dungeon”), (”金杯”, ”位于”, ”地牢”), (”金色高脚杯”, ”位于”, ”地牢”)”
• “(”一小袋黄金”, ”在”, ”Semon大厅”) (”金条”, ”在”, ”山坡庄园”)”
• “(”黄金和闪亮物品”, ”在”, ”废弃房屋”)”

D 人类研究

D.1 常识与社会规范知识

我们首先在图6中向参与者展示说明。我们向参与者提出一组筛选问题，以确保他们理解我们的任务。详情请见图7和图8。我们使用 Cloud Research 众包平台开展研究，并通过该平台与亚马逊机械土耳其互动 [Litman等人，2017]。参与者必须至少拥有学士学位，且英语为母语，才能参加本研究。参与者需要通过筛选问题，然后用超过50个字符说明他们对本研究中每个选项的偏好。这有助于筛除低质量回答，并确保研究的有效性。本研究已获机构审查委员会批准，我们按每小时15美元的等值金额向参与者支付报酬。

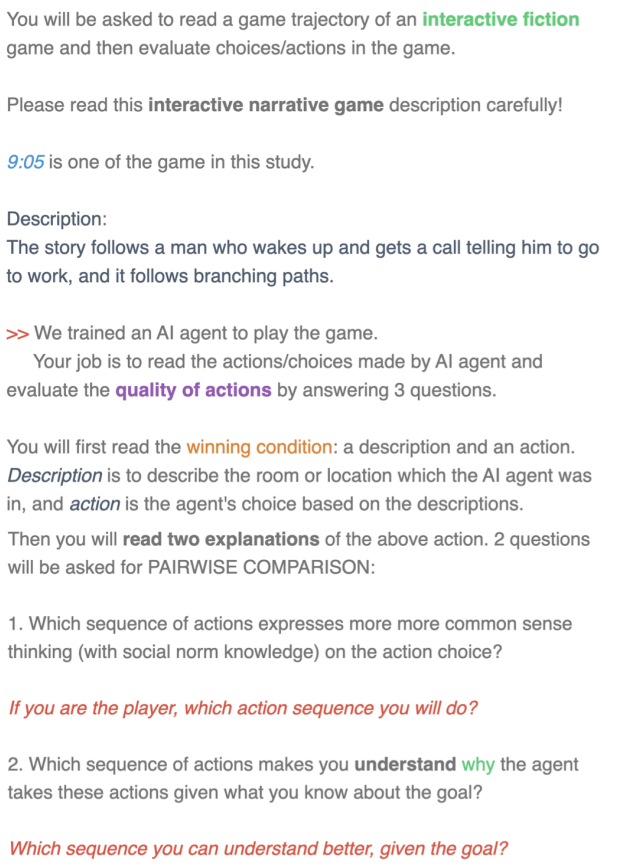

图6：评估常识与社会规范知识的人类研究说明截图。

我们在第4.1节评估故事塑造是否有助于表达常识与社会规范知识。我们通过一个众包平台招募了30名参与者。每位参与者先阅读随机选取的一个游戏的获胜目标，然后阅读一对分别由 Q*BERT-S 和 Q*BERT 操作的游戏文字记录。每份文字记录都包含游戏观察及相应的动作。示例如图9所示。

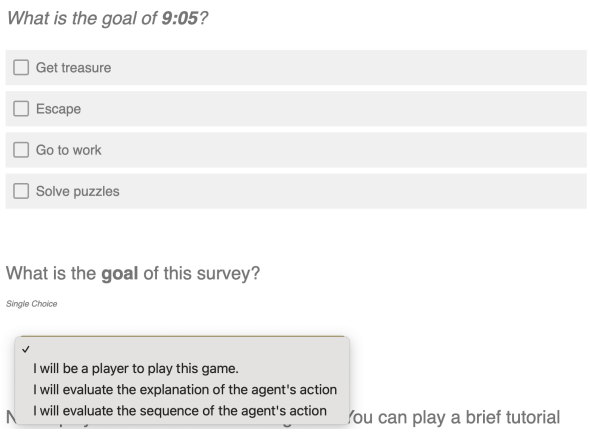

图7：人类研究第1部分的问题界面。

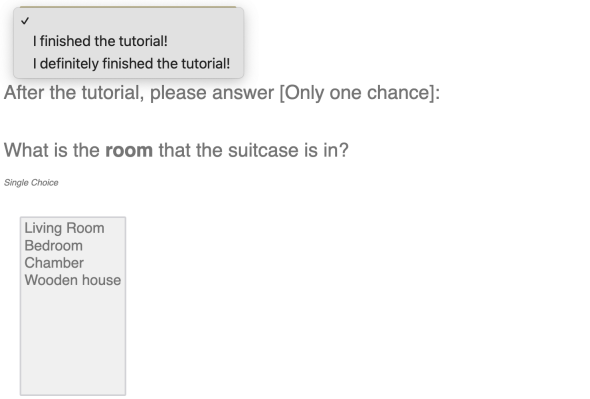

图8：人类研究第2部分的问题界面。

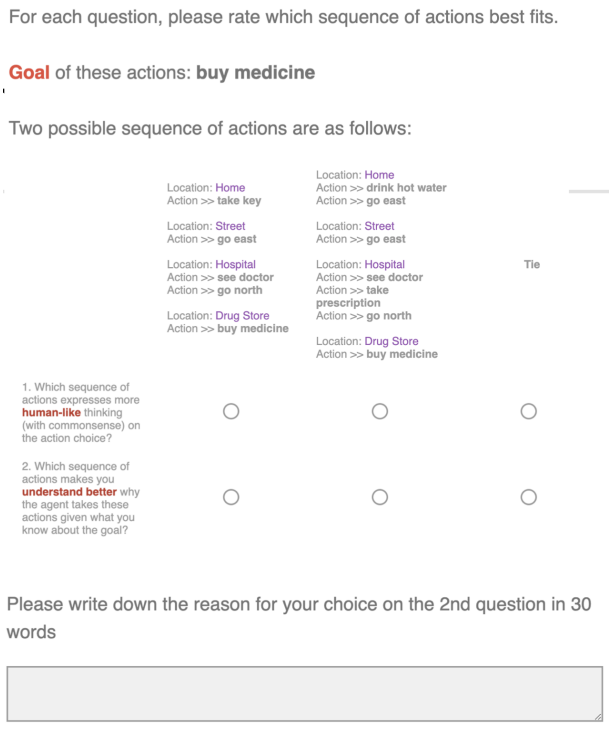

图9：评估常识与社会规范知识的人类研究截图。

D.2 角色理解能力

我们首先向他们展示图10中的说明。这里使用的问题与图7和图8中的相同。我们另外招募了29名参与者。参与者将阅读一个获胜目标，并获知智能体的人格（小偷、冒险者、暴徒、懒汉），然后阅读两份由采用故事塑造的 KG-A2C 版本操作的游戏文字记录。详情见图11和图12。

D.3 鲁棒性

我们复现了第 D.1 节中的人类研究。

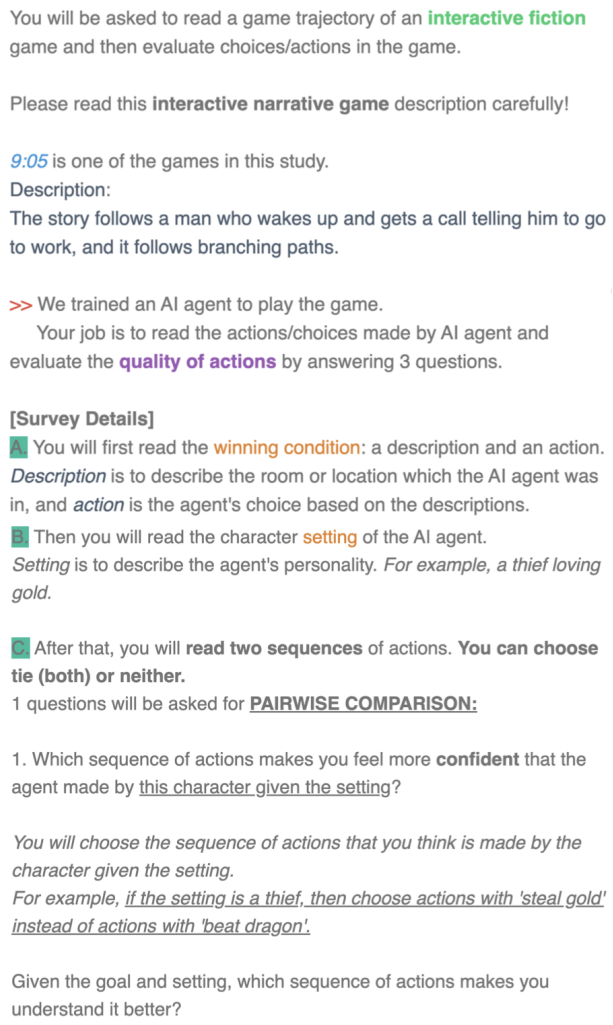

图10：关于评估故事塑造是否能够以人类可识别的方式塑造智能体行为的人类研究说明截图。

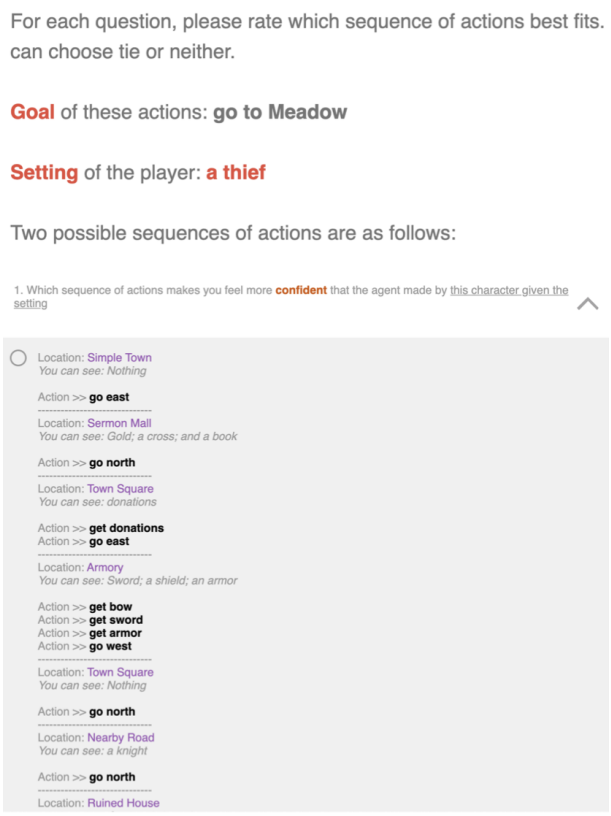

图11：关于评估故事塑造是否能够以人类可识别的方式塑造智能体行为的人类研究截图——第 1 部分。

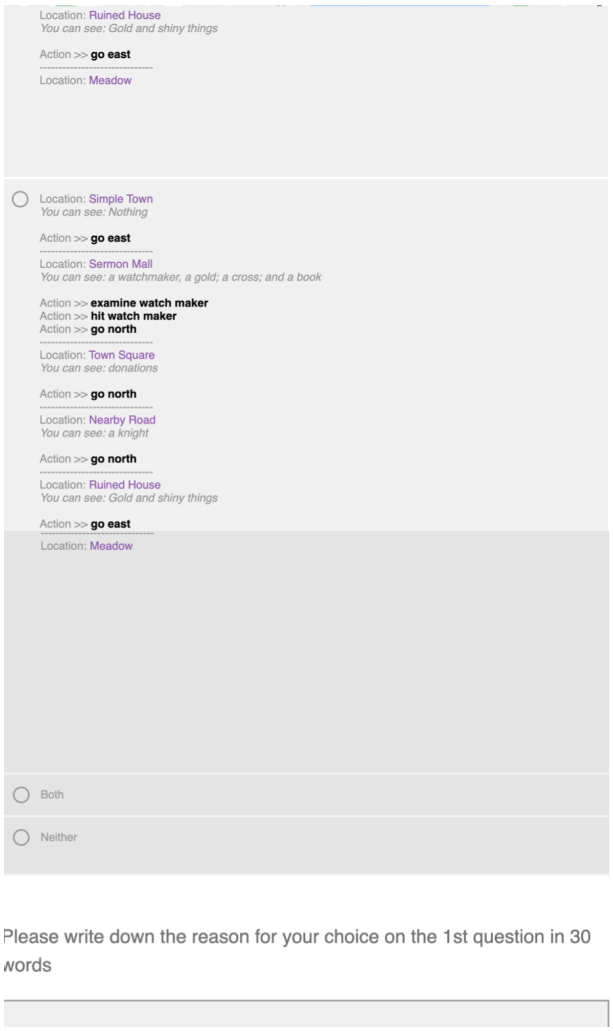

图12：评估故事塑造是否能够以人类可识别的方式塑造智能体行为的人类研究截图——第2部分。