2023 届本科生学士学位论文　　学校代码：10269

華東師範大學

East China Normal University

本科生毕业论文

面向儿童安全教育的交互式讲故事智能体：设计、实现与评估

Interactive Storytelling Agents for Child Safety Education: Design, Implementation and Evaluation

姓　　名：雷颖

学　　号：10195102413

学　　院：计算机科学与技术学院

专　　业：计算机科学与技术

指导教师：孙玉灵

职　　称：副研究员

2023 年 4 月

华东师范大学学位论文诚信承诺

本毕业论文是本人在导师指导下独立完成的，内容真实、可靠。本人在撰写毕业论文过程中不存在请人代写、抄袭或者剽窃他人作品、伪造或者篡改数据以及其他学位论文作假行为。

本人清楚知道学位论文作假行为将会导致行为人受到不授予/撤销学位、开除学籍等处理（处分）决定。本人如果被查证在撰写本毕业论文过程中存在学位论文作假行为，愿意接受学校依法作出的处理（处分）决定。

承诺人签名：雷颖　　日期：2023 年 5 月 23 日

华东师范大学学位论文使用授权说明

本论文的研究成果归华东师范大学所有，本论文的研究内容不得以其它单位的名义发表。本学位论文作者和指导教师完全了解华东师范大学有关保留、使用学位论文的规定，即：学校有权保留并向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅；本人授权华东师范大学可以将论文的全部或部分内容编入有关数据库进行检索、交流，可以采用影印、缩印或其他复制手段保存论文和汇编本学位论文。

保密的毕业论文（设计）在解密后应遵守此规定。

作者签名：雷颖　导师签名：　日期：2023 年 5 月 23 日

目录

面向儿童安全教育的交互式讲故事智能体：设计、实现与评估

摘要：

基于儿童安全教育的重要性和目前面临的挑战，本文旨在使用以人为本的人工智能技术，构建面向儿童安全教育的交互式讲故事智能体，来克服儿童安全教育中存在的挑战，提升儿童安全教育的质量，守护儿童的健康成长。本文的主要工作分为四个部分。首先，通过形成性研究分析了儿童安全教育的现状和用户需求，提出了一种将儿童安全教育融入到交互式讲故事智能体中的解决方案。其次，通过半自动化的方式构建了儿童安全知识语料库，并基于 FairytaleQA 构建了安全知识匹配和安全问答数据集。第三，设计了一个基于 GPT-3 的外部安全知识增强的童话故事安全问答生成的方法，并通过实验验证了模型的有效性。最后，通过用户实验评估了该方法的整体效果，采用了访谈和问卷两种方式收集用户的评价和反馈。结果表明，用户对该方法给予了积极的评价，在各个维度上都显示出了该方法的优势和潜力。同时，用户也提出了一些改进意见，为未来的研究提供了方向。

关键词：儿童安全教育，讲故事，GPT-3，小样本上下文学习，知识检索，问答生成

Interactive Storytelling Agents for Child Safety Education: Design, Implementation and Evaluation

Abstract：

Based on the importance of child safety education and the current challenges, this paper aims to use Human Centered Artificial Intelligence (HCAI) technology to build an interactive storytelling agent for child safety education as a means to overcome the challenges in child safety education, improving the quality of child safety education and guarding the healthy growth of children. The main work of this paper is divided into four parts. Firstly, we analyze the current situation of child safety education and user needs through a formative study, and proposed a solution of integrating child safety education into an interactive storytelling agent. Secondly, we constructed a corpus of children's safety knowledge through a semi-automated approach, and built a safety knowledge matching and safety question and answer (QA) dataset based on FairytaleQA. Thirdly, we designed a GPT-3 based external safety knowledge augmented fairytale safety question and answer generation (QAG) method and verified the effectiveness of the model through experiments. Finally, we evaluate the overall effectiveness of the method through a user study, using both interviews and questionnaires to collect users' evaluations and feedback. The results show that users gave positive evaluations of the method, showing its strengths and potential in all dimensions. At the same time, users also offered some suggestions for improvement, which provide directions for future research.

Keywords： Children's Safety Education, Storytelling, GPT-3, Few-shot In-context Learning, Knowledge Retrieval, Question and Answer Generation

1、 绪论

1.1 研究背景

儿童安全教育是指通过各种方式向儿童传授如何预防和应对各种意外伤害的知识和技能，它对于儿童的健康成长来说十分重要。根据美国疾病控制和预防中心（Centers for Disease Control and Prevention，CDC）的数据，意外伤害是儿童和青少年死亡的首要原因，其中包括交通事故、窒息、溺水、中毒、火灾和跌倒等，然而这些伤害往往是可以通过有效的安全教育来预防的[1]。

因此，提高儿童安全教育的质量和效果是政府、专业组织、和学术界一直以来共同关注的重要课题。为了实现这一目标，一方面，政府或专业组织，如美国卫生与公众服务部（U.S. Department of Health and Human Services）[2]、全球儿童安全组织（Safe Kids Worldwide）[3]、安全之家（SafeHome）[4]等，都在其网站上提供了丰富的儿童安全教育的资源和指南；另一方面，很多研究人员也对儿童安全教育进行了深入的探索和评估，分析了现有的问题和挑战，提出了改进的方法和建议[1-19]。

然而，现有研究表明，儿童安全教育仍然面临着一些困难和局限性。例如，从儿童的角度来看，他们可能缺乏对安全知识的兴趣，或者难以理解和记忆抽象的概念和规则[7]；从安全教育的提供者来看，尤其是父母[14]，他们可能缺乏儿童安全教育的意识，不清楚如何选择合适的内容和方式来引导儿童学习安全知识，或者没有足够的时间和精力来持续地进行安全教育[20]。

1.2 研究内容

基于儿童安全教育的重要性和挑战性，本工作旨在使用人机交互和人工智能技术来克服儿童安全教育中存在的挑战，提升儿童安全教育的质量，守护儿童的健康成长。我们首先通过形成性研究调研了儿童安全教育的现状和用户需求，然后探索出一种将儿童安全教育融入到给儿童讲故事的交互智能体中的解决方案。例如，当智能体讲到“白雪公主吃了女巫的毒苹果中毒身亡”的情节时，相关算法可以生成有关“是否能吃陌生人给的食物”这方面的安全问答。具体来说，本工作的研究内容主要包括以下四点：

研究内容一：用户需求分析。本文通过形成性研究，调研了儿童安全教育的现状和用户需求，提出了儿童安全教育的设计建议，包括日常反复进行、提供现实场景或者想象空间、以及通过智能代理来代替或者协助父母进行安全教育等。

[1]https://www.cdc.gov/injury/features/child-injury/index.html

[2]https://www.hhs.gov/

[3]https://www.safekids.org/safetytips

[4]https://www.safehome.org/resources/child-safety/

研究内容二：语料库和数据集的构建。本文构建了一个儿童的安全知识语料库，包含了 141 条由主题、危险后果、场景、危险行为和安全行为组成的安全知识，并标注了一个基于 FairytaleQA[21] 数据集的安全知识匹配和安全问答数据集，包含了 35 个故事段落和对应的 157 条安全知识和问答对。

研究内容三：算法设计和模块测评。本文提出了一个外部安全知识增强的童话故事安全问答生成方法，它由两个模块组成：知识检索模块和问答生成模块。知识检索模块利用神经协同过滤和对抗损失实现，能够根据故事文本匹配合适的安全知识；问答生成模块利用基于 GPT-3 的小样本上下文学习实现，能够根据故事文本和安全知识生成有教育意义的问答。本文通过实验验证了这两个模块的效果，发现它们都远超过基线模型。

研究内容四：用户实验。本文通过用户实验评估了该方法的整体效果，采用了访谈和问卷两种方式收集用户的反馈和评价，并通过定性和定量分析相结合的方法进行分析。结果表明，用户对该方法给予了积极的评价，在各个维度上都显示出了该方法的优势和潜力。同时，用户也提出了一些改进意见和建议，为未来的研究提供了方向。

1.3 研究意义

本工作旨在使用人机交互和人工智能技术克服儿童安全教育中存在的挑战，提升儿童安全教育的质量，守护儿童的健康成长。本工作具有重要的理论意义、技术意义和实践意义。

从理论意义上来说，本工作以人工智能、人机交互理论为指引，关注儿童安全教育存在的困难和挑战，探索了一种将儿童安全教育融入到给儿童讲故事的交互智能体中的解决方案，为儿童安全教育提供了一种新颖有效的方式，丰富了人机交互和人工智能领域的相关理论和方法。

从技术意义上来说，本工作构建了儿童安全知识语料库和基于童话故事的知识匹配和安全问答数据集，设计并实现了一个外部安全知识增强的童话故事安全问答生成方法，提供了将安全知识融入给儿童讲故事的交互智能体的数据和技术基础。

从实践意义上来说，本工作提供了一种将儿童安全教育融入到给儿童讲故事的交互智能体中的解决方案，并通过用户实验验证了其效果，为儿童安全教育提供了一种有趣易用的潜在工具，可以帮助儿童提高安全知识水平，增强安全意识和行为能力，减少意外伤害风险。

1.4 章节安排

本文共分为七章，具体结构安排如下：第一章为绪论，介绍了本工作的研究背景、内容、意义，以及章节安排等；第二章为相关工作，依次介绍交互式讲故事智能

体、知识增强的问题生成或者问答、以及基于大语言模型的小样本上下文学习这三个方面的研究现状、代表性工作、以及研究空白；第三章为用户需求分析，介绍了形成性研究的方法过程，有关儿童安全教育现状和用户需求的关键结论，最后得出了将安全教育融入给儿童讲故事的过程中的解决方案；第四章为语料库和数据集的构建，介绍了具体的构建过程、规则、标注、和示例等等；第五章为算法设计与模块测评，介绍了知识检索模块和问答生成模块的原理、实现和评估；第六章为用户实验，介绍了用户实验的设计、执行和分析过程，并给出了用户实验的结果和反馈；第七章为结论与展望，总结了本文的主要工作和贡献，指出了本文的不足和局限，并提出了未来的研究方向和建议。

2、 相关工作

本文依次从交互式讲故事智能体、知识增强的问题生成或者问答、以及基于大语言模型的小样本上下文学习这三个方面的研究现状、代表性工作、以及研究空白（Research Gap）等角度介绍本文的相关工作。

2.1 交互式讲故事智能体

交互式讲故事智能体是指能够根据给定的内容，与听众进行交互式的问答或反馈的人工智能系统。在儿童教育场景下，交互式讲故事智能体也成为一种有效的辅助工具，帮助培养儿童的想象力、创造力、语言能力、社交能力等[22]。近年来，随着人工智能技术在自然语言处理、计算机视觉、人机交互等领域的快速发展和应用，构建面向儿童的交互式讲故事智能体已经成为一个热门且具有挑战性的研究方向。目前已经有一些代表性的研究工作涉及到这一领域，例如童话故事问答数据集 FairytaleQA[21]、基于神经网络的问答或者问题生成算法[23-24]、基于对话系统的故事应用[20] 等。然而，这些研究工作还存在一些不足和局限性。例如，从教育学的理论和视角出发，给儿童讲故事不仅是为了娱乐和启发，也是为了传递一些有价值的知识和道德[25-27]。因此，交互式讲故事智能体不应该只是简单地生成和回答与故事文本相关的问题，而应该能够结合外部常识和实际生活情境，生成和回答一些具有教育意义的问题，帮助儿童拓展知识面，增强生活技能，塑造价值观。

研究空白：交互式讲故事智能体是进行儿童安全教育的潜在的方案，然而目前还没有研究将外部常识和实际生活情境融入到交互式讲故事智能体中，包括本文关注的安全常识知识。因此，本文将探索将儿童安全教育融入到给儿童讲故事的交互式智能体中来填补这一研究空白。

2.2 外部知识增强的问题生成或者问答

外部知识增强的问答或者问题生成技术是指利用外部知识来提升问答或问题生成任务的性能和效果的技术。近年来，随着大规模知识库和数据集的出现和发展，研究学者们开始探索如何将外部知识融入到问答或问题生成的模型中，以解决一些需要常识或专业知识的问题。常见的外部知识可以分为结构化知识和非结构化知识，其中常见的结构化知识包括 ConceptNet[28-29] 、ATOMIC2020[30] 等，常见的非结构化知识包括 Wikipedia [1] 、Dbpedia[31]、Wiktionary [2] 等，或者自动生成相关知识图谱的 COMET[32]。常见的基于知识的问答或者问题生成的数据集包括 CommonsenseQA[33] 、OpenbookQA[34]、OK-VQA[35] ，K-VQA[36] 等。在这些知识库和数据集的基础上，研

[1]https://www.wikipedia.org/

[2]https://en.wiktionary.org

究学者们提出了一系列相关方法。例如，在问答方面，Xu 等学者[37] 结合了 ConceptNet[28-29]和 Wiktionary [1] 来优化常识问答，Yasunaga 等学者[38] 把问答数据和检索出的外部常识性子图进行联合推理，Feng 等学者[39] 重点研究了基于知识的常识性问答中的多跳问题，等等。在问题生成方面，Jia 等学者[40] 使用 ConceptNet[28-29] 优化了问题生成的效果，Uehara 等学者[36] 研究了基于知识的视觉问答生成任务，等等。

研究空白：知识增强的问答生成是将儿童安全教育融入到给儿童讲故事的交互是智能体中的技术方案，然而目前的研究集中于知识增强的问题生成或者问答，尚未有知识增强的问答生成的相关工作，尤其在交互式讲故事智能体的算法研究领域。另外，目前还没有专门用于儿童安全教育的儿童安全知识语料库，以及相关的数据集以供参考和使用。因此，本文通过构建儿童安全知识语料库和基于童话故事的安全知识匹配和安全问答数据集，并设计相应的知识增强的安全问答生成的算法，来填补这一研究空白。

2.3 基于大语言模型的小样本上下文学习

基于大语言模型的小样本上下文学习（Few-shot In-context Learning，FSIL）[41] 是一种利用预训练的大规模语言模型（Large Language Model，LLM），如 GPT-3[42]，GPT-4[43]，Alpaca[2]等, 在新任务上进行快速适应和泛化的方法，而无需对模型进行微调或参数更新。该方法的核心思想是通过给模型提供一些任务相关的示例（如输入输出对），让模型从中学习任务的模式，并据此对新输入进行预测。具体来说，在推理过程中，新任务 y 的目标直接以给定的上下文 C 和新任务的输入 x 作为文本序列生成任务。注意这里的 C，x，y 都是文本序列，例如，$y = (y^1, \cdots, y^T)$。因此，在每个解码步骤 t 处，有：

$$y^t = argmax_{yt} p_{LM}(y^t \mid C, x, y^{<t})$$

其中 LM 表示预先训练的语言模型的权重，对于所有新任务都被固定住（frozen）。上下文 $C = h, x_1, y_1, ..., x_n, y_n$，包括一个可选择的头部提示信息（Prompt Head）h，以及 n 个新任务的例子（$\{x_i, y_i\}_{i=1}^n$）作为上下文信息（In-context Examples）。

基于大语言模型的小样本上下文学习具有数据效率高、计算成本低、泛化能力强等优点，因此在近年来受到了越来越多的关注和探索，涉及到多种自然语言处理任务，如文本分类[44] 、摘要生成[42]、问答[42]、翻译[42]、代码生成[42,45] 等。但这种方法也存在一些挑战和局限性，如如何选择合适的示例[42,46]、如何生成有效的提示[47]、如何评估和解释模型的行为[48]等。

[1] https://en.wiktionary.org

[2] https://crfm.stanford.edu/2023/03/13/alpaca.html

研究空白：由于大语言模型的强大小样本上下文学习的能力和其包含的多样化的知识，基于大模型的小样本上下文学习是将儿童安全教育融入到给儿童讲故事的交互是智能体中生成安全问答任务的潜在解决方案，但目前使用此方法生成问答的工作还十分有限，尤其在交互式讲故事智能体的算法研究领域。因此，本文通过基于GPT-3 的小样本上下文学习的方式实现基于故事段落和安全知识的安全问答生成，并验证其有效性，来填补这一研究空白。

3、 用户需求分析

为了进一步从用户的角度理解儿童安全教育的现状和需求，我们进行了一项形成性研究（Formative Study）[20]，收集了 1）对儿童安全教育的重视程度，2）其孩子参与过的安全教育类型、场景、内容以及用到的工具，以及 3）用户对于现有的儿童安全教育的看法等等。和[3,7]类似，我们选择 5-9 岁的儿童为例进行集中研究，一方面正如 SafeKids Worldwide [1] 所示，不同年龄阶段的安全教育具有不小的差异，另一方面该年龄段的儿童开始拥有除家庭以外的更大活动区域，经常面临缺乏监管的情况，是儿童安全事故高发的年龄段。

3.1 研究方法

3.1.1 参与者

本实验中，我们招募了拥有 5-9 岁年龄段孩子的父母作为用户实验的参与者。我们没有对他们的安全意识或对儿童安全教育的参与或重视程度设定特定的要求，以便招募到具有不同背景和经验的人群。我们主要通过两种方法招募参与者：1) 个人社交网络，和 2) 滚雪球抽样（Snow Sampling）[49]。我们首先从个人社交网络（例如，朋友、同学和家人）中邀请符合条件的人参与实验，然后随机要求他们在他们的社区或社交网络中推荐其他可能感兴趣的人。为了确保参与者的多样性，我们尽量招募具有不同性别、年龄、地点等人口统计学特征的参与者。最终，我们共招募了 7 名参与者，其中 3 名女性和 4 名男性，年龄在 30 岁到 50 岁之间，分别来自辽宁、湖南、安徽、黑龙江、上海等城市。他们的孩子年龄在 5 岁到 9 岁之间，有不同的性别和年龄。表 3-1 列出了参与者的基本信息。

表 3-1　研究参与者的人口统计学信息

Table 3-1　Demographic Information of Participants

参与者	性别	年龄	地点	教育水平	职业	孩子性别	孩子年龄
P1	女	30-40	辽宁	本科	公司职员	男	5
P2	男	40-50	湖南	本科	销售	男	9
P3	男	40-50	安徽	硕士	公司职员	女	8
P4	女	30-40	黑龙江	本科	教师	男	6
P5	男	40-50	上海	硕士	IT 开发人员	女	7
P6	女	30-40	重庆	本科	个体经营户	女	9
P7	男	30-40	山东	高中	司机	男	5

[1] https://www.safekids.org/

3.1.2 数据收集

我们采用半结构化访谈（Semi-structure Interview）的形式收集用户的反馈和评价。我们预先设计了一些一般性的问题，以适应具有不同经验水平的受访者。当受访者提到一些有趣的观点或先前的经验时，我们会跟进更多的细节和具体的例子。在2023 年 2 月期间，所有访谈均由作者亲自或远程通过腾讯会议、微信语音通话或电话进行。每次访谈持续 20 分钟左右，每位参与者获得 20 元酬金。在参与者的同意下，我们录音并转录了所有的访谈内容。

3.1.3 数据分析

对于半结构化访谈的内容，我们采用主题分析方法（Thematic Analysis）[50] 进行分析。主题分析是一种定性研究方法，旨在从数据中识别、分析和报告重要的主题或模式。我们在收集数据的同时开始进行编码（coding）和分析，并使用 Excel 来编码和跟踪主题。在最初的分析阶段，我们仔细审查了收集到的数据，并标记了与研究问题相关的想法。然后，我们结合了最初的代码列表，反复迭代几次，并生成了最终的主题图。在关键结论部分，我们将介绍这些主题的全部细节，并使用有代表性的引用来支持我们的发现。为了保护参与者的身份，我们使用 PX 来表示参数。

3.2 关键结论

3.2.1 家长对儿童安全教育的重视程度

我们使用口头版的 5 点李克特量表（5-point Likert scale）[51-52]在访谈中询问参与者对儿童安全教育的重视程度，并用饼状图展示了结果，如图 3-1 所示。图中绿色越深表示越重视，蓝色越深表示越不重视，灰色表示一般。结果显示，有 29% 的参与者非常重视儿童安全教育，14% 的参与者一般重视，而剩下的 57% 的参与者持中立或不重视的态度。这说明参与者对儿童安全教育的意识和知识还有很大的提升空间。

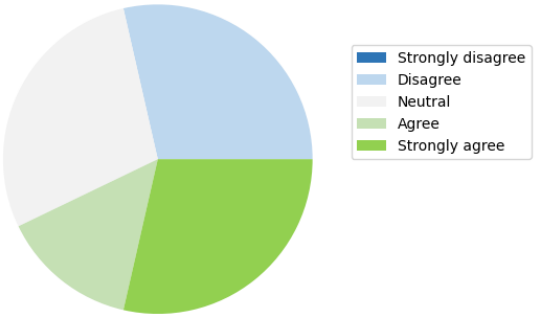

图 3-1 参与者对儿童安全教育的重视程度

Figure 3-1 The Importance that Participants Place on Child Safety Education

通过访谈分析，我们发现参与者对儿童安全教育缺乏足够重视的原因主要有三方面：一是自己本身对安全教育就不够重视，只是偶尔在生活中提醒孩子注意安全；二是自己也缺乏相关的系统教育或者知识储备，只能用自己的生活经验给孩子传授安全知识；三是自己工作很忙，没有足够的时间陪伴孩子进行亲子互动和安全教育。例如，P6 表示：“因为我自己也不怎么重视，比如说我看到他手上湿了，然后就碰电器，我想到有这么个事就有可能会说一下，如果我想不起来这个事，我可能也不会”。P2 表示：“我就只能用我的生活经验，生活中碰到一个 case 就说一个 case”。P6 表示：“自己工作很忙，没有足够的时间陪伴孩子”。因此，从对参与者对儿童安全教育重视程度的分析可以看出，家长给儿童的安全教育存在意识、知识和时间上的局限性。

3.2.2 儿童安全教育的类型、场景、内容和工具

通过访谈，我们总结了参与者的孩子参与过的儿童安全教育的类型、场景、内容和工具，以及他们对此的看法，如表 3-2 所示。

表 3-2 儿童参与的安全教育的类型、场景、内容和工具

Table 3-2 Types, Scenarios, Content and Tools of Safety Education in Which Children Are Involved

类型	内容
类型	家长/老师提醒 [P1]、防灾演练（消防，地震逃生）[P1，3，4]、专家讲座 [P1，3，4，5，6]、教育视频 [P1，2，3，5]、公众号推文 [P1]、教材手册百科全书 [P2，4]、动画片 [P2，P5]、参与亲子活动 [P3，6]、互联网资料 [P4]、视频访谈类节目 [P4]、政府党委官方宣传 [P4]、事故视频 [P4]、学校的课程 [P4] 等
场景	家庭生活 [P1，2]、公众场所 [P1，P4，P6]、社区街道 [P3] 等
内容	安全出口 [P1]、防火防溺水 [P1，4，6]、防烫伤 [P2]、运动带护具 [P2]、地震 [P3]、陌生人教育（个人信息，食物诱拐，跟走）[P3，4，6]、自然灾害（泥石流）[P6]、用电安全 [P6]、踩踏 [P6] 等
工具	书籍 [P2，4]，电视互联网资料 [P1，2，3，4，5]，Ipad [P5] 等

从表中可以看出儿童参与的安全教育类型、场景、内容和工具是多样化的。但即便针对儿童的安全教育如此丰富多样，用户表示这些教育类型大多数具有一定的局限性：一方面，安全知识概念对来小孩来说是抽象的、缺乏场景支撑的，不能给小孩带来很强的触动；另一方面，一些安全教育受各方面限制较大，需要能够在日常中反复强调。

首先，在安全知识概念的理解方面，教育效果受限于缺乏具体的场景支撑、儿童

对抽象概念的理解能力、以及本身的知识储备，例如 P3 表示“网上看的这种给你的那种触动，并不如真实环境当中那么强”，P4 表示“孩子对抽象概念的理解能力很差，你很难用长的语言描述去让它明白一个事情”，P6 提供了一个具体的例子“比如说洗个手不能去触碰电器，我觉得对于 5 9 岁的小孩来说，这个概念比较抽象。”相反，几乎所有的参与者都承认现身说法的重要性，即让儿童真实地或者想象处在具体的场景中 [P3，4，5，6]，例如 P4 表示“孩子不具备很多社会性的概念，需要给他提供一个具体的、容易理解的想象空间”，P5 则强调具体实践、P6 强调生活经历、P3 强调防灾演习。

另外，参与者也普遍赞同在日常中反复强调的重要性，例如 P3 觉得防灾演习虽然能够提供具体的场景，促进儿童的理解，但“实地演练这个东西这个机会并不是很多，你得抽时间，家里、工作上都很忙，而且你要看小孩子有没有上学、有没有空…受各方面限制比较大”，P6 则表示“你给他讲一次两次，不一定能让他记住，这种东西它最好还是说在教育的过程中，它能反复的去强调了解这种概念”。

因此，从对儿童安全教育的类型、场景、内容和工具的分析可以看出当前不少类型的安全教育对于儿童的效果有限，而提供具体简单的现实场景或者想象空间以促进儿童理解，以及在日常中反复强调来深化印象是儿童安全教育中比较重要的因素。

3.2.3 设计建议

由以上的分析，我们得出几点关于儿童安全教育在技术方面的设计建议：（1）对儿童的安全教育需要在日常反复进行，因此可以尝试把安全教育融入到现实生活场景中，例如讲故事、看动画片；（2）儿童对抽象概念缺乏理解，需要以儿童容易理解的方式为他提供提供具体简单的现实场景或者想象空间，例如存在于童话故事和动画片中的场景；（3）家长给儿童提供的安全教育具有意识、知识和时间方面的局限性，可以设计一款儿童安全教育的自动代理来替代或者辅助家长对儿童进行安全教育。

3.3 设计方案

基于关键结论中的设计建议，我们考虑在儿童讲故事的过程中融合安全教育，设计一款基于童话故事的儿童安全教育智能体，理由如下：

第一，如相关工作所述，讲故事是儿童阶段在日常反复进行的常见教育方式和亲子互动行为，目前已被用来增长儿童的语言能力、逻辑思维等方面，因此将常识教育，包括安全教育，融入讲故事的过程中是很自然的行为。尽管现实场景或者操作演习能带来更加真实的体验，但讲故事作为一种更日常的行为可以作为一种有效补充，能让他们在听故事的的同时反复吸收安全知识，强化安全意识。因此，讲故事是一种适合儿童的日常安全教育的形式。

第二，相比传统的专家讲座、视频、以及教材手册，童话故事场景以一种更为儿

童接受的方式为儿童提供了一个具体的、简单的想象空间，使得他们能够结合具体的故事情节，自然地学习现实场景下的安全知识。童话故事中往往有着寓意深刻、形象生动、情节曲折、人物鲜明的特点，能够激发儿童的兴趣和好奇心，也能够帮助儿童建立正确的价值观和道德观。通过在讲故事的过程中融入安全教育，我们可以让儿童在享受故事乐趣的同时，也能够学习到有用的安全知识。所以，童话故事是一种适合儿童安全教育的内容载体。

第三，目前已有不少工作设计了给儿童讲故事的自动代理来替代或者辅助父母给儿童讲故事，因此我们也可以通过自动问答或者给家长提供安全知识提示的方式自动化儿童安全教育的过程，来弥补家长给儿童进行安全教育在意识、知识和时间方面的局限性。自动代理可以在讲故事的适当的时机插入安全教育相关的问题或者建议，引导儿童思考和回答。同时，自动代理也可以给家长提供一些安全教育相关的提示或建议，帮助家长更好地与儿童沟通和互动。通过这样的方式，我们可以让儿童安全教育变得更加智能化和便捷化。因此，我们认为自动代理是一种适合儿童安全教育的工具。

总之，通过以上分析，我们考虑设计一款基于童话故事的儿童安全教育智能体，它可以在儿童讲故事的过程中融合安全教育，利用童话故事的形式和内容来引导儿童学习安全知识，同时利用自动代理的技术来实现智能化和便捷化的安全教育。具体方案为：首先，我们将准备儿童安全知识语料库，以及基于童话故事的安全知识匹配和问答数据集。因为目前还没有专门用于儿童安全教育的安全知识语料库和基于童话故事的安全故事数据集。然后，我们将构建安全知识检索和安全问答生成两个模块的模型，实现输入一段故事文本，从安全知识语料库中匹配合适的安全知识，再基于故事和知识生成安全知识问答的目的，并使用自动测评的方法分别对这两个模块进行测试。最后，我们使用用户测评对整个方法的有效性进行了验证，并得出未来设计的建议。

4、 儿童安全教育语料库与数据集构建

由于目前还没有专门用于儿童安全教育的安全知识语料库和基于童话故事的安全知识匹配和安全问答数据集，我们首先需要做一些数据准备工作，包括（1）构建面向儿童安全教育的安全知识语料库，（2）构建基于童话故事的安全知识匹配和安全问答数据集。具体来说，本工作采用半自动化的方式构建安全知识语料库，并以段落为单位对童话故事标注对应的安全知识和问答。

4.1 安全知识语料库的半自动化构建

4.1.1 安全知识来源检索和筛选

为了收集儿童安全知识，我们在互联网上进行安全知识来源的检索和筛选，大致步骤如图 4-1 所示。首先，我们以“child safety education”，“child safety tips”，“child safety rules”等关键词在互联网上搜索，优先选择政府或权威机构的网站。然后，我们浏览搜索结果，直到没有新的信息出现，并按照以下标准筛选资料：（1）与主题相关性；（2）来源可信度；（3）目标受众，如儿童和父母；（4）可访问性，如是否有地区限制等。

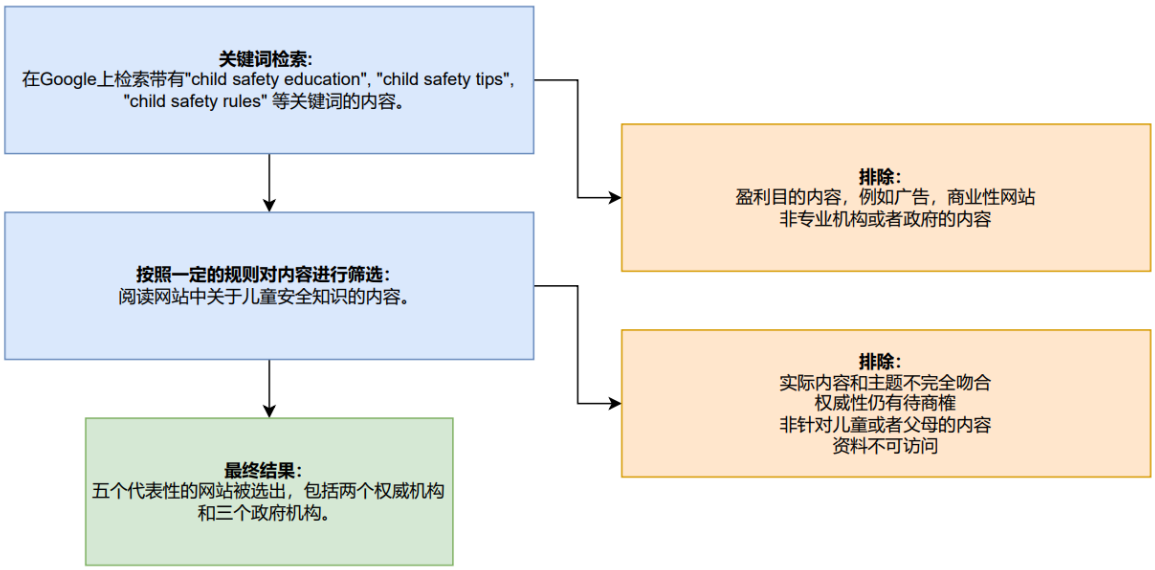

图 4-1 安全知识来源的检索和筛选的过程

Figure 4-1 The Process of Retrieving and Filtering Safety Knowledge Sources

经过筛选，我们最终选取了五个具有代表性的网站作为儿童安全知识来源，分别为 Safekids Worldwide[1]，SafeHome[2]，美国疾病控制和预防中心 (Centers for Disease

[1]https://www.safekids.org/big-kids

[2]https://www.safehome.org/resources/child-safety/

Control and Prevention, CDC) [1]，爱尔兰官方政府网站 [2]，美国国家失踪与被剥削儿童中心 (National Center for Missing and Exploited Children, NCMEC)[3]。

4.1.2 安全知识内容的提取和整合

为了避免手工进行简单重复的操作，我们使用爬虫对各个网站中的知识进行抽取（PDF 版本的资料除外），大致包含三个步骤：（1）基于 Request 包获取网页内容；（2）基于 BeautifulSoup 包对网页内容进行解析，提取需要的内容；（3）以 Excel 的格式保存提取到的结果。获取各来源的知识后，我们首先以来源于 Safekids 的安全知识为主进行整理，然后将来自于其他来源的、不被 Safekids 的安全知识所包含的内容有选择性地依次整合到其中，最终得到了 119 条面向儿童的安全知识。

4.1.3 安全知识特征提取和重构

当前对于每一条安全知识，基本的组成元素都是主题和描述，其中描述又由危险后果，场景，危险行为，安全行为这四个元素中的部分或者全部组成。例如，现有一条安全知识，主题为“Water Safety”，描述为“Whether you’re swimming in a backyard pool or in a lake, teach children to swim with a partner, every time. Do not allow children to swim alone”，其中的危险后果为“drowning”，场景为“swim; water, like pool, lake, river, or ocean”，危险行为为“swim alone”，安全行为为“swim with a partner every time”。由于场景和危险后果是将故事情节和安全知识进行匹配的重要元素，而危险行为和安全行为对于安全知识问答的生成也必不可少。因此，为了方便后续的知识检索和问答生成，针对每一条安全知识，我们从其描述部分提取出了危险后果，场景，危险行为，安全行为等特征。注意，针对安全知识缺少其中个别元素的情况，使用 unknown 进行填充。因此，经过特征提取的过程，每一条安全知识由“主题 | 描述”的模式转化为了“主题 | 危险后果 | 场景 | 危险行为 | 安全行为”的模式。然而，由于对于每一条安全知识，它可能属于多个主题、包含多个危险后果、场景，同时，不同安全行为或者危险行为可能又对应着同样的主题、危险后果和场景。因此，我们使用 Python 对提取完特征后的安全知识进行了处理，使得每条知识的主题、危险后果和场景对各异，对应一到多条危险和安全行为。最终，我们得到了 141 条面向儿童安全教育的安全知识，举例如表 4-1 所示。

4.2 基于 FairytaleQA 的数据集构建

基于构建的面向儿童安全教育的安全知识语料库，我们构建基于童话故事的安全知识匹配和安全问答数据集，将故事情节和安全知识、安全问答建立一到多的映射

[1] https://www.cdc.gov/parents/children/safety.html

[2] https://www.nidirect.gov.uk/articles/keeping-children-safe-while-out-and-about

[3] https://www.missingkids.org/

表 4-1　最终版安全知识的示例

Table 4-1　Examples of the Final Version of Safety Knowledge

类型	内容
主题	Water Safety
危险后果	Drowning
场景	Water, like pool, lake, river, or ocean
危险行为 1	Not wear personal flotation device (PFD)
安全行为 1	Always wear a U.S. Coast Guard-approved personal flotation device
危险行为 2	Without supervision of an adult
安全行为 2	Never go near or in water without an adult present
危险行为 3	Swim alone
安全行为 4	Swim with a partner every time

关系。考虑到时间和工作量的因素，我们意在构建一个小样本数据集。根据统计学中的大数定律[53]，我们确定 24 为最小的样本数值。具体构建步骤如下：

首先，我们确定 FairytaleQA[21] 为童话故事文本来源。FairytaleQA[21] 是一个专注于幼儿园到八年级学生叙事理解的数据集，包含来源于 15 本故事书的 278 个儿童友好故事，从儿童年龄和故事内容、多样性方面均适合于本任务，因此选择其为童话故事的文本来源。

然后，我们从 FairytaleQA[21] 中随机选取了来自于 14 个童话故事的 35 个段落。考虑到句子所包含的信息有限，而整个故事又过长，因此我们选择以段落（section）为单位进行标注。另外，在着 35 个故事段落中，其中 30 个用于模型的训练和测试，剩下 5 个用于后续的用户实验。

最后，我们对每一个故事段落标注一到多个安全知识标签，并针对每一个故事段落和安全知识标签对标注上对应的问答，得到了 157 个样本，每一个均包含故事段落、安全知识、安全问题、安全答案四部分的内容。

5、 面向安全教育的知识检索与问答生成算法

本文在第 4 部分描述了构建面向儿童安全教育的安全知识语料库和基于童话故事的安全知识匹配和安全问答数据集的总体流程。基于此，为了达到输入一段故事文本生成对应安全问答的目的，首先需要根据特定的故事段落检索到合适的安全知识，然后再基于故事文本和安全知识生成对应的安全问答。因此，本部分将重点描述安全知识检索和安全问答生成的算法设计与模块测评，总体框架如图 5-1 所示。

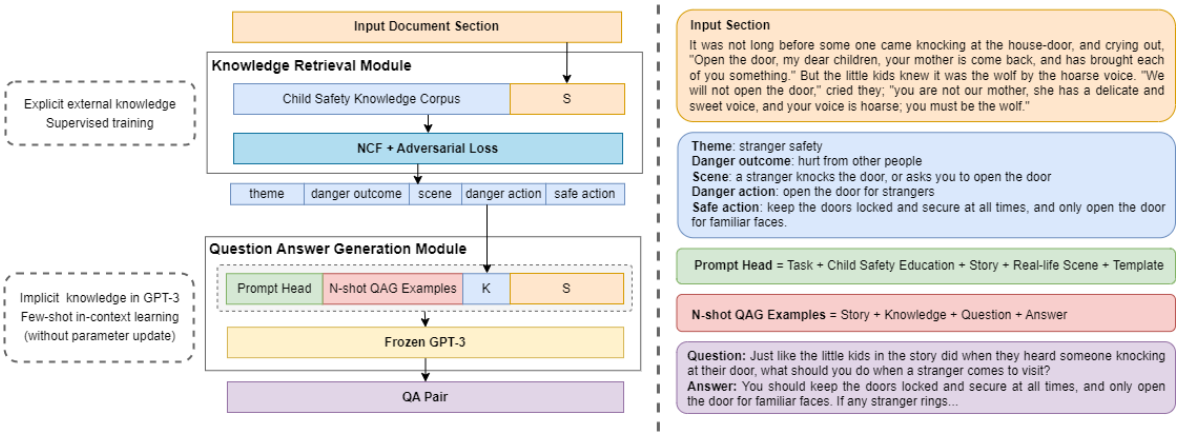

图 5-1　知识检索和问答生成的整体模型框架

Figure 5-1　General Model Framework for Knowledge Retrieval and Question Answer Generation

5.1　基于神经协同过滤和对抗损失的安全知识检索

5.1.1　方法描述

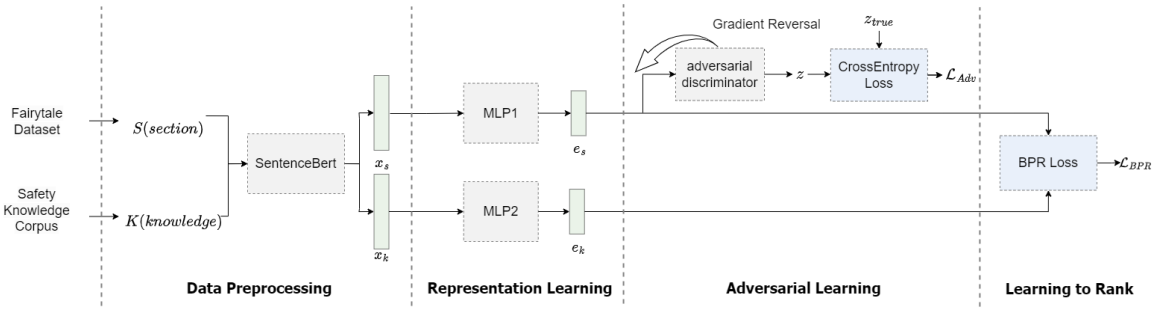

图 5-2　基于神经协同过滤和对抗损失的安全知识检索的模型框架

Figure 5-2　Model Framework of Safety Knowledge Retrieval based on Neural Collaborative Filtering and Adversarial Loss

安全知识检索的算法总体框架如图 5-2 所示。该框架主要分为四部分，依次是数据预处理、表征学习、对抗学习和排序学习。数据预处理模块使用预训练模型 SentenceBert[54] 对故事文本和安全知识进行处理得到向量表示。表征学习的模型主体是双塔多层感知机（Multi-layer Perceptron, MLP）[55]，用于从数据中学习有用的表示。对

抗学习的模型主体是一个鉴别器，结合梯度逆转用以阻止模型学习可能只有对特定领域才重要的表征，避免对训练域的分布“过拟合”。排序学习使用的是通过贝叶斯个性化排序损失（Bayesian Personalized Ranking Loss，BPR Loss）[56] 来对故事段落和正负知识的匹配程度排序进行学习。

5.1.1.1 数据预处理

SentenceBert[54] 作为经典的句子嵌入预训练模型，结合余弦相似度被广泛用于语义文本相似度计算、语义搜索或释义挖掘等任务中。因此，在数据预处理模块，我们使用[54] 将训练和测试所需的 30 个故事文本（表示为 section），及 141 条安全知识文本（表示为 knowledge）处理为固定维度的向量 x_s 和 x_k。公式如下：

$$x_s = SentenceBert(section)$$

$$x_k = SentenceBert(knowledge)$$

5.1.1.2 表征学习

表征学习的目的通过训练排序（Learning to Rank，LTR）模型，从数据中学习有用的表示。相比传统排序模型[57]，深度神经排序模型对学习对目标任务更高级的表示很有用。因此，模型的主体借鉴了信息检索经典模型神经协同过滤（Neural Collaborative Filtering，NCF）[58]，通过双塔多层感知机（Multi-layer Perceptron，MLP）[55] 分别对输入的故事段落向量表示 x_s 和安全知识向量表示 x_k 进行学习，得到输出 e_s 和 e_k。公式如下：

$$e_s = MLP_1(x_s)$$

$$e_k = MLP_2(x_k)$$

其中 MLP_1，MLP_2 是双塔 MLP 模型。

5.1.1.3 对抗学习

对抗学习的目的是防止模型对训练域的分布进行“过拟合”，尤其针对本工作的小样本数据集。深度神经排序模型能从训练数据中学习特征的能力是一个强大的属性，它使他们能够潜在地发现没有被手工制作的特征所捕获的新关系，然而学习新特征的能力可能会以在训练过程中没有观察到的领域的泛化和性能差为代价[59]。例如，模型可能会观察到某些关联——例如，“狼/小山羊”（来自《狼和七只小山羊》）和“陌生人敲门”这条安全知识——在训练语料库中比其他关联更容易同时出现，或者该模型可能会得出结论，根据它在训练查询中出现的相对频率，学习“狼/小山羊”比学习其他更重要，例如，同样带有“陌生人敲门”情节的“白雪公主/王后”（来自《白雪公主》）更重要。尤其针对本工作的小样本数据集，模型容易对。如果我们的目

标是在训练的故事上实现最佳性能，那么这些相关性和分布很重要，但如果我们想要在新的故事中实现“开箱即用”性能，模型必须学习对它们更健壮。相比之下，传统检索模型做出更少的分布假设，通常表现出更稳健的跨域性能。注意，该方法中的域指的是不同的故事段落。

近年来，对抗性学习（Adversarial Learning）被证明是一种有效的适用于分类任务的跨域正则化器（Cross Domain Regularization）[60-62] 。我们采用了类似的策略来迫使神经排序模型学习更多的故事域不变表示，即将排序模型与一个对抗性鉴别器（discriminator）相结合。我们在训练中采用留一交叉验证（leave-one-out cross validation，LOOCV）法[63]，我们表示训练集中的域为 $D_{train} = d_1, ..., d_{k-1}, d_{k+1}, ..., d_n$，测试集的域为 $D_{test} = d_k$，其中 n 为用于训练和测试的样本总数。该鉴别器试图基于排序模型学习到的表示来预测训练样本的域 d_{true}，即故事段落的分类。当通过排序模型的各层反向传播时，来自对抗性组件的梯度被逆转，从而为排序模型提供了一个负反馈信号，以阻止它学习可能只有对特定领域才重要的表征。具体来说，该鉴别器是一个分类器，它检查排序模型的隐藏层的输出，并试图预测训练样本的域，并使用一个标准的交叉熵损失（Cross Entropy Loss）[64] 进行训练。对抗损失 L_{adv} 的表示如下：

$$L_{adv}(e^s, e^k, \theta_D) = -log\ (p(d_{true} \mid e^s, e^k, \theta_D))$$

$$p(d_{true} \mid e^s, e^k, \theta_D) = exp\ (z_{true}) / \sum_{j \in D_{train}} exp\ (z_j)$$

其中 θ_D 是对抗学习模型的参数，z_j 表示鉴别器最后一层输出的结果，即对数几率（log probability, logits），z_{true} 表示正确的对数几率。

鉴别器模型的参数更新通过反向传播实现，为了阻止模型学习可能只有对特定领域才重要的表征，我们在模型中加入一个梯度逆转层（Gradient Reversal Layer，GRL）[65]。该层将标准梯度 $\delta L_{adv}/\delta\theta_{rel}$ 转换为 $-\delta L_{adv}/\delta\theta_{rel}$ ，其中 θ_{rel} 是表征学习模型的梯度。这导致了 θ_{rel} 最大化了域识别损失，同时仍然允许 θ_D 区分域。

5.1.1.4 排序学习

排序学习的目的是用来学习故事段落对应安全知识的最佳排序，从而实现检索。由于故事段落对应安全知识具有稀疏性，我们采用贝叶斯个性化排序（Bayesian Personalized Ranking，BPR）[56] 作为排序学习的算法。BPR 是基于矩阵分解的，不是做全局的评分优化，而是一种 pairwise 的排序算法，适合矩阵稀疏性。具体来说，在我们的任务中，BPR 针对每一个三元组 $< s, k^+, k^- >$，希望能够使故事段落 s 对相关知识 k^+ 和不相关知识 k^- 的差异更明显。训练过程中，针对每一个 epoch，我们都对故事段落对应的每个相关知识 k^+ 进行负采样，随机得到不相关知识 k^-，从而获取多

个三元组 $< s, k^+, k^- >$ 用于训练。因此，这样可以学到更加全面的排序关系，相比关注故事段落和所有安全知识拟合的具体数值损失最小，BPR 实现了更好的排序效果。对于特定故事段落 s，BPR 损失 L_{BPR} 的表示如下：

$$L_{BPR} = -\sum_{k^+ \in K_s} log\ (\hat{y}_{sk^+} - \hat{y}_{sk^-})$$

$$y_{sk^+} = e_s e_{k^+}^T$$

$$y_{sk^-} = e_s e_{k^-}^T$$

其中 K_s 是故事段落 s 对应的相关知识的集合。

5.1.2 实验配置

5.1.3 数据集

本实验采用第 4 部分描述的安全知识语料库和童话故事数据集，包括 141 条儿童安全知识和 30 个标注有安全知识和问答标签的故事段落。在本次实验中，我们主要使用童话故事数据集中每个故事段落及其对应的安全知识标签，其中标签的重要性不分先后，而问答标签将在下一个问答生成模块的实验中使用。

5.1.3.1 评测指标

为了评价算法的表现，我们采用多个常见的评价指标对算法给每个故事段落匹配到的知识排名列表进行评分，包括命中率（Hit Ratio，HR）[66]，归一化折损累计增益（Normalized Discounted Cumulative Gain，NDCG）[67]，准确率（Precision）和召回率（Recall）。其中 HR@K 可以直观地衡量测试安全知识标签是否出现在前 K 名的列表中；NDCG 通过给排名最高的测试项目分配更高的分数来解释命中者的位置；Precision 即为预测为正确的数据中真实值为正确的比例，衡量的是查准率；Recall 即为在所有的真实值为正确的数据中, 有多少能预测正确，衡量的是查全率。我们计算了所有故事段落样本的这四个指标，并报告了平均分数。

5.1.3.2 基准模型

我们将本方法与一些基准模型进行了比较，包括词频-逆文档频率（Term Frequency -Inverse Document Frequency，TF-IDF）[68]，以及关键词匹配算法。TF-IDF 是一种统计方法，用以评估一个字词对于一个语料库中的一份文档的重要程度，常用于信息检索与数据挖掘领域。在安全知识检索的任务中，我们首先通过 TF-IDF 得到安全知识和故事数据的表示，然后使用余弦相似度得到某条知识对于某输入段落的重要程度，从而得到每个输入故事段落对应的知识相关性的排序。关键词匹配算法则是将经 SentenceBert[54]处理后的故事段落的关键词和知识中用于检索的关键词（即危险

后果和场景）计算余弦相似度，最后同样得到每个输入故事段落对应的知识相关性的排序。

5.1.3.3 参数设置

我们基于 PyTorch 深度学习框架实现了基于神经协同过滤和对抗损失的知识检索算法。为了确定该方法的超参数，结合小样本数据集的特点，我们采用留一交叉验证（leave-one-out cross validation，LOOCV）法[63]对模型效果进行测试，从而确保结果的可靠性和稳定性。具体来说，对于每个故事段落样本，我们将它的数据作为测试集，并利用剩余样本数据作为训练集进行训练，最后取所有次数的平均作为最终的结果，依次调整模型的超参数。模型在训练的过程中是通过上述的联合损失进行学习的，并使用 Adam 优化器对模型进行优化，共计训练 100 个轮次。为了兼顾训练效率和训练后期的稳定性，我们将学习率设置为 1e-3，且分别在训练轮次为 5 的倍数处设置了衰减系数为 0.8 的学习率衰减，即每训练 5 个轮次学习率就变为原来的 0.8 倍。

5.1.4 结果分析

在四个测评指标上（其中 Hit@K 的 K 取 1 和 6，6 为故事段落对应的安全知识数量的平均向上取整），我们对基于神经协同过滤和对抗学习的知识检索算法 NCF+Adversarial Learning，以及两个基准算法 TF-IDF 和基于 SentenceBert 和余弦相似度的关键词匹配算法的效果进行了比较，如表 5-1 所示，可以看到，我们提出的模型显著超过两个基准模型。

表 5-1　不同的模型在安全知识检索任务上的效果比较

Table 5-1　Comparison of the Effectiveness of Different Models on the Safety Knowledge Retrieval Task

Model	Hit@1	Hit@6	NDCG	Precision	Recall
TF-IDF	0	0.259	0.124	0.043	0.046
SentenceBert+Cosine-sim	0.500	0.667	0.565	0.178	0.221
NCF+Adversarial Learning (Ours)	0.533	0.833	0.670	0.372	0.466

此外为了验证对抗损失部分的效果，我们进行了消融实验，对去掉对抗学习和反向传播部分后模型 NCF 的效果进行了测试，如表 5-2 所示。可以看到，去掉对抗学习模块后的模型在每个指标上都低于包含对抗学习的模型，从而验证了对抗学习模块的有效性。另外，NCF 模块的测试效果，除了 Hit@1 指标外，在其他指标上都显著高于两个基准模型。

表 5-2　对抗学习模块的消融实验结果

Table 5-2　Results of the Ablation Study on the Adversarial Learning Module

Model	Hit@1	Hit@6	NDCG	Precision	Recall
NCF	0.467	0.800	0.642	0.339	0.412
NCF+Adversarial Learning (Ours)	0.533	0.833	0.670	0.372	0.466

5.2　基于 GPT-3 小样本上下文学习的安全问答生成

5.2.1　方法描述

我们采用基于 GPT-3 的小样本上下文学习（Few-shot In-context Learning，FSIL）进行安全问答的生成，图 5-1 的 Question Answer Generation Module 展示了模型的基本框架。GPT-3 的输入提示信息是一个单词序列，它包括上下文 C（包括头部提示信息 h 和 n 个上下文示例（$\{x_i, y_i\}_{i=1}^{n}$）以及 QAG 的输入 x，如灰色框所示。其中 QAG 的输入是一段故事文本（“Story: ... ‘I’ll easily get rid of my apples. Here, I’ll give you one of them.” ”No,” said Snow White, ”I cannot accept anything from strangers.’ ...”），知识匹配阶段检索到的知识（“theme: ...; danger outcome: ⋯; danger action: ⋯; safe action: ...”）以及空白的 QA（“Q: A:”），如蓝色框中所示。QAG 的目标输出 y 是 QA（“Q:Just like Snow White accepted a poisoned apple from the stranger in the story, when/if a stranger offers you food, what should you do?;A: You should politely refuse anything given by strangers, because it could actually have something dangerous in it...”），如紫色框所示。此 QA 是以一种开放式的文本生成方式产生的，即 QA 可以包含从 GPT-3 的整个词汇表中选择的任意数量的单词。上下文 C 由头部提示信息 h 开头，它是一个对于所有样本都适用的固定字符串，如绿色框所示，具体将在下面的 Prompt Head 处描述。上下文 C 的剩余部分是 n 个上下文示例的字符串（$\{x_i, y_i\}_{i=1}^{n}$）的拼接，如红色框所示。然后我们将上下文 C 和 QAG 的输入 x 拼接起来共同生成提示信息。GPT-3 将构建好的提示信息的文本作为输入，隐式地从语言模型中检索和推理知识，以开放式文本生成任务的方式预测目标输出 y，即 QA。

5.2.1.1　头部提示信息

头部提示信息（Prompt Head）包含任务描述（Task），用于儿童安全教育的目的（Child Safety Education），结合故事场景（Story Scene）和现实场景（Real-life Scene）的提示信息，即 Prompt Head = Task + Child Safety Education + Story Scene + Real-life Scene + Template，具体内容如表 5-3 所示。

表 5-3　头部提示信息的具体内容

Table 5-3　The Specifics of Prompt Head

类型	内容
Task	Please generate a question and answer pair for children aged between 5 to 9 according to the bove context.
Child Safety Education	Please note that the question and answer pair is generated for children’safety education, not only for understanding the story itself.
Story Scene	Please summarize the related content in the story about the knowledge and combine it to the generated question in a natural way.
Real-life Scene	Please note that the question and answer pair should not be limited in the story, but encourage children to associate how to respond in related real life scenes, so that they can apply the safety knowledge into reality.
Template	Therefore, for example, you can use a template like ‘Just like xx happened in the story, we can associate a real-life scene like when/if you xxx, what you should xx?’

5.2.1.2　上下文示例

按照经验来说，提供更多的上下文例子通常导致更好的小样本学习的效果[42]。然而，新任务中可用的示例数量和模型的最大输入长度共同限制了输入提示中最大示例数量。在我们的任务中，使用的模型 GPT-3 输入序列实际上固定在 2048 个 token[42]，因此输入的序列的令牌（token）长度需要小于等于 2048，空的位置算法会自动补全。如此前所述，GPT-3 的输入部分包括提示头部信息（Prompt Head）、多个上下文例子（In-context Examples）、以及用于测试的故事段落、匹配的一条安全知识，注意需要预测输出的问题和答案不属于输入序列。其中 Prompt Head 的 token 长度 $L_{Prompt-Head}$ 为 126，一个上下文例子的最长 token 长度 $L_{In-context-Example} = L_{story} + L_{knowledge} + L_{question} + L_{answer}$，以每个样本为单位统一计算得 509，用于测试的故事段落和安全知识都取对应的最长 tokens 长度 $L_{test} = L_{story} + L_{knowledge}$，以每个样本为单位统一计算得 421。由于 GPT-3 输入序列的平均长度 $L_{Input-Prompt} = L_{Prompt-Head} + N * L_{In-context-Example} + L_{test}$ 需满足小于 GPT-3 最长的输入长度 2048，因此估计得到最大上下文示例数量 N 需小于等于 2.95，即 $N = 2$。在本任务中，由于生成的问题和安全知识强相关，而在生成问答的阶段，故事段落对应的安全知识已通过知识检索模

块得知，因此可用根据这些安全知识在候选 In-context examples 中进行选择，即挑选具有相同安全知识的样本作为 In-context Examples，其中一个样本包括故事段落、安全知识以及问答。经统计得知，训练测试数据集中知识标签出现的最多次数为 6，即使用知识标签进行检索时得到的 In-context Examples 的数量为 5（$> N$）。因此在我们的任务中，GPT-3 最大 token 输入长度更加限制我们可以取的上下文例子的数量 N，也就是说，存在有语言模型可以取的更多可用的例子的情况。

为了更好地利用可行的上下文示例，我们可以仔细挑选合适的上下文示例。在我们的任务中，我们试图在训练数据集中挑选和测试示例 x 具有类似问题特征的例子，这里的特征包括对应的知识标签，以及故事段落本身。首先，我们选择具有与示例 x（包括故事段落、知识）相同知识的示例（包括故事段落、知识、问题和答案）作为候选示例。然后，我们需要我们基于测试的故事段落和候选示例的故事段落计算相似度，来对候选示例进行筛选和补充。如果候选示例的数量大于 N，还需要通过故事段落的相似度进行二次筛选。如果候选示例的数量等于 N，则不需要进行额外的处理。如果候选示例的数量小于 N，则取故事段落的相似度排名靠前的样本作为候选示例的补充。

5.2.2 实验配置

5.2.2.1 数据集

本实验采用第 4 部分描述的安全知识语料库和童话故事数据集，包括 141 条儿童安全知识，以及 30 个标注有安全知识和问答标签的故事样本。在本次实验中，我们主要使用的是 30 个故事样本中的 157 个包含故事段落、安全知识、安全问题、安全答案的样本。注意，本实验的主要目的是检测问答生成模块的效果，因此故事段落对应的知识标签是给定的。

5.2.2.2 评测指标

为了评价算法的表现，我们采用常见的评价指标对算法给生成的问答对进行评分，包括召回率为导向的评估指标（Recall-Oriented Understudy for Gisting Evaluation，ROUGE）系列之一的 ROUGE-L[69]，以及基于语言模型的评价指标 BERTScore[70]。

Rouge-L 是一种基于最长公共子序列的摘要评估指标，它计算候选摘要和参考摘要之间的最长公共子序列的重合率，反映了句子级别的顺序和连贯性。它的优点是不需要指定 n-gram 的长度，适用于短摘要提取。在本文的问答生成任务中，Rouge-L 主要用于衡量生成答案和参考答案之间的表面形式上的重合度，反映出生成问答的完整性和流畅性。

BERTScore 是一种用于评估文本生成任务中结果与参考文本相似度的评估指标。

BERTScore 利用了 BERT 的预训练上下文嵌入，通过余弦相似度计算候选句子和参考句子中每个单词的相似度。BERTScore 具有以下优点：（1）它能够捕捉到单词之间的语义关系，而不仅仅是表面的匹配；（2）它能够兼顾召回率和准确率，而不偏向于任何一方；（3）它能够抵抗对抗性重写的干扰，即能够区分出与参考答案语义不一致的候选答案。在本文的问答生成任务中，生成问题的评估需要考虑候选答案和参考答案之间的语义相似度和表述多样性，因此 BERTScore 也是一种适合于问答生成任务的评估指标。

5.2.2.3 基准模型

我们将本方法与基准模型——基于 GPT-3 的零样本学习（Zero-shot Learning）的童话故事安全问答生成进行了比较，我们的方法和基线模型输入提示的内容如图 5-3 所示。可以看到，相较于文本提出的方法，基准模型的输入提示的内容缺少了 In-context examples。反之，依据给定的知识标签，基准模型将 Prompt Head，及其中一条给定的知识标签对应的知识（包含主题、危险后果、危险行为和安全行为）直接作为 GPT-3 的输入，生成对应的问答对。

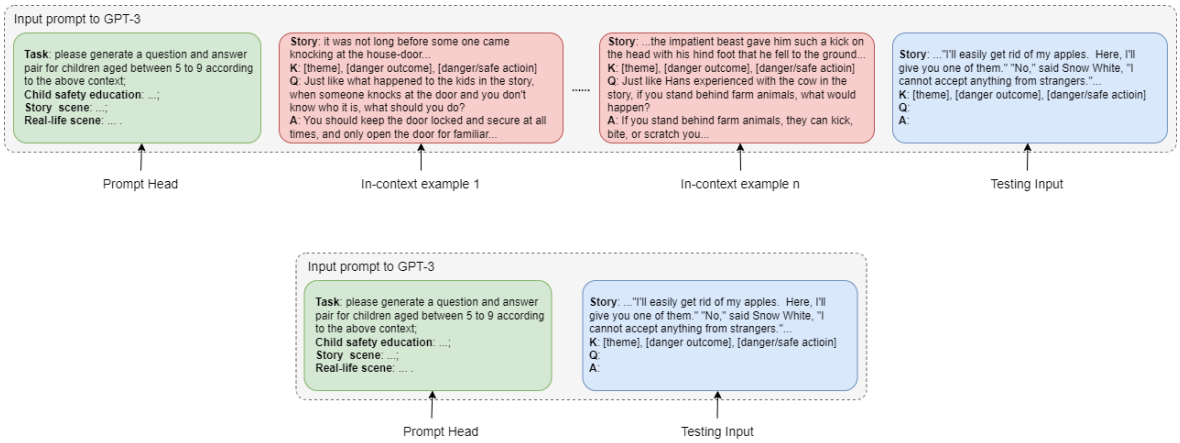

图 5-3　基于 GPT-3 推理时的小样本/零样本学习生成问答对的输入提示的比较

Comparason of Input Prompt Example for Generating Question-Answer Pair Based on Few-shot/Zero-shot Learning during GPT-3 Inference

5.2.2.4 参数设置

我们基于 PyTorch 深度学习框架实现了基于 GPT-3 小样本上下文学习的问答生成算法。由于模型无需参数的更新，因此我们只需要设计实验来验证模型的有效性。我们采用留一交叉验证（leave-one-out cross validation，LOOCV）法[63]对模型效果进行测试，从而确保结果的可靠性和稳定性。具体来说，对于每个故事-知识-问答对（总数为 157），我们将其作为测试示例，并利用剩余样本数据作为选取上下文示例的来源，为 GPT-3 生成测试示例的问答提供有效的提示。

5.2.3 结果分析

在基于 ROUGE-L 和 BertScore 各自的三个测评指标 Precison、Recall 和 F1-score 上，我们对基于 GPT-3 的小样本上下文学习问答生成（Zero-shot），以及基准模型——基于 GPT-3 的零样本上下文学习问答生成（Few-shot）算法的效果进行了比较，如表 5-4 所示。可以看到，我们提出的模型在所有的指标上表现优秀，且均显著超过基准模型。

表 5-4　不同的模型在安全问答生成任务上的效果比较

Table 5-4　Comparison of the Effectiveness of Different Models for Safety Question and Answer Generation Task

Model	R-Precision	R-Recall	R-F1	B-Precision	B-Recall	B-F1
Zero-shot	0.623	0.624	0.617	0.857	0.863	0.860
Few-shot (Ours)	0.812	0.815	0.811	0.956	0.958	0.957

6、 用户实验

在 5.1 和 5.2 两个模块中，我们分别验证了安全知识检索和安全问答生成两个模块的有效性，下面我们通过用户实验对整个算法流程的效果从多个维度进行验证，并与基线模型进行比较。具体来说，本章节依次描述了数据准备、数据收集、数据分析的方法和过程，最终展示了测评结果，并给出了未来的设计建议。

6.1 实验准备

为了模拟用户使用该方法的现实场景，我们在数据集中提前划分了 5 个数据样本用于用户实验。针对每个故事段落，我们分别使用基线模型——基于 GPT-3 的零样本学习，和我们提出的方法——基于 GPT-3 的安全知识增强的小样本上下文学习模型进行问答的生成，每种方法各生成 3 个问答对用作实验数据。具体来说，在使用我们方法的情况下，我们首先使用知识检索模块得到排名前 3 的知识，然后依据知识和故事段落从标注好的 157 个用于训练测试数据集中挑选上下文示例（In-context Examples），再依次组合头部提示信息（Prompt Head）、上下文示例、故事段落、以及安全知识一起作为 GPT-3 的输入，经过模型的推理从而得到对应的问答。我们对该故事对应的 3 个知识采用上述相同的方法进行处理，最终针对每个故事段落得到了对应的 3 个安全问答输出。在使用基线模型的情况下，我们直接将头部提示信息（Prompt Head）和故事段落作为输入，重复 3 次得到对应的问答输出。总之，我们为每位用户准备了 5 组实验，每组实验包含 1 个故事段落，2 组由不同算法生成的数量为 3 的问答对。

6.2 实验方法

6.2.1 数据收集

我们依次通过问卷和访谈的方法对形成性研究中的同一批用户进行数据的收集。问卷实验用于对生成的问答效果进行人工评测（Human Evaluation），包含两个维度的内容：1）技术维度，人工测评是自然语言处理研究领域中对生成的内容进行评分的方法之一，尤其上述自动测评部分的 ROUGE-L 和 BERTScore 指标在反映生成问答的语义匹配度和逻辑一致性方面具有局限性，因此我们这里将从常见的几个维度以及结合任务需求对两种方法生成的问答进行评分，包含问答可读性、知识相关性、问题相关性、答案相关性等等，从而对自动测评的效果进行有效的补充；2）儿童安全教育维度，侧重于问答对儿童安全教育这一任务场景的适用性，包含问答中安全知识的正确性/专业性、问答和儿童所处的现实场景的相关性、问答对儿童所处的现实场景的普适性、问答中展示故事中所体现的安全知识的全面性等等。访谈的目的是收集用户关于基于童话故事对儿童进行安全教育的总体使用体验，我们主要咨询了他

们的对该方法有用性/意义性、知识匹配程度、喜好程度等的看法，并和传统的儿童安全教育的方法进行了比较，最后询问了他们对未来相关接口/产品的使用意愿，在这个过程我们同时也收集了一些有关基于童话故事进行儿童安全教育的未来设计建议（design implications）。

问卷实验部分我们主要采用了 5 点李克特量表（5-point Likert scale）[51-52]。5 点李克特量表是一种用于衡量人们对某个主题或问题的态度或意见的量表，它由一组相关的陈述或问题构成，每个陈述或问题都有五个答案选项，分别是“非常同意”、“同意”、“不确定”、“不同意”和“非常不同意”。相比只提供两个答案选项的二元问题，5 点李克特量表可以更精确地收集数据，并且适用于多种统计方法和分析目的。和形成性研究类似，访谈部分采用的同样是是半结构化访谈。我们预先设计了一些一般性的问题，以适应具有不同经验水平的受访者。当受访者提到一些有趣的观点或先前的经验时，我们会跟进更多的细节和具体的例子。在 2023 年 4 月期间，所有访谈均由作者亲自或远程通过腾讯会议、微信语音通话或电话进行。每次填写问卷和访谈持续 40 分钟左右，用户获得 40 元酬金。在用户的许可下，我们通过录音记录了所有的访谈内容，并转录成中文。

6.2.2 数据分析

6.2.2.1 定量分析

在问卷实验和半结构化访谈的部分问题中，我们使用 5 点李克特量表（5-point Likert scale）[51-52] 收集到了有关用户的量化数据。在数据分析部分，首先我们对数据进行了分组，包含三个主题：（1）用户总体对两种不同方法生成的安全问答的爱好程度及比较（问卷）；（2）用户对生成问题在技术和儿童安全教育方面的评分（问卷）；（3）用户的总体的体验和看法，包括可用性等等（访谈）。然后，我们使用基于 Python[1]的 Numpy 库[2]进行数据处理，并使用 Matplotlib 库[3]进行数据可视化。

6.2.2.2 定性分析

定性分析采用了形成性研究中的同样的主题分析方法（Thematic Analysis）[50] 。即在收集数据的同时开始进行编码（coding）和分析，并使用 Excel 来编码和跟踪主题。在最初的分析阶段，我们仔细审查了收集到的数据，并标记了与研究问题相关的想法。然后，我们结合了最初的代码列表，反复迭代几次，并生成了最终的主题图。在测评结果的半结构化访谈数据分析部分，我们将介绍这些主题的全部细节，并使用有代表性的引用来支持我们的发现。为了保护参与者的身份，我们使用 PX 来表示参

[1]https://www.python.org/

[2]https://numpy.org/

[3]https://matplotlib.org/

数。

6.3 实验结果

6.3.1 问卷结构数据分析

我们在问卷实验中使用了 5 点李克特量表（5-point Likert scale）[51-52] 的形式收集了用户对我们提出的方法（ours）和基准方法（baseline）生成的问题的评分，包含用户对两组问答喜好程度的分析、技术维度评分以及儿童安全教育相关的评分。

6.3.1.1 喜好程度分析

我们使用条形图的形式可视化了用户对两种方法生成问答的喜爱程度，如图 6-1 和图 6-2 所示。其中纵坐标的代表 5 点李克特量表的数值，数值越高表示用户的喜爱程度越高，红蓝色分别代表了 baseline 和 ours 方法的效果。在图 6-1 中，我们可视化了各被试对所有故事两组问答的平均喜好程度，横坐标代表被试的实验编号。可以看到，除了用户 P3 稍微偏向喜好 baseline 的问答外，其余 6 名用户喜好 ours 远超过于 baseline。在图 6-2 中，我们统计了全部被试对所有故事两组问答的平均喜好程度和方差，横坐标代表生成问答所使用的方法类型，黑色的误差棒的长度代表着方差的大小。可以看到，ours 方法的平均用户喜好在 5 点李克特量表上表示喜欢到非常喜欢之间，数值比 baseline 方法的高一个点以上，同时用户喜好 ours 方法的方差小于用户喜欢 baseline 方法的方差。根据以上的分析可以初步得出结论，用户普遍喜欢 ours 方法，另外相对于 baseline 方法，用户总体上更加喜欢 ours 方法。

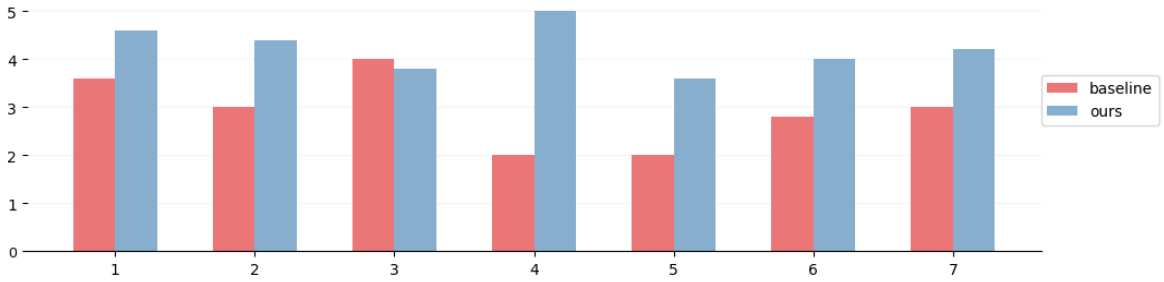

图 6-1 各用户对所有故事两组问答的平均喜好程度

Figure 6-1 Average Preference of Each User for Both Sets of QA for All Stories

6.3.1.2 技术维度评分的分析

我们使用雷达图的形式可视化了全部用户对所有故事两组问答在技术上维度的平均评分，如图 6-3 所示，包含问答可读性（即单词语法等使用的正确性，QA Readability）、知识相关性（即安全知识和故事本身的匹配程度，Knowledge Relevance）、问题相关性（即问题和故事文本以及安全知识的匹配程度，Question Relevance）、以及答案相关性（生成的答案与问题的匹配程度，Answer Relevance）。其中红色折线

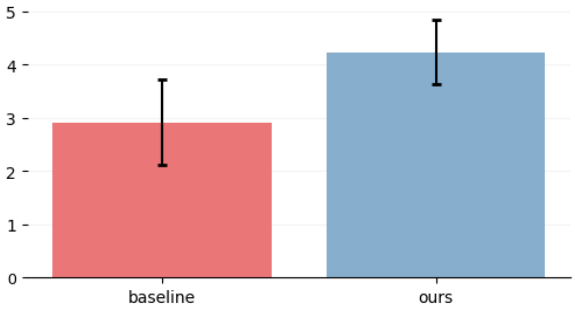

图 6-2　全部用户对所有故事两组问答的平均喜好程度和方差

Figure 6-2　Mean Preference and Variance of All Users for Both Sets of QA for All Stories

条表示用户对 baseline 方法生成问答的各维度评分，蓝色折线条表示用户对 ours 方法生成问答的各维度评分，而对应的灰色网格线，即 5 点李克特量表，数值越高代表着用户对问答在该维度上的认可度越高。可以看到，ours 方法在各个维度上的评分数值都大于 4，普遍高于 baseline 方法，尤其在 Question Relevannce、QA Readability 和 Knowledge Relevance 这三方面。根据算法部分的实验结果，可以推测出是知识检索算法和上下文小样本学习的效果。根据以上的分析可以初步得出结论，用户对 ours 方法在技术维度上比较认可，另外，相对于 baseline 方法，用户对 ours 方法在 Question Relevannce、QA Readability 和 Knowledge Relevance 这三方面认可程度的差异性更大。

图 6-3　全部用户对所有故事的两组问答在技术维度上的平均评分

Figure 6-3　Average Scores of All Users for Both Sets of QA of All Stories on the Technical Dimension

6.3.1.3　儿童安全教育维度评分的分析

与技术维度的评分分析类似，我们使用雷达图形式可视化了全部用户对所有故事两组问答在儿童安全教育维度上的平均评分，如图 6-4 所示，包括问答中安全知识的正确性/专业性（Knowledge Expertise）、问答和儿童所处的现实场景的相关性（Reality Relevance）、问答对儿童所处的现实场景的普适性（Reality Universality）、以及问答中

展示故事体现的安全知识的全面性（Knowledge Comprehensiveness）。可以看到 ours 方法在各个维度上的评分数值都大于 4，显著高于高于 baseline 方法，同样可以推测出是知识检索算法和上下文小样本学习的效果。根据以上的分析可以初步得出结论，用户对 ours 方法在儿童安全教育的维度上比较认可，另外相对于 baseline 方法，ours 方法在儿童安全教育维度上的优势比在技术上的优势明显。

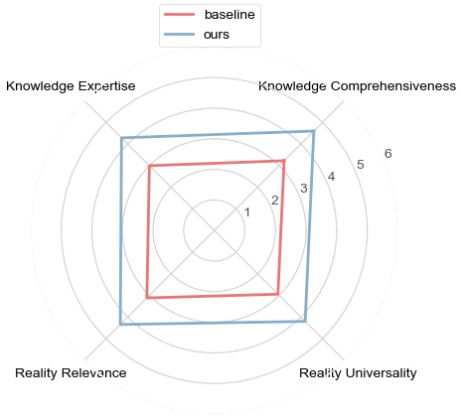

图 6-4　全部用户对所有故事的两组问答在儿童安全教育维度上的平均评分

Figure 6-4　Average Scores of All Users for Both Sets of QA of All Stories on the Child Safety Education Dimension

6.3.2　半结构化访谈数据分析

我们的研究结果是基于对访谈内容进行主题分析得到的主题图。由于访谈的目的主要是了解用户对于本文提出方法的看法以及建议，因此我们的主题图包括有用性（usefulness）、知识的匹配性（knowledge matching）、喜爱程度（likeness）、与传统方法的比较（better than tradition）、使用意愿（future use）和一些未来的设计建议（design implication）。除了文字结论外，我们还使用口头版 5 点李克特量表收集了用户对部分问题看法的量化信息，如图 6-5 所示。

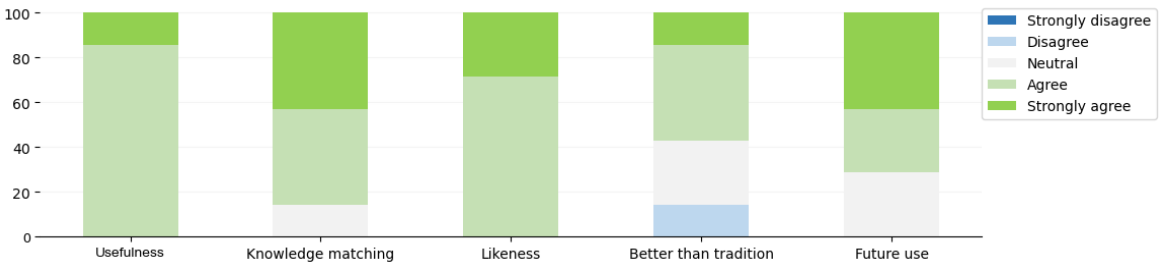

图 6-5　全部用户对基于童话故事的儿童安全教育在各维度上的平均看法

Figure 6-5　Average Scores of all users on Each dimension of the Fairy Tale based Child Safety Education

6.3.2.1 方法的有用性

如图 6-5 中 Usefulness 部分所示，所有的用户都觉得该方法对于儿童安全教育是有用的，有意义的。例如 P2 表示童话故事场景从心理上比较容易让儿童接受，而不会像传统说教中产生逆反心理。“这个场景比较生动，小孩子也会比较愿意听，就平时你直接跟小孩讲，小孩嫌烦，你讲童话故事讲完的时候跟小孩反思两句，我觉得这比较容易让小孩接受。”。P5 肯定的原因是童话故事和现实场景的匹配性，“童话故事体现的场景是跟这个年龄段的儿童可能会遇到的这个场景是有符合的，是有相干的，所以我觉得用故事来教育是绝对是有用。”而 P4 从儿童教育的角度提出了一些其他的看法，同时也指出了它的局限性，“童话故事它的一个作用是它可以给孩子提供一个比较简单的想象场景，然后提供一个比较简单的动作指导，让它变得不那么难理解，但是它是有局限性的，它终究还是一个文本的内容，它不会直观的给孩子呈现出来。”总之，大家都觉得该方法是有用的有意义的，无论从儿童的接受程度、场景的现实意义、儿童的教育等等方面，但也存在一定的局限性有待改进。

6.3.2.2 知识的匹配性

如图 6-5 中 Knowledge matching 部分所示，绝大部分用户都觉得该方法的知识匹配度较高。例如 P1 表示“基本准确”，P3 表示“应该都匹配到了”。另外 P2 和 P5 都强调了额外的人工参与的重要性。P2 表示可以在亲子共读的过程中有选择性地进行提问，“虽然在三个问题中偶尔存在比如一个不准确问题的情况，但问题不大，比如我在场就可以在问正确的那个”。P5 也强调了类似的观点，“可以让它（算法）来生成一些问题，然后还是要结合人工来进行一个筛选出来，就是一些比较优质的那种经过人工核验以后那种比较优质的问题，然后他们再进行一个提供，就给儿童们来做，这样的话我觉得是一个比较好的方法。”总之，大家大部分大都觉得文本和知识的匹配程度较高，而对于存在的瑕疵，用户认为可以通过家长共读，或者提前的人工筛选来解决。

6.3.2.3 方法的喜爱程度

如图 6-5 中的 Likeness 部分所示，所有的用户都喜爱该方法。一方面，有的用户表示在讲故事的过程中进行安全教育拓展了他们进行儿童安全教育的思路，例如 P1 表示“因为之前没有想到了童话故事，还能进行安全教育。我觉得这是一个很新颖的点，这带给了我启发！”，P2 也表示“主要是我觉得开阔了我的思维！”。另一方面，有的用户认为该方法让新技术更好得惠及大家，例如 P5 表示“我觉得这个确实是一个很好的东西，虽然它有点瑕疵，但我还是很喜欢！首先因为 OpenAI 它做的这么好，现在这些大模型实际上对我们很触手不可及，现在我们每个人都可以用！”还

有一些用户表示该方法用于亲子共同阅读的过程中可以增加和孩子的互动，例如 P6 表示“如果我不在现场，小孩可以学习面对如何在各种情况下保护自己，我觉得挺好的。如果我在场明它可以给我提供参考素材，我再对小孩进行提问…在讲故事时我不可能去光讲故事吧，我还希望跟小孩儿去互动，一方面在安全方面对孩子进行了指导，另一方面也是做一个沟通，算是增加感情的一种方式，我觉得这样的工具还是非常不错的！”这和 Storybuddy[20] 中得出的结论类似。也有用户表示生成的问题很好得做到了故事情节和现实场景的结合和分离之间的平衡，例如 P6 表示“结合是为了从故事情节中引出安全知识，但为了避免一个模糊的浪漫的故事被完全定性成了某种安全问题，问答采用了联想的方法将情节和问答进行了分离，例如读到某个情节可以联想到某个现实场景下的安全问题。”另外，还有一些用户表示虽然很喜欢，也有待改进，例如 P4 表示“总体上好的，它也有待改进的，例如文字不够直观，动画或者图像更好一些，所以也具有一定局限性。”总之，尽管有一些瑕疵，所有的用户都表示了对该方法的喜爱，无论是处于新的安全教育思路、新的技术、还是促进和孩子的互动等等。

6.3.2.4 与传统方法的比较

如图 6-5 中的 Better than tradition 部分所示，用户认为该方法比传统方法好的观点比较分散，其中认可和非常认可的比例占 57%，而中立及以下的比例占 43%。其中部分用户认为相对传统方法，该方法让儿童更感兴趣，例如 P1 表示“它比较让人有兴趣一些，而且印象更深刻一些”，P3 也表示“安全教育宣传册现在说实话很多拿回来就拿来垫桌子…小孩子们就喜欢看动画片、听故事，你给他看个（安全教育的）纪录片，他看了看了估计睡着了。”还有用户认为在讲故事的过程中进行安全教育在传统的安全教育中是一种有效的补充，例如 P6 表示“是故事的延申，也是一种安全教育的补充，说补充是因为它不够系统全面…但是一旦儿童的安全知识体系搭建起来，这种讲故事其实就是在强化它的一些细枝末节上，都可以让这种体系或者说进一步更加的深刻。”

然而有的用户还是更加倾向实地演练，例如 P5 表示“线下来说可能会有更多的那种体验方式，比如说不仅有讲，可能还会让你去参与互动，比如消防演练。“也有用户认为该方法的互动因素不多，和传统的讲解形式差别不大，需要更加侧重引导和互动，例如 P4 表示“我们在做幼教，特别是 K12 的时候，其实会特别注重跟孩子互动的一个过程，而且互动要做得非常频繁它才会有效。因为成年人的思维模式真的你很难和 K12 阶段的儿童去匹配，你是不能一直给他灌输你的想法的，你需要和他进行一个思维的交换，所以说你一定要多跟他去互动。（在我们的方法中）设计互动会是更好的一个方式，如果没有的话，其实我觉得这种问答的形式跟我们单纯的文本讲

知识的形式，它其实区别不是没有那么大…也就是说他应该有一些那种引导式的，然后小孩说一个，然后你再给他反馈，然后再就来回互动。”其中在[71]中也有得出类似的结论，“现有的针对儿童的语音接口（Voice User Interface，VUI）应用程序未能进行开放式对话，并提供长期的来回机会。”

总之，基于讲故事的安全教育是传统安全教育的一种有效补充，尤其是在唤起儿童兴趣、日常强化儿童安全意识的角度，但在交互方面还存在着很大的局限性。

6.3.2.5 方法的使用意愿

如图 6-5 中的 Future use 部分所示，大部分用户（72%）都希望在未来使用到此技术用于儿童的安全教育，例如 P5 表示“这挺好的，我肯定是会去乐于接受这个技术的”。但仍有一些用户持中立态度，例如 P4 强调“我觉得我不排斥它，但是也不会很积极的去主动用，除非它能提供更多的一些互动。”因此，尽管大多数用户都对未来使用该技术表示期待，但我们还可以在交互上提供更多的改进来满足更大范围用户的需求，详细的改进意见我们将在设计建议部分详细说明。

6.4 设计建议

根据以上用户测评，我们总结了一些设计建议。（1）从设计主题来看，目前的基于童话故事的安全教育方法对于专业的安全教育来说不够系统、不够全面，但可以作为童话故事的延申，或者儿童安全教育的有效补充；（2）从呈现方式来看，相对于文字，动画更能给予直观的呈现，而实地演习更能让人印象深刻但比较受限，因此后续可以考虑以动画交互、或者是虚拟现实技术（Virtual Reality，VR）[72]和增强现实技术（Augument Reality，AR）[73]代替现实等方式优化给与儿童呈现的内容；（3）从交互设计来看，可以考虑设计单独面向儿童、以及亲子共读两种模式，在第一种模式中讲故事的代理直接对小孩儿进行提问，在第二种模式中讲故事的代理将为父母提供候选的安全知识及对应的问答，供父母选择性提问，促进父母和儿童的交互；（4）从技术设计来看，单一的问答被提出不能显著区别于传统的传授型安全教育，最好能够以来回互动的方式对儿童进行引导和反思，可以考虑使用近期研究的热点大语言模型实现。

7、 总结和展望

7.1 总结

本文探索了一种将儿童安全教育融合进给儿童讲故事的过程中的方法，旨在提升儿童安全教育的效果。首先，本文通过形成性研究，调研了儿童安全教育的现状和需求，提出了儿童安全教育的设计建议，并根据这些建议设计了数据和算法两个部分。然后，我们通过半自动化的方式构建了儿童安全知识语料库，并标注了基于FairytaleQA[21] 的安全知识匹配和问答生成数据集。其次，我们设计了外部安全知识增强的童话故事安全问答生成方法，包括基于 NCF 和对抗损失的知识检索模块和基于 GPT-3 小样本上下文学习的问答生成模块，并通过实验验证了这两个模块的有效性。最后，我们通过访谈和问卷两种方式的用户实验，评估了该方法的整体效果，并通过定性和定量分析得到了用户较为积极的反馈，以及未来的设计建议。

7.2 局限性和未来工作

本文的工作有四个局限性，分别体现在用户实验方面、数据准备方面、模型设计方面、以及交互设计层面。在用户实验方面，受限于时间等各种条件我们只对 7 名被试进行了实验，实验结果可能存在一定的偏差。另外，我们的研究采用了滚雪球抽样（Snowball Sampling）[49] 的方法来招募参与者，虽然已经采用了各种策略来确保参与者的多样性，但该方法在参与者的多样性方面仍然存在局限性，例如我们没有招募儿童等其他利益相关者。我们未来的调查会扩大用户招募的数量和范围，纳入更多数量、更多不同角色的利益相关者的观点。在数据准备方面，儿童安全知识语料库和童话故事安全问答数据集的规模和质量还有待提高，需要更多的专业人士和众包工人合作参与构建和验证。在模型设计方面，知识检索和问答生成模块还可以进行进一步的优化，考虑到最新最优的方法，以提高知识匹配和问答生成的效果。交互设计层面，针对交互形式单一，仅存在单次问答的方式，缺乏来回的互动的问题，我们未来的工作会考虑采用大模型的方法实现讲故事智能代理对儿童的引导和来回互动，同时设计单人阅读和亲子共读两种模式帮助替代或者辅助家长进行安全教育。此外，童话故事的呈现方面文字存在不够直观的特点，可以考虑采用动画和 VR 等形式进一步提升用户的交互体验。

参考文献

[1] KENDRICK D, YOUNG B, MASON-JONES A J, et al. Home safety education and provision of safety equipment for injury prevention[J]. Evidence-based child health: a Cochrane review journal, 2013, 8(3): 761-939.

[2] CARLIN J B, TAYLOR P, NOLAN T. School based bicycle safety education and bicycle injuries in children: a case-control study[J]. Injury Prevention, 1998, 4(1): 22-27.

[3] RACE K E. Evaluating pedestrian safety education materials for children ages five to nine[J]. Journal of school health, 1988, 58(7): 277-281.

[4] DUPERREX O, BUNN F, ROBERTS I. Safety education of pedestrians for injury prevention: a systematic review of randomised controlled trials[J]. Bmj, 2002, 324(7346): 1129.

[5] WILKS J, KANASA H, PENDERGAST D, et al. Beach safety education for primary school children[J]. International journal of injury control and safety promotion, 2017, 24(3): 283-292.

[6] SOLOMON R, GIGANTI M J, WEINER A, et al. Water safety education among primary school children in Grenada[J]. International journal of injury control and safety promotion, 2013, 20(3): 266-270.

[7] LURIA J W, SMITH G A, CHAPMAN J I. An evaluation of a safety education program for kindergarten and elementary school children[J]. Archives of pediatrics & adolescent medicine, 2000, 154(3): 227-231.

[8] STOKES S C, MCFADDEN N R, SALCEDO E S, et al. Firearm injuries in children: a missed opportunity for firearm safety education[J]. Injury prevention, 2021, 27(6): 554-559.

[9] ZHOU W J, XU X L, LI G, et al. Effectiveness of a school-based nutrition and food safety education program among primary and junior high school students in Chongqing, China[J]. Global health promotion, 2016, 23(1): 37-49.

[10] GOLDMAN J D, COLLIER-HARRIS C A. School-based reproductive health and safety education for students aged 12–15 years in UNESCO’s (2009) International Technical Guidance[J]. Cambridge Journal of Education, 2012, 42(4): 445-461.

[11] RICCI F. Strategies for teaching safety education to children with special needs[J]. International journal of trauma nursing, 2000, 6(4): 129-132.

[12] LUO H, YANG T, KWON S, et al. Performing versus observing: Investigating the effectiveness of group debriefing in a VR-based safety education program[J]. Computers & Education, 2021, 175: 104316.

[13] O’NEILL S, FLEER M, AGBENYEGA J, et al. A cultural-historical construction of safety education programs for preschool children: Findings from SeeMore Safety, the pilot study[J]. Australasian Journal of Early Childhood, 2013, 38(2): 74-84.

[14] LEE J N, JUNG M, PARK J W. Effects of school safety education on safety behavior among elementary school students[J]. Child Health Nursing Research, 2006, 12(4): 506-513.

[15] KITAMURA Y. The possibility of holistic safety education in Japan: from the perspective of education for sustainable development (ESD)[J]. IATSS research, 2014, 38(1): 40-47.

[16] BARR J, SALTMARSH S, KLOPPER C. Early childhood safety education: An overview of safety curriculum and pedagogy in outer metropolitan, regional and rural NSW[J]. Australasian Journal of Early Childhood, 2009, 34(4): 31-36.

[17] HARTIKAINEN H, IIVARI N, KINNULA M. Children’s design recommendations for online safety education[J]. International Journal of Child-Computer Interaction, 2019, 22: 100146.

[18] EDWARDS S, NOLAN A, HENDERSON M, et al. Young children’s everyday concepts of the internet: A platform for cyber-safety education in the early years[J]. British journal of educational technology, 2018, 49(1): 45-55.

[19] KOLKO D J. Efficacy of cognitive-behavioral treatment and fire safety education for children who set fires: Initial and follow-up outcomes[J]. The Journal of Child Psychology and Psychiatry and Allied Disciplines, 2001, 42(3): 359-369.

[20] ZHANG Z, XU Y, WANG Y, et al. Storybuddy: A human-ai collaborative chatbot for parent-child interactive storytelling with flexible parental involvement[C]. in: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems. 2022: 1-21.

[21] XU Y, WANG D, YU M, et al. Fantastic Questions and Where to Find Them: FairytaleQA–An Authentic Dataset for Narrative Comprehension[C]. in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2022: 447-460.

[22] SIM S, BERTHELSEN D. Shared book reading by parents with young children: Evidence-based practice[J]. Australasian Journal of Early Childhood, 2014, 39(1): 50-55.

[23] ZHAO Z, HOU Y, WANG D, et al. Educational Question Generation of Children Storybooks via Question Type Distribution Learning and Event-centric Summarization[C]. in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2022: 5073-5085.

[24] YAO B, WANG D, WU T, et al. It is AI’s Turn to Ask Humans a Question: Question-Answer Pair Generation for Children’s Story Books[C]. in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2022: 731-744.

[25] ZEVENBERGEN A A, WHITEHURST G J. Dialogic reading: A shared picture book reading intervention for preschoolers[J]. On reading books to children: Parents and teachers, 2003, 177: 200.

[26] GANOTICE JR F A, DOWNING K, MAK T, et al. Enhancing parent-child relationship through dialogic reading[J]. Educational Studies, 2017, 43(1): 51-66.

[27] XU Y, WANG D, COLLINS P, et al. Same benefits, different communication patterns: Comparing Children's reading with a conversational agent vs. a human partner[J]. Computers & Education, 2021, 161: 104059.

[28] LIU H, SINGH P. ConceptNet—a practical commonsense reasoning tool-kit[J]. BT technology journal, 2004, 22(4): 211-226.

[29] SPEER R, CHIN J, HAVASI C. Conceptnet 5.5: An open multilingual graph of general knowledge [C]. in: Proceedings of the AAAI conference on artificial intelligence: vol. 31: 1. 2017.

[30] HWANG J D, BHAGAVATULA C, LE BRAS R, et al. (comet-) atomic 2020: On symbolic and neural commonsense knowledge graphs[C]. in: Proceedings of the AAAI Conference on Artificial Intelligence: vol. 35: 7. 2021: 6384-6392.

[31] AUER S, BIZER C, KOBILAROV G, et al. Dbpedia: A nucleus for a web of open data[C]. in: The Semantic Web: 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007+ ASWC 2007, Busan, Korea, November 11-15, 2007. Proceedings. 2007: 722-735.

[32] BOSSELUT A, RASHKIN H, SAP M, et al. COMET: Commonsense Transformers for Automatic Knowledge Graph Construction[C]. in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019: 4762-4779.

[33] TALMOR A, HERZIG J, LOURIE N, et al. CommonsenseQA: A Question Answering Challenge Targeting Commonsense Knowledge[C]. in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). 2019: 4149-4158.

[34] MIHAYLOV T, CLARK P, KHOT T, et al. Can a Suit of Armor Conduct Electricity? A New Dataset for Open Book Question Answering[C]. in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018: 2381-2391.

[35] MARINO K, RASTEGARI M, FARHADI A, et al. Ok-vqa: A visual question answering benchmark requiring external knowledge[C]. in: Proceedings of the IEEE/cvf conference on computer vision and pattern recognition. 2019: 3195-3204.

[36] UEHARA K, HARADA T. K-VQG: Knowledge-aware Visual Question Generation for Commonsense Acquisition[C]. in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2023: 4401-4409.

[37] XU Y, ZHU C, XU R, et al. Fusing Context Into Knowledge Graph for Commonsense Question Answering[C]. in: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. 2021: 1201-1207.

[38] YASUNAGA M, REN H, BOSSELUT A, et al. QA-GNN: Reasoning with Language Models and Knowledge Graphs for Question Answering[C]. in: Proceedings of the 2021 Conference of

the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2021: 535-546.

[39] FENG Y, CHEN X, LIN B Y, et al. Scalable Multi-Hop Relational Reasoning for Knowledge-Aware Question Answering[C]. in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020: 1295-1309.

[40] JIA X, WANG H, YIN D, et al. Enhancing question generation with commonsense knowledge [C]. in: Chinese Computational Linguistics: 20th China National Conference, CCL 2021, Hohhot, China, August 13–15, 2021, Proceedings. 2021: 145-160.

[41] YANG Z, GAN Z, WANG J, et al. An empirical study of gpt-3 for few-shot knowledge-based vqa [C]. in: Proceedings of the AAAI Conference on Artificial Intelligence: vol. 36: 3. 2022: 3081-3089.

[42] BROWN T, MANN B, RYDER N, et al. Language models are few-shot learners[J]. Advances in neural information processing systems, 2020, 33: 1877-1901.

[43] OpenAI. GPT-4 Technical Report[J]. arXiv preprint arXiv:2303.08774, 2023.

[44] SCHICK T, SCHÜTZE H. Exploiting Cloze-Questions for Few-Shot Text Classification and Natural Language Inference[C]. in: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. 2021: 255-269.

[45] NOROUZI S, TANG K, CAO Y. Code generation from natural language with less prior knowledge and more monolingual data[C]. in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers). 2021: 776-785.

[46] LIU J, SHEN D, ZHANG Y, et al. What Makes Good In-Context Examples for GPT-3?[C]. in: Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures. 2022: 100-114.

[47] LIU P, YUAN W, FU J, et al. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing[J]. ACM Computing Surveys, 2023, 55(9): 1-35.

[48] MIN S, LYU X, HOLTZMAN A, et al. Rethinking the Role of Demonstrations: What Makes In-Context Learning Work?[J]. arXiv preprint arXiv:2202.12837, 2022.

[49] GOODMAN L A. Snowball sampling[J]. The annals of mathematical statistics, 1961: 148-170.

[50] BRAUN V, CLARKE V. Thematic analysis.[M]. American Psychological Association, 2012.

[51] JOSHI A, KALE S, CHANDEL S, et al. Likert scale: Explored and explained[J]. British journal of applied science & technology, 2015, 7(4): 396.

[52] DAWES J. Do data characteristics change according to the number of scale points used? An experiment using 5-point, 7-point and 10-point scales[J]. International journal of market research, 2008, 50(1): 61-104.

[53] LOÈVE M. Probability theory[M]. Courier Dover Publications, 2017.

[54] REIMERS N, GUREVYCH I. Sentence-bert: Sentence embeddings using siamese bert-networks [J]. arXiv preprint arXiv:1908.10084, 2019.

[55] MINSKY M, PAPERT S. Perceptrons: An Introduction to Computational Geometry[M]. MIT press, 1969.

[56] RENDLE S, FREUDENTHALER C, GANTNER Z, et al. BPR: Bayesian personalized ranking from implicit feedback[J]. arXiv preprint arXiv:1205.2618, 2012.

[57] ROBERTSON S, ZARAGOZA H, et al. The probabilistic relevance framework: BM25 and beyond [J]. Foundations and Trends® in Information Retrieval, 2009, 3(4): 333-389.

[58] HE X, LIAO L, ZHANG H, et al. Neural collaborative filtering[C]. in: Proceedings of the 26th international conference on world wide web. 2017: 173-182.

[59] MITRA B, CRASWELL N. An Introduction to Neural Information Retrieval[M]. Now Publishers Inc, 2018.

[60] GANIN Y, USTINOVA E, AJAKAN H, et al. Domain-adversarial training of neural networks[J]. The Journal of Machine Learning Research, 2016, 17(1): 2096-2030.

[61] TZENG E, HOFFMAN J, SAENKO K, et al. Adversarial discriminative domain adaptation[C]. in: Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 7167-7176.

[62] COHEN D, MITRA B, HOFMANN K, et al. Cross domain regularization for neural ranking models using adversarial learning[C]. in: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. 2018: 1025-1028.

[63] WONG T T. Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation[J]. Pattern Recognition, 2015, 48(9): 2839-2846.

[64] ZHANG Z, SABUNCU M R. Generalized cross entropy loss for training deep neural networks with noisy labels[C]. in: Advances in neural information processing systems. 2018: 8778-8788.

[65] GANIN Y, LEMPITSKY V. Unsupervised domain adaptation by backpropagation[C]. in: International conference on machine learning. 2015: 1180-1189.

[66] YANG X, STECK H, GUO Y, et al. On top-k recommendation using social networks[C]. in: Proceedings of the sixth ACM conference on Recommender systems. 2012: 67-74.

[67] DISTINGUISHABILITY C. A Theoretical Analysis of Normalized Discounted Cumulative Gain (NDCG) Ranking Measures[J]., 2013.

[68] RAMOS J, et al. Using tf-idf to determine word relevance in document queries[C]. in: Proceedings of the first instructional conference on machine learning: vol. 242: 1. 2003: 29-48.

[69] LIN C Y. Rouge: A package for automatic evaluation of summaries[C]. in: Text summarization branches out. 2004: 74-81.

[70] ZHANG T, KISHORE V, WU F, et al. Bertscore: Evaluating text generation with bert[J]. arXiv preprint arXiv:1904.09675, 2019.

[71] XU Y, BRANHAM S, DENG X, et al. Are Current Voice Interfaces Designed to Support Children's Language Development?[C]. in: Proceedings of the 2021 CHI conference on human factors in computing systems. 2021: 1-12.

[72] BURDEA G C, COIFFET P. Virtual reality technology[M]. John Wiley & Sons, 2003.

[73] BILLINGHURST M, CLARK A, LEE G, et al. A survey of augmented reality[J]. Foundations and Trends® in Human–Computer Interaction, 2015, 8(2-3): 73-272.

致谢

在华东师范大学计算机科学与技术学院度过的四年时光，是我人生中最宝贵的财富之一。在这里，我受益于学院在人工智能、人机交互等领域的优秀师资和先进设备，也感受到了学校自由开放、浓厚人文的教育理念。在完成本科毕业论文之际，我想向所有关心和帮助过我的老师、同学、朋友和家人表示衷心的感谢。

首先，我要感谢我的论文指导老师孙玉灵老师。孙老师不仅在专业知识和科研方法上给予了我悉心指导，还以其严谨务实的工作态度和宁静致远的学术追求影响着我。孙老师对我的教诲和关怀，将成为我今后工作和生活中的宝贵财富。

感谢本科期间与我合作过的周爱民老师、王大阔老师、麻晓娟老师、刘峰博士生、马帅博士生、张嘉淏学长等。他们在科研项目中为我提供了宝贵的经验和建议，在生活中也给予了我很多支持和鼓励。他们是我走进科研世界的引路人，也是我成长过程中的良师益友。

感谢实验室的所有同学们，他们在学习上互相切磋，在生活上互相照顾，在困难中互相鼓励，在快乐中互相分享。他们让我度过了一个充满收获和欢乐的本科生涯。

感谢四年来一直陪伴我的朋友们，他们在我遇到挫折时给我安慰，在我取得进步时给我祝福，在我迷茫时给我指引，在我忙碌时给我陪伴。他们是我人生路上最可靠的伙伴，也是我最珍惜的财富。

感谢我的家人和男友，他们是我最坚强的后盾，也是我最温暖的港湾。感谢家人对我的无私奉献和无条件支持，感谢男友对我的理解和鼓励。他们让我感受到了爱的力量，也激励着我不断前进。

最后，感谢本文引用著作中所有作者对本领域做出的重要贡献，以及百忙之中审阅本论文并提出宝贵意见的专家学者们。