arXiv:2303.03235v1 [cs.HC] 6 Mar 2023

支持文本解读的论证图可视化

预印本

Hanadi Mardah*, 1, Oskar Wysocki[1,2],Markel Vigo[1],和 André Freitas[1,2,3]

[1]计算机科学系，曼彻斯特大学，曼彻斯特，英国 [2]数字实验癌症医学团队，癌症生物标志物中心，CRUK曼彻斯特研究所，英国[3]Idiap研究所，马蒂尼，瑞士

摘要

自然语言处理（NLP）方法近来的发展，尤其是论证挖掘领域的发展，有望改变我们与文本互动的方式，并支持对复杂话语和辩论的解读与分析。复杂论证的图形化可视化能否促进对论据进行更批判性的解读？本研究重点分析论证图（AG）与常规文本相比，在支持论证解读方面的影响。我们发现，在大多数用户体验问卷量表（UEQ）指标上，AG的表现优于外在指标；在NASA-TLX工作负荷的所有维度上也优于常规文本，唯独时间需求和身体需求方面例外。尽管文本模型和AG模型在批判性解读方面取得了相近的结果（包括工作记忆表现和参与者决策的改变），但更多参与者更喜欢AG模型。解读过程涉及在AG中参考论证图式（与批判性问题（CQs）相关联）。值得注意的是，当参与者对论证主题较不熟悉时，他们会选择更多批判性问题（借助AG中的论证图式），使AG图式在某些量表上对解读过程表现出（相对的）支持作用。因此，AG被认为能够为论证解读带来更具批判性的方法，尤其适用于不熟悉的主题。根据本研究的25名参与者的结果，论证图似乎对论证解读过程总体产生了积极影响。

关键词 论证结构；Walton分类方案；论证挖掘；论证图；可视化图，图交互

1 引言

论辩文本是一种非常普遍的话语类型，其过程是劝说他人接受某种态度或观点。“论证”一词指由两种主要类型的话语单元（句子）组成的句子子集：“前提”和“主张”（在某些情况下称为“结论”）。这些句子通过由理由构成的合理链条彼此关联，用于支持或攻击某一主张，或推导出进一步的主张 [2]。论证的最终主张通常称为其结论，即某人提出并视为真实的、可以为真或为假的命题，见补充图 A.2。

论证结构被表示为可废止的前提—结论结构以及一组批判性问题，它们有助于基于正反论证的权重，测试论证的强度和可接受性的过程。在文本中自动识别论证及其相关组成部分（主张和前提）被称为论证挖掘或论证提取。这是自然语言处理（NLP）中一个快速

*通讯作者：hanadi.mardah@manchester.ac.uk。博士生兼乌姆·古拉大学讲师，电子邮箱*homerdah@uqu.edu.sa*。网站：https://uqu.edu.sa/en/Profile/homerdah 或 https://hanadi-mardah.sitelio.me/

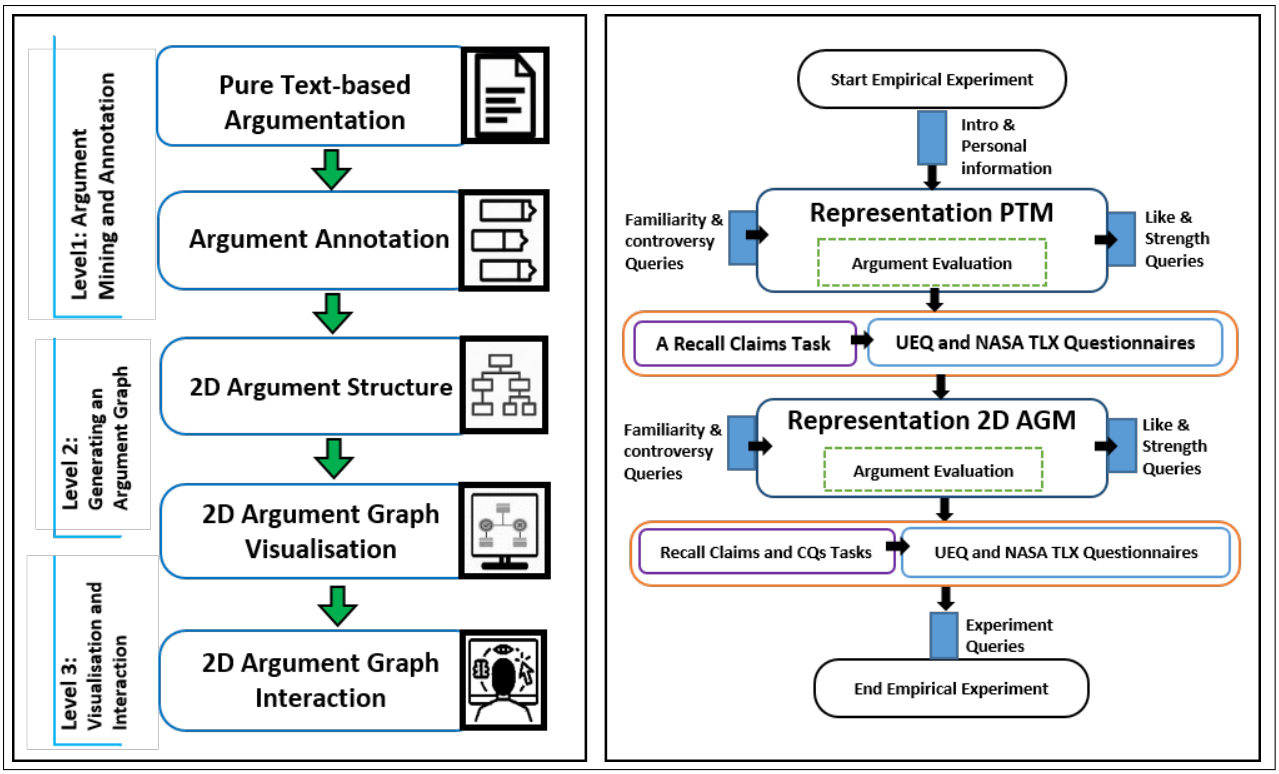


图 1：（左）基于 DAGGER 框架的研究流程 [1] 以及（右）抽象的实证实验工作流程，更多详情见图 A.11

发展的子领域，并得到多个标注语料库 [3, 4] 的支持。在这些分类器的支持下，可以解析论据以构建支持性的图表示，即论证图谱——逻辑论证结构的可视化表示，见补充示例 A.1。论证图谱或论证图可将复杂论证分解为相互独立的论证单元，从而可能促进对论据的解读与分析 [5]。

本研究探讨论证图（AG）的可视化表示对论证解读的影响。具体而言，研究重点关注可视化的影响，以及在较小程度上交互策略对论证图结构的影响。此外，我们旨在通过提出一种论证可视化模型，降低阻碍用户以批判性方式理解、分析和辩论各种现有论证的障碍；该模型结合了论证挖掘领域近期形成的能力加以开发。研究流程框架（图 1（左））基于 DAGGER 框架[2] [1]，包括三个层级：（i）使用论证挖掘进行人工标注；（ii）生成论证图；以及（iii）对论证图进行可视化和交互。研究对以下内容进行了批判性评价：（1）与文本形式相比，结构化论证图表示是否有助于论证解读；以及（2）所提出的论证图对论证解读任务的影响。因此，本研究包含一项实证分析，将纯文本模型（**PTM**）作为基线模型，与所提出的模型——二维论证图模型（**2D AGM**）进行比较。参见图 1（右），更多详情见图 A.11。

从实用角度来看，本 研究旨在回答以下研究问题（RQs），这些问题分为两个领域 ：

1. **论证表示：**
   - RQ1：话语层图表示如何支持对复杂论证和辩论的解释?
   - RQ2：哪些类别和关系会对解释过程产生积极影响?
   - RQ3：论证图式和论证类型如何支持这种表示的构建?
2. **可视化与交互：**

[2]DAGGER 是一种用于生成 从 Datalog+/- 中获取论证框架 一致性知识库的工具。

- RQ4：与纯文本解读相比，所提出的模型在内在和外在评估指标方面表现如何?
- 研究问题5：用户对所提解决方案的偏好，以及其易用性、有效性和实用性有何感知?
- RQ6：主题的熟悉度、强度和争议性会如何影响论证的批判性解读过程?

总而言之，本研究的贡献如下：

- 我们提出了一种采用节点链接图的二维论证图模型。这种方法与现有的可视化和空间配置工具有所不同，并对其起到了补充作用 [6, 7, 8, 9, 10, 11]。
- 作为方法论的一部分，我们遵循 Walton 分类方案来表示论证图，其中包含三个主要节点（层级和颜色各不相同）以及四种核心关系（颜色各不相同）。随后，我们使用可视化量表工具，让用户根据核心主张的正方论据（+）和反方论据（-）的权重，确定总体论证主题。这一点优于其他工具 [10, 6] ，因为最终决策映射这一明确功能能够支持对论证组成部分进行内在评估。我们的假设是，二维论证图模型在论证的批判性一致性评估（决策）方面表现更佳。
- 我们比较并对照这两种模型在论证批判性解读过程中的主要特征，例如所需时间和工作记忆。实验设置了涉及回忆主张和选择批判性问题（CQs）的任务。我们使用 A-BLEU分数评估每种模型中回忆出的主张与参考文本的相似程度。我们的假设是，与 PTM 相比，二维论证图模型在时间和工作记忆方面会取得更好的结果。

为实现本研究的目标，我们必须纳入以下内容：

- 我们对25名用户进行了实证分析，这些用户与两种模型表示进行交互：PTM和二维论证图模型。每种模型均包含一致性评估、任务以及两份调查问卷（NASA-TLX和UEQ）。实验包括一项调查问卷，用于了解用户对主题的熟悉度、争议性和强度，以及用户是否喜欢该模型。此外，还包括关于模型和实验的询问。
- 我们通过两份调查问卷以及若干关于论证主题的熟悉度、争议性和强度的询问，从工作负荷和用户感知方面进行了外在评估。这些询问用于确定这些特征对两种模型中解释过程的影响。这有助于支持理性评估，并为用户提供更直观的解释。我们的假设是，与PTM相比，二维论证图模型将实现更优的工作负荷表现和用户感知，因为使用论证图的根本原因在于分解复杂论证并促进解释。与PTM相比，论证图对论证解释的影响将更大，而这种影响取决于主题的熟悉度、争议性及其强度。

2 背景与文献综述

论证旨在通过理性且基于证据的信念来证明结论与行动的合理性，阐明分歧、证明真理并理解多种视角 [12]。尽管自然语言处理（NLP）和论证挖掘（AM）领域在自动构建论证结构及可视化方面取得了进展，但我们对这种新增结构与可视化方法相结合，究竟能在多大程度上支持人们解读复杂论证，仍缺乏充分的理解与量化研究。因此，本研究介绍了论证图的可视化与交互（VIAGs），这是计算机支持的论证可视化（CSAV）的一个子主题 [13]。本节回顾并总结了文献中有关论据表示以及论证映射／可视化模型的方法（图 1（左））。

2.1 论证结构与 Walton 的分类方案

论据的分类和结构化面临一项挑战 [14, 2]）。目前尚无“结构化论证”的普遍定义，这使得如何找到一种分类系统来识别不同的论证模式，并针对各种情境和目的构建论证，成为一项根本性挑战[15]。人们曾多次尝试对图式结构进行分类 [16, 17, 18, 19, 20],，但最常用的方案由 Walton 提出 [21]。根据 Lawrence 等人的一项调查 [22], ，可以依据与 Walton 方案相对应的批判性问题来评估论据。Can 等人[23] 也发现，沃尔顿模型有助于支持论证和批判性评价。本研究将采用 Walton 分类方案 [21, 2, 24] 来表示论证文本。

2.2 论证挖掘与标注方法

论证挖掘（AM）是一种文本挖掘技术 [25]，包含两个主要阶段：识别论证话语并预测论证关系 [14, 26]。论证挖掘可以完全自动化完成（[27, 28, 29, 30, 31, 20, 32, 33,34], TARGER[35], MARGOT ARGs[36], 或 ConToVi [37]），也可以结合人工与自动化方法（[38,39]）。然而，该领域缺乏标准化 [40]；由于未能充分纳入语义和领域知识，完全自动化的方法准确性不足。因此，专家目前仍依赖耗时的人工标注 [41]。本研究旨在完善 Walton 分类体系 [2]，使用 BRAT 标注工具 [43] 和 VIANA 框架，分析来自“ProCon.org” [42] 的辩论主题，重点提升论证挖掘中的文本解读能力。

2.3 论证映射或可视化

论证结构可视化也称为论证表征、映射或图，是一种文本可视化形式 [44]。论证图可以呈现为节点链接图或树形图，其中节点链接图适用于表示层级结构[45, 21]。根据 [12], , 框／节点通常包含完整、符合语法的陈述句：理由／前提（支持某些主张的证据）；主张（被认为真实的观点）；结论（由理由支持的最终主张）；或反驳意见（反对结论的证据）。节点之间的关系通过线条／箭头表示推理关系。因此，本研究将沃尔顿论证结构构建并表示为一棵树（节点链接图），包含三个主要节点：前提、主张和结论，以及四种核心关系：攻击、支持、反驳和削弱[2], ，并将其可视化为二维基于图的论证模型，使用 Unity 框架进行渲染 [3]。

2.4 论证分析方法

若干论证挖掘系统包含可用于分类、比较和衡量论证结构的分析方法，其中包括面向论证用户的**非数值分析**。例如：Neva [47], 、 Carneades [48], 以及由 [49] 提供的主张数据集；面向普通用户的 Kialo [50], 和 PEOPLES [52]; 以及旨在帮助学生和教师的 AGORA-net [53], 和面对面课堂 [54]。另外，还有基于论证图式的系统 [55], Carneades [56] 和 Parmenides [57]。

论证结构数值分析，根据正方论据和反方论据的权重比较论证结构，例如 [6] 和 designVUE [10]。然而，这些评估过程复杂，且缺乏对论证组成部分的评估，也没有提供支持内在或外在分析的评估。相比之下，本研究评估论证组成部分，并采用基于正负权重的简单数值过程（明确提供映射决策功能）。

2.5 评估　论证结构表示法

该研究的核心部分评估了两种论证表示方式：基于文本的论证，以及基于二维图、以沃尔顿论证结构为基础的论证。此前已有多项研究以不同方式开展了类似的评估过程： [8] 中的研究结果加深了对可视化如何支持逻辑推理的理解。或者，[9]对两种可视化表示——PCM和HTNM——进行了定性与定量评估。或者， [11] 开展了一项案例研究，其中包含基于两种认知思维模式实施的四个系统。此外，[13] 评估了以五种不同可视化形式呈现推理的效果。或者，AVISE [7] 支持对现实世界论证的评估。这些研究均未基于沃尔顿论证结构开展评估，也不支持内在评估或外在评估，亦未提供用于映射决策的明确功能。相比之下，本研究旨在展示，与基于文本的论证相比，基于二维图的论证在

[3]Unity [46] 是一款由 Unity Technologies 开发的跨平台游戏引擎，于 2005年6月 在苹果全球开发者大会上首次公布并发布，最初是一款适用于 Mac OS X 的游戏引擎。此后，该引擎逐渐扩展，支持多种桌面、移动、主机和虚拟现实平台。Unity 具有许多功能，研究人员可以轻松尝试这些功能，从而让可视化更加灵活。Unity 具备以下功能：1- Unity 是一个免费、灵活且易于使用的平台，可用于简单的二维可视化，也可用于非常复杂的可视化。2- 在 Unity 中，你可以浏览数量庞大的类型、子类型和风格。3- Unity 的真实感表现能力非常强大，因此许多开发者也将其用于游戏开发以外的任务。4- Unity 支持两种常见的编程语言：C♯ 和 JavaScript。任何有 C♯ 背景的人都可以快速上手 Unity 并开始编写脚本。5- 在 Unity 中编程的一大特点是，你可以直观地使用用户界面为脚本添加功能，而不必与特定的变量列表交互。6- Unity 一直优先考虑支持任意平台的开发，平台选择（iOS、Android、个人电脑或主机）也在不断增加。Unity 支持 Oculus Rift、HTC Vive、Microsoft Hololens 等设备。6- 获取资源十分容易。Unity 没有内置的资源，都可以在资源商店中找到。7- 为满足未来需求，Unity 服务提供了一组新功能，让你更轻松地构建、分享和销售项目。Unity Cloud Build 和 Unity Collaborate 是用于备份整个项目并构建多个版本的工具，而且不会拖慢系统运行。

论证解读、一致性评估、绩效、工作记忆、用户感知和用户体验方面的显著差异及影响。该研究使用可视化量表工具评估论证（作为内在评估），并使用NASA任务负荷指数和UEQ两份问卷评估论证模型（作为外在评估）。

3 方法

3.1 二维论证图（AGs）模型的开发——二维论证图模型。

目标模型为二维论证图模型，该模型需要以下三个维度。

论证挖掘、语料库选择与标注的回顾。每篇论证文本均被视为一场包含正反双方的辩论，双方各自具有相应的主张和前提。文本的组织方式是先呈现主要结论，随后每个正方陈述之后紧接其反方陈述；所有论证主题均遵循这一结构。从辩论主题中随机选择并提取了非结构化论证文本[4]。所选的论证挖掘模式所依据的结构基于Walton分类方案 [2], ，该方案经Milz[24]修订和简化。我们在Brat工具中进行了配置，定义了实体（主张、前提和结论）、关系（支持、攻击、削弱和反驳关系）以及属性（类比、知情立场、专家意见、公众意见、因果、证伪、正面后果和负面后果），这些均基于论证图式。所使用的论证图式包括：专家意见论证、知情立场论证、公众意见论证、类比论证、基于正面／负面后果的论证、因果论证以及基于证伪的论证。事实上，每个主题都有两个结论，一个为正方结论，另一个为反方结论。因此，我们对正反双方的所有主张及其前提进行了标注，并将其链接至双方的结论。每个前提的节点均被标注，根据Walton 分类方案与其自身属性（图式类型）相连接，并链接至其主张。每种图式的标准均由Walton [21]描述。这形成了一个标注语料库，随后对其进行解析以构建论证图（见图1（左））。

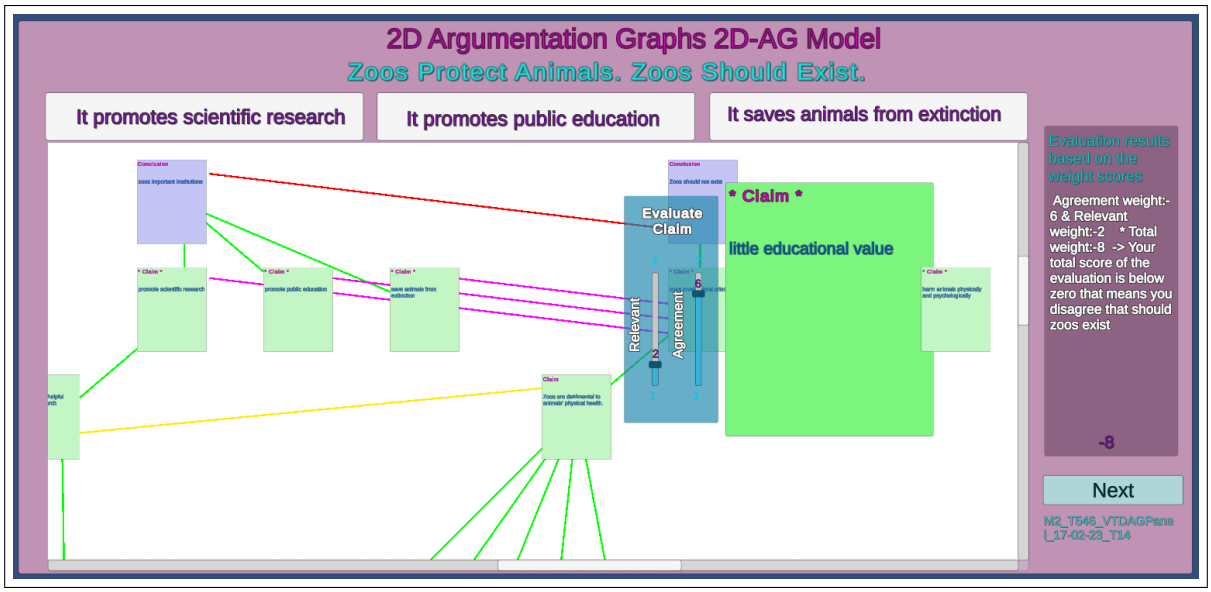
构建论证图结构。 这些论证结构被映射到树形结构 [21] 中，以表示论据 [45, 12]。具体而言，树形结构采用一种非空间填充式变体，可体现层级关系，并通过节点链接图表示。节点链接图受两个因素影响：分支数和深度，这使得渲染深度不过深的树形结构更加容易。本研究采用节点链接图，利用论证图结构中的部分类别和关系构建二维论证图模型。因此，标注语料库中的每一行都提供一个节点、链接或与类型相关的属性。彩色矩形形状表示一个节点，可能是结论节点（紫色）、主张节点（绿色）或前提节点（灰色）。每个前提节点都与一个（属性）论证图式相关联（例如，专家意见论证、类比论证等）[2]。明亮颜色的节点表示参与者尚未点击这些节点，而深色表示参与者已点击该节点。每条彩色链接表示节点之间的关系，这些关系分别为支持（绿色）、削弱（粉色）、反驳关系（黄色）或攻击（红色）。在论证图中，空间和颜色配置会随支持类别而变化（见图2）。

在二维论证图模型中渲染论证图结构。在 Unity 中，我们使用一种常见的渲染工具——矩形面板，在屏幕上呈现论证图。面板左侧是论证图面板，其中包含三个或两个主要主张（标签页）。每个标签页都包含两个论证图子树：支持论证图子树及其反驳反方论证图子树。论证图子结构渲染在名为“滚动视图”的资源中，用户可以通过上下和／或左右滚动来浏览图。对于每个**主张节点**（绿色节点），我们使用两个李克特量表滑块（分别用于评估相关性和一致性），衡量某个主张与结论的相关程度，以及用户对该主张的认同程度。评分范围为 1 至 7，其中 1 = 表示不同意／不相关， 7 = 表示非常同意／相关。每个**前提节点** （灰色节点）都关联一种图式类型及相关的批判性问题，用于检验论证的强度和可接受性。矩形面板右侧显示整个论证的评估信息，该信息基于所有主张的汇总得分（见图 A.6）。用户每次为主张评分时，评估信息都会自动更新。因此，将所有主张的得分进行算术求和，即可得出该论证的最终判定。此外，还设有一个隐藏计时器，用于计算用户在模型上花费的总耗时。

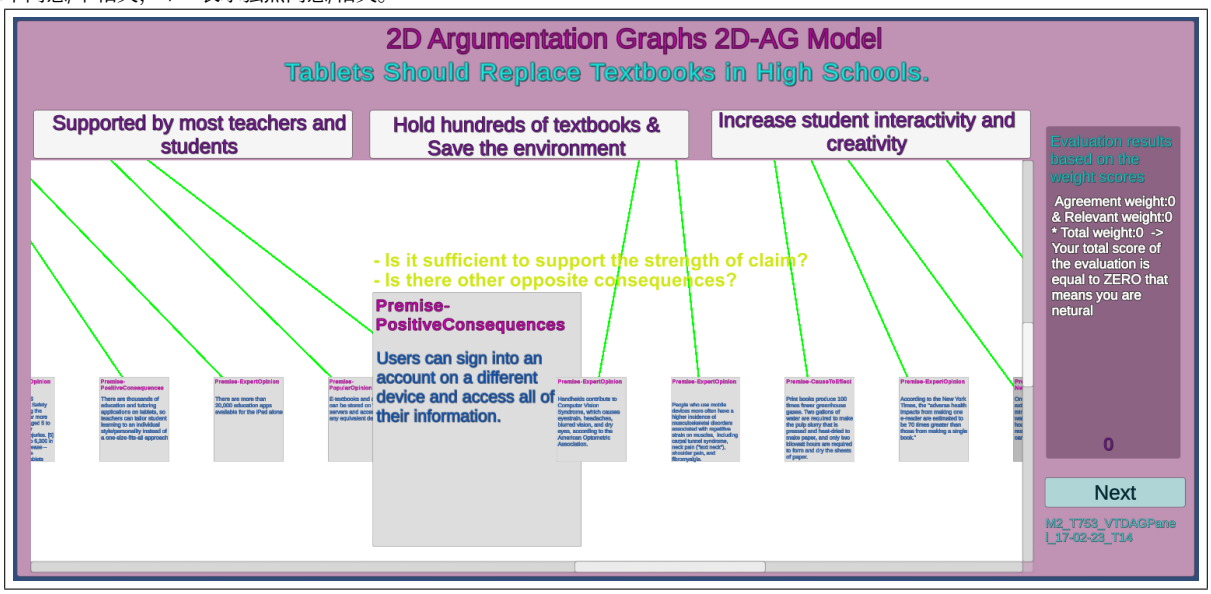
3.2 纯文本模型（PTM）的开发。

基线模型——纯文本论证（PTM）——通过对文本进行简单渲染来呈现（使用同一组辩论主题），不包含额外结构。

[4]



（a）二维论证图模型中的一项主张（绿色节点），关联着两个李克特量表滑块，用于评估相关性和一致性（1 至 7）；其中，1 =表示不同意/不相关，7 = 表示强烈同意/相关。



（b）二维论证图模型中的一个前提（灰色节点），关联着适当的图式类型和相关的批判性问题。例如，论证图式类型包括专家意见论证、类比论证等。[2]

图 2：二维论证图模型（2D AGM）包含左侧的论证图面板，其中设有三个或两个主要主张（标签页）。每个标签页包含两个论证图子树：支持论证图子树及其反对论证图子树。彩色矩形表示一个节点，可以是结论（紫色）、主张（绿色）或前提（灰色）节点。每条彩色连线表示节点之间的关系，分别表示支持（绿色）、削弱（粉色）、反驳关系（黄色）或攻击（红色）。右侧显示相关一致性得分总体权重的评估信息。

在 Unity 中，我们使用一个矩形面板在屏幕上呈现论证文本。矩形面板内的另一个名为“Scroll View”的组件中包含一个名为“TextMeshPro”的组件，用于让用户滚动查看文本。文本以一页的长度呈现在深色背景上，以便于阅读。此外，左侧设有一个李克特量表滑块，用户可以使用它按照 -3 到 +3 的量表对论证进行评价，其中 -3 表示反对，+3表示一致性，0 表示中立。此外，还有一个隐藏计时器，用于计算用户的总耗时（见图 3）。

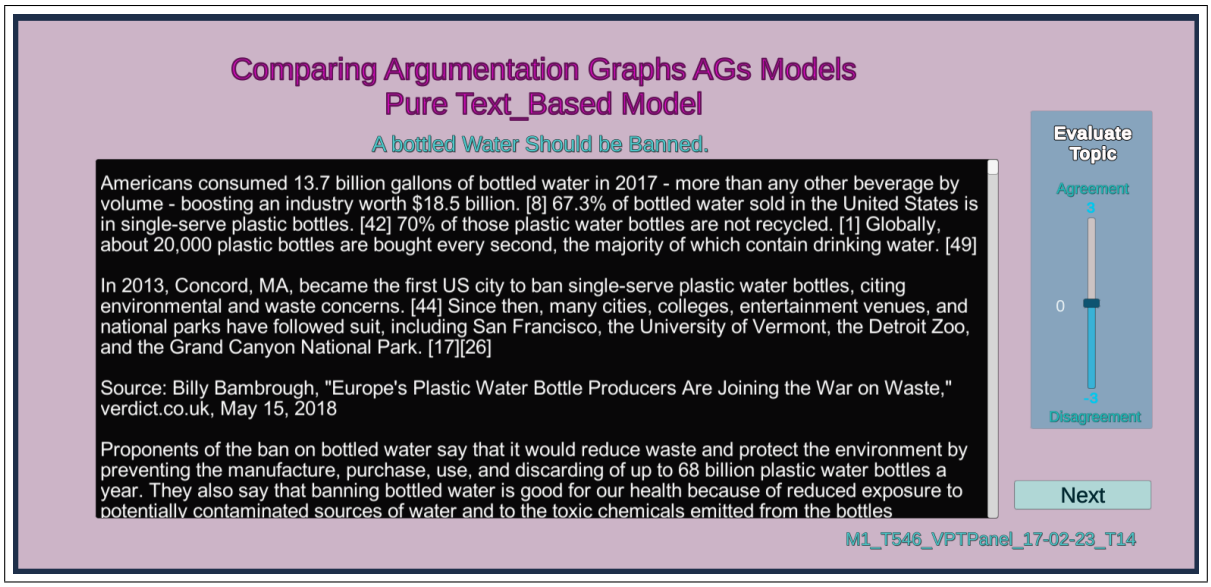


图3：纯文本模型（PTM，基线模型），左侧显示论证文本。右侧，用户通过移动李克特量表滑块来表示对某项论证的一致程度（-3 到 +3），其中 -3 表示反对， +3 表示赞同，0 表示中立。

4 实验设计

实验采用被试内设计。每位参与者依次体验两种模型：首先是 PTM，然后是二维论证图模型（2D AGM）；从 10 个论据中随机为每位参与者选取两个（每个论据来自不同主题，见补充材料 8）。

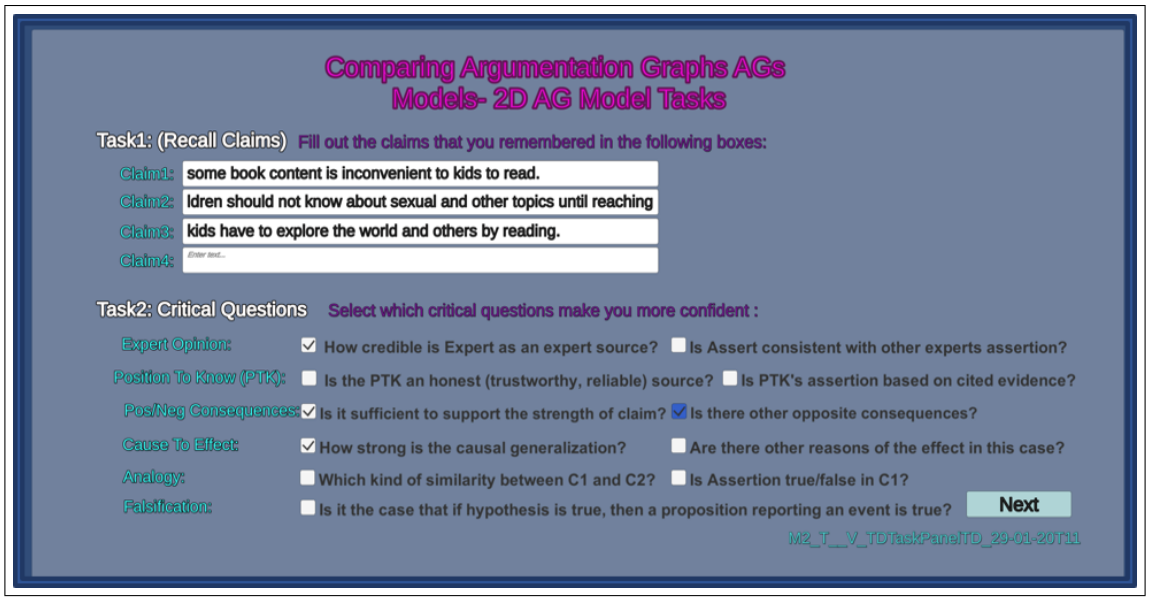
实验的第一部分评估 PTM。首先，询问用户对主题争议性的感知、对辩论的熟悉度以及最初的立场。接着，用户在文本模型的滚动视图中阅读一篇论证文章，并使用李克特量表控件表示自己对论证的同意程度（任务1，见图3）。之后，询问用户对主题强度的看法，以及是否喜欢该主题和模型。然后，用户在四个文本框中填写论点陈述，以写下记住的主张（下文称为任务1：回忆主张，见图4b）。最后，用户填写两份简短的封闭式结构化问卷：NASA-TLX 8.1 和 UEQ 8.2，以评估工作负荷、使用体验和意见（见图 A.9 和 A.10）。

第二部分中，用户与二维论证图模型进行交互。与实验的第一部分类似，在显示二维论证图之前，系统会询问用户对该主题的熟悉度、其争议程度以及用户最初的立场。二维论证图以二维彩色树状图的形式呈现在滚动视图中（图2）。用户使用附在每个主张上的李克特量表控件，标明自己对论证的一致性以及每个主要论点的相关性。一致性的最终得分由所有分数聚合得出（见第3.1节）。
接下来，系统要求用户完成两项任务：任务1：回忆主张；任务2：选择批判性问题。在任务1中，与PTM相同，用户需要在四个包含论点陈述的文本框中写下记住的主张。在任务2中，系统会向用户展示一组带复选框的批判性问题（摘自Walton的著作），并要求用户勾选在二维论证图模型的论证中用到的问题（见图4a）。最后，与PTM一样，用户需要完成两份问卷：NASA-TLX和UEQ。有关实验环境的详情，请参阅补充方法8、NASA-TLX 8.1和UEQ 8.2。为确保可复现性，完整的实验工作流程见补充图A.11。

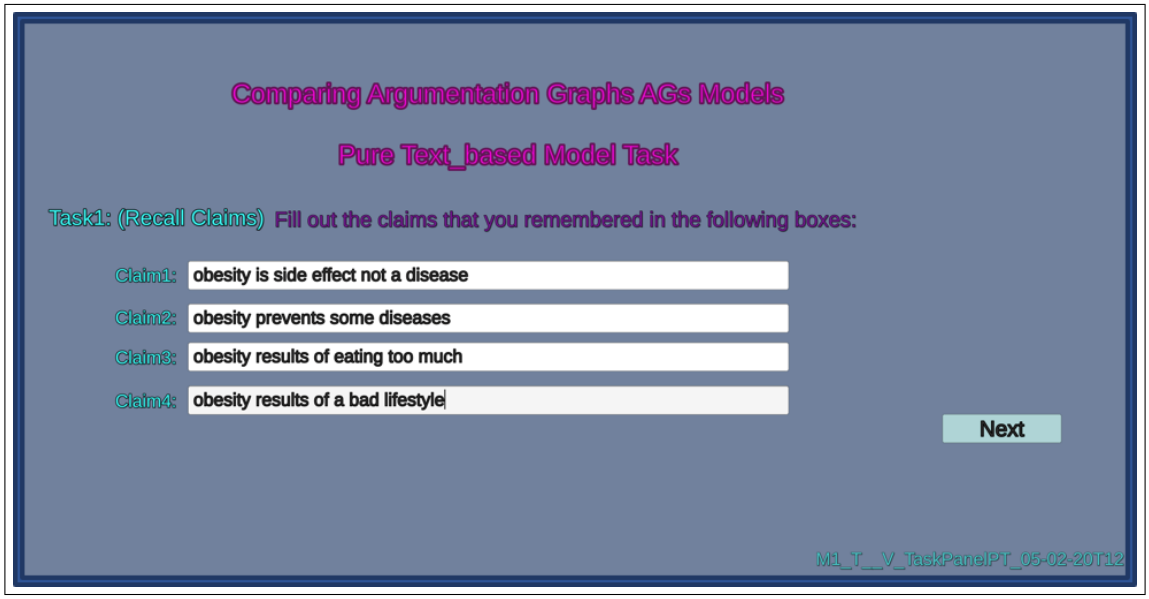
4.1 参与者、指标与设置

数据：PTM 和二维AGM模型中呈现了代表不同论据的 10 个主题（见补充材料第 8 节）。

参与者：我们从 [匿名] 工程学院招募了25名研究生，并向参与者发放亚马逊礼品卡作为报酬（每位参与者10英镑）。共有15名女性和10名男性参与，其中23



(a) 二维论证图模型任务 - 任务1：回忆主张包含四个用于输入回忆主张的输入框。任务2：为二维论证图模型选择批判性问题



(b) PTM任务 - 任务1：PTM模型的回忆主张包含四个用于输入回忆主张的输入框。

图4：模型任务——任务1：回忆主张，针对(a)二维论证图和(b) PTM模型，包含四个用于输入回忆主张的字段。任务2：为二维论证图模型选择批判性问题。

参与者中有名博士生和2名硕士生。参与者的年龄在20至40岁之间。所有参与者的视力均正常或矫正至正常，且无色觉缺陷。

指标：实验旨在评估以下方面：

1. 参与者在辩论中的初始立场（在查看 PTM 或二维论证图模型之前）。2. 对论证的同意程度（在阅读 PTM 或二维论证图模型时），参见第 3.2 节和第 3.1 节。3. 完成任务所需的时间（在模型上花费的总时间，以分和秒计）。4. 任务1：工作记忆（短期记忆）（A-BLEU 以 0 到 1 之间的小数计算，参见补充材料第 8.1 节）。

5. 任务2：选定的批判性问题。（选定的批判性问题总数，以整数表示（0-11）。参见补充材料第 8、8.2 节和第 4 节）6. 参与者对主题的熟悉度（关于该主题的先验知识）、争议性（公众之间复杂且涉及广泛的辩论）、强度（支持主张和结论的理由——前提陈述）以及偏好（对该主题的喜好度）量表。计算李克特量表控件评分，范围为 1-7，其中 1 表示非常负面，7 表示非常正面。7. 工作负荷（NASA-TLX），包括时间需求[5], 身体需求[6], 心理需求[7], 绩效[8],挫折感[9], 和努力程度[10]。每项的分类结果为整数（0-500），总工作负荷为小数（0-100）。8. 用户体验（UEQ），包括吸引力[11], 清晰度[12], 效率[13], 可靠性[14], 和刺激性[15]。每项的评分结果为小数（-3 至 3）。9. 用户体验（UEQ），包括基准[16]。预期结果为小数（-3 至 3）。

设置：实验于白天在安静的房间内进行。项目以 Unity 游戏引擎为平台运行。参与者通过 Microsoft Teams 使用鼠标和键盘完成在线任务。结果和数据通过一个名为“Uni-Excel”的程序化嵌入式插件包收集，该插件包具有显著优势并提供了帮助，详见补充材料 6.7。

5 结果

5.1 与两个模型交互后的意见和印象。更多参与者更青睐二维论证图模型

如图5所示，尽管部分参与者更青睐 PTM（N = 9，36%），但更多参与者更青睐二维论证图模型（N = 21，84%）——该评估基于李克特量表值（>3）。与 PTM 相比，参与者对二维论证图模型的偏好评分显著更高（$p = 0.005$，威尔科克森符号秩检验）。这一结果表明，就目标指标和用户感知而言，二维论证图模型的影响力更大；这两方面均属于 RQ4 和研究问题5 的考量范围。同样，这也进一步印证了 [58] 针对 HTNM 模型得出的类似发现。

5.2 变更后的决策与一致性：PTM 与二维论证图模型之间无显著差异

我们评估了参与者在看到模型后是否改变了立场（从赞同（5–7）转为反对（1–3）或中立（4），反之亦然）。在 PTM 中，有 10 名（40%）参与者改变了立场，而在二维论证图模型中，这一人数为 13 名（52%）。立场变化按正方与反方两个方向定义：PTM 中有 6 次正向变化，二维论证图模型中有 8 次；PTM 中有 4 次负向变化，二维论证图模型中有 5 次（见图 6）。然而，比较两种模型的变化值时，这一差异并不显著（$p=0.29$，见表 1）。尽管这些差异仅达到边缘显著性，但仍表明，二维论证图模型支持论证解读与构建（RQ1 和 RQ3），对解释过程有积极效应（RQ2），并且在论证一致性评估方面表现更好：更多参与者改变了立场（内在指标），相较于 PTM

[5]时间需求：由于任务各要素出现的频率或节奏，您感受到了多大的时间压力？节奏是缓慢悠闲，还是快速紧张？ [6] 身体需求：需要进行多少身体活动（例如推、拉、转动等）？任务是轻松还是 demanding？缓慢还是 brisk？松弛还是 strenuous？ [7]心理需求：需要进行多少心理感知活动（例如思考、搜索、记忆等）？任务是轻松还是 demanding？简单还是复杂？ [8]绩效：您认为自己在完成实验者（或您自己）设定的任务目标方面取得了多大成功？对于完成这些目标时的表现，您满意程度如何？ [9]挫折感：在任务过程中，与感到安全、满足、放松或自满相比，您有多大程度上感到不安、气馁、恼怒、紧张或烦躁？ [10]努力程度：为了达到您的绩效水平，您需要付出多大的努力（心理和身体上的）？ [11] 吸引力是指：对模型的总体印象。用户喜欢还是不喜欢它？ [12]清晰度是指：用户是否容易熟悉该模型并学会如何使用它？ [13]效率是指：用户能否无需不必要的努力即可完成任务？它的响应速度快吗？ [14]可靠性是指：用户是否感觉能够控制交互？它是否安全且可预测？ [15]刺激性是指：使用该模型是否令人兴奋并富有激励性？使用起来有趣吗？ [16]基准意味着该量表是参照基准数据集中的现有数值设定的。该数据集包含来自 468 项研究、涉及不同产品（软件、网页、网上商店等）的 21,175 人的数据

表 1：PTM 与二维论证图模型之间的比较。Sig——显著性：p<0.001 时为***，p<0.01 时为**，p<0.05 时为*，p>0.05 时为 ns（不显著）；p 值来自威尔科克森符号秩检验，以及用于 UEQ 和 NASA任务负荷指数的 t检验。

		PTM	二维论证图模型	p	Sig
解读一项论证	中位数 [第一四分位数, Q3]	11:13 [5:57, 17:12] min	13:05 [9:33, 23:13] min	0.015	*
任务1 ‘回忆主张 ’	A-BLEU分数 中位数 [第一四分位数,Q3]	0.07 [0.02、0.14]	0.10 [0.04, 0.2]	0.247	ns
参与者的意见 参与者 (李克特量表 [1-7])	有争议的话题，n	17	20	0.148	ns
	中位数 [第一四分位数,Q3]	5 [3, 6]	6 [4, 7]		
	熟悉的话题，n	16	12	0.33	ns
	中位数[第一四分位数,Q3]	5 [3, 6]	4 [2.5, 5]		
	强烈认同的话题，n	22	23	0.84	ns
	中位数 [第一四分位数,Q3]	6 [5, 7]	6 [5, 7]		
	喜欢的话题, n	17	19	0.3	ns
	中位数 [第一四分位数,Q3]	6 [4, 6.5]	6 [4.5, 7]		
	喜爱模型, n	9	21	0.005	**
	中位数 [第一四分位数,Q3]	3 [2, 5]	6 [4, 7]		
	初始位置变化，n	10	13	-	-
	变化的绝对幅度 中位数 [第一四分位数,Q3]	3 [2.75, 4]	2 [1.5, 3]	0.29	ns
UEQ （评分 [-3,3]）	吸引力，中位数 [第一四分位数，Q3]	-0.5 [-1.25，1.5]	1 [0.63，1.75]	0.0026	**
	清晰度，中位数 [第一四分位数，Q3]	1 [0, 1.67]	1.33 [0.83，1.67]	0.059	ns
	效率，中位数 [第一四分位数,Q3]	0.5 [-0.5, 2]	1.5 [0.75, 2.25]	0.036	*
	可靠性，中位数 [第一四分位数,Q3]	0 [-0.5, 1.5]	1.5 [0, 1.75]	0.113	ns
	刺激性，中位数 [第一四分位数，Q3]	0.5 [-0.25，0.88]	0.75 [0.5，1.13]	0.035	*
NASA任务负荷指数 （评分 [0-500]）	心理需求，中位数 [第一四分位数,Q3]	300 [210, 387.5]	120 [65, 205]	0.0002	***
	身体需求，中位数 [第一四分位数,Q3]	0 [0, 60]	0 [0, 95]	0.50	ns
	时间需求，中位数 [第一四分位数，Q3]	100 [50, 180]	100 [75, 210]	0.32	ns
	绩效，中位数 [第一四分位数，Q3]	100 [65, 232。5]	130 [27.5, 232.5]	0.91	ns
	努力程度，中位数 [第一四分位数，Q3]	200 [90, 300]	140 [72.5, 210]	0.07	ns
	挫折感，中位数 [第一四分位数，Q3]	120 [7.5, 197.5]	65 [37.5，167.5]	0.51	ns
NASA任务负荷指数 （评分 [0-100]）	总体工作负荷得分，中位数 [第一四分位数，Q3]	64.33 [53.67，69.5]	48.33 [39.167, 62.67]	0.01	*

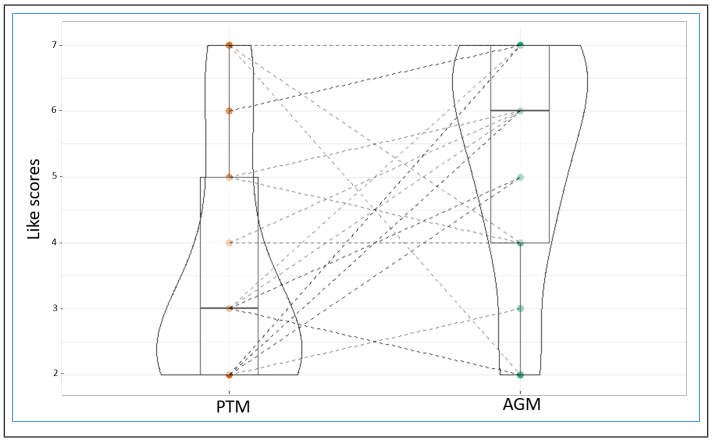
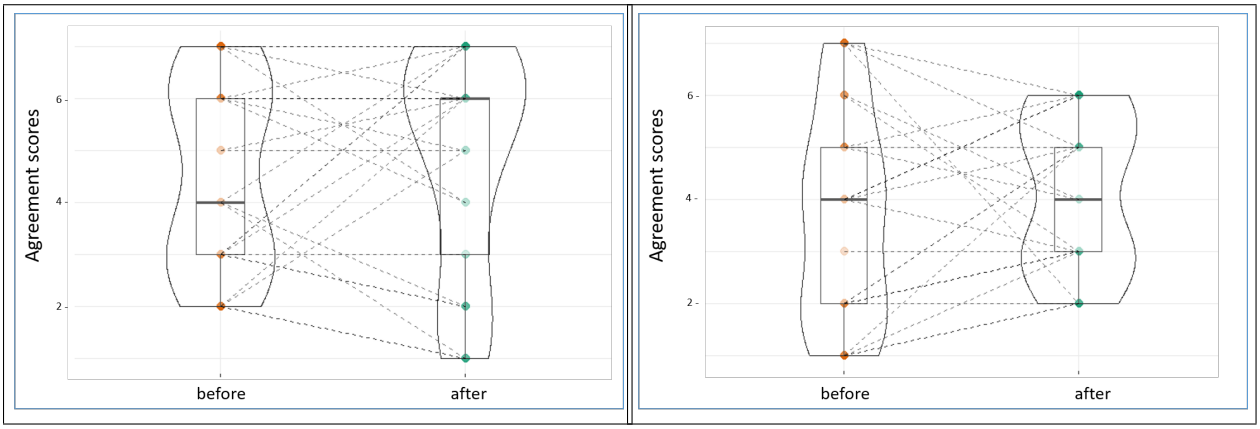


图5：两个模型的 Like模型得分配对图基于李克特量表值（>3）。差异采用威尔科克森符号秩检验的 p 值进行测量，Sig– 表示显著性：p < 0.001 时为 ***，p < 0.01 时为 **，p< 0.05 时为 *，p > 0.05 时为 ns（不显著）)

（RQ4）。尽管其他一些研究（例如[6] 和 designVUE [10], ）依据正方与反方论据的权重比较论证结构，但它们对论证结构进行了复杂且具有挑战性的数值评估

这一过程不包含任何支持论证一致性评估或论证模型评估。相比之下，二维论证图模型会评估论证组成部分，并基于正负权重采用简单的数值过程（为映射决策提供明确的功能）。



(a) PTM。 (b) 二维论证图模型。

图 6：与两种模型交互前后的 一致性得分：(a) PTM，(b) 二维论证图模型。虚线连接各参与者的得分。

5.3 任务1 回忆主张：衡量参与者的工作记忆。二维论证图模型与 PTM 对工作记忆的影响相似。

我们通过 A-BLUE 得分比较了这些模型对参与者工作记忆的影响（见表 1）。两种模型的 A-BLUE 得分均较低，且差异不显著（$p = 0.247$）；PTM 的中位数为 $= 0.07$，二维论证图模型的中位数为 $= 0.10$。不过，我们认为，通过增加交互、练习以及对设计配置的熟悉度，可以提高二维论证图模型的 A-BLUE 得分。因此，就积极影响工作记忆并支持解释过程而言，这些结果支持 RQ1、RQ2 和 RQ3。
类似研究也报告了这一结果，例如 [8]，发现，围绕协作论证可视化组织良好的教学实践，能够显著提升学生的分析推理技能；此外，[11] 发现，叙事与图形表征对参与者的知识建构绩效没有影响。更多信息请参见表 A.1，其中列出了两种模型中获得最高 A-BLEU 得分的候选句——PTM 中的一位参与者和二维论证图模型中的一位参与者。表A.2列出了 PTM模型的参考文本，表 A.3 列出了二维论证图模型的参考文本。图 A.7 展示了所有参与者的 A-BLEU分数。

5.4 二维论证图模型－任务2：选定的批判性问题数量与主题熟悉度的相关性 y

我们发现，选定的批判性问题数量与对主题的熟悉度呈负相关（使用斯皮尔曼相关测量，$p = 0.038$，$r = -0.42$，见图8a）。主题越熟悉，选定的批判性问题就越少，这与 RQ1、RQ2、RQ3 和 RQ6 相关。这表明，图式类型（包括批判性问题）有助于参与者更批判地解读，尤其是在面对熟悉的主题时。然而，主题强度与批判性问题数量之间（$p = 0.99$，图8c）、立场偏好（喜好度）与批判性问题数量之间（$p = 0.93$，图8d），以及主题争议性与批判性问题数量之间（$p = 0.87$，图8b）均不存在其他显著相关性。

5.5 用户体验问卷 UEQ（衡量用户感知）：在吸引力、清晰度和效率方面，二维论证图模型显著优于 PTM，在可靠性和刺激性方面也表现更好。

通过对 UEQ 答案进行统计比较，可以较为轻松地比较两个模型 [59]。本研究从用于评估交互式软件模型质量的 26 个条目中，仅选取了 15 个相关条目；更多详情请参见图 A.10 和 8.2。我们发现，在吸引力（$p = 0.0026$）、效率（$p = 0.036$）和刺激性（$p = 0.035$）方面，二维论证图模型显著优于 PTM。然而，我们发现两个模型在清晰度（$p=0.059$）或可靠性（$p=0.113$）方面没有显著差异，见表 A.5。这一点与 RQ4 和研究问题5有关。

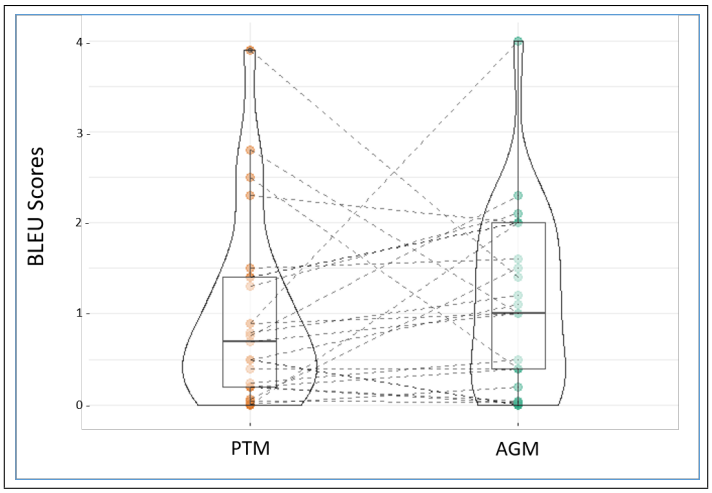
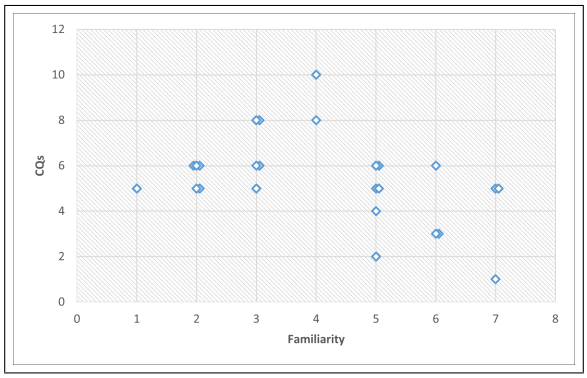
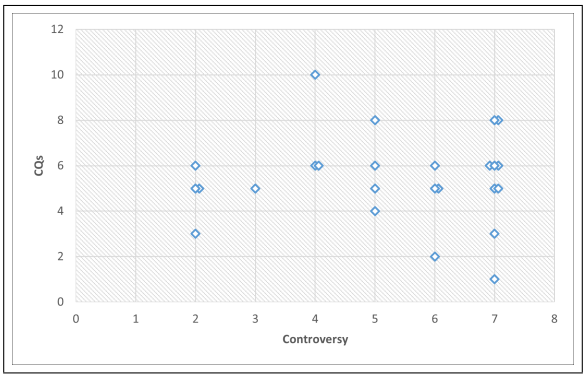


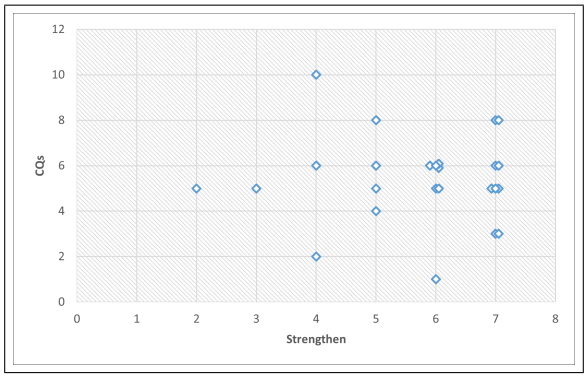
图7：两种模型在任务1“回忆论点”中的A-BLEU分数配对图。虚线连接各参与者的分数。PTM和二维论证图模型中，参与者在任务1取得的分数之间没有显著差异（p=0.247）。



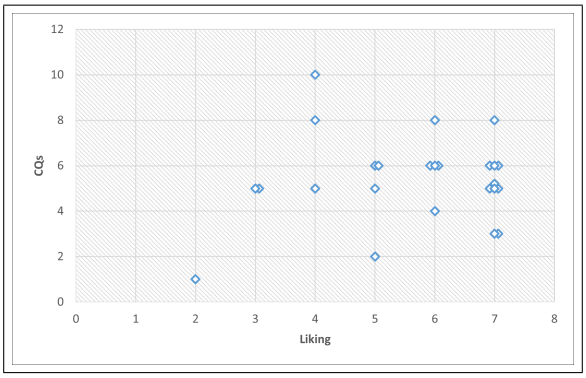
（a）主题熟悉度与批判性问题数量



（b）争议性与批判性问题数量

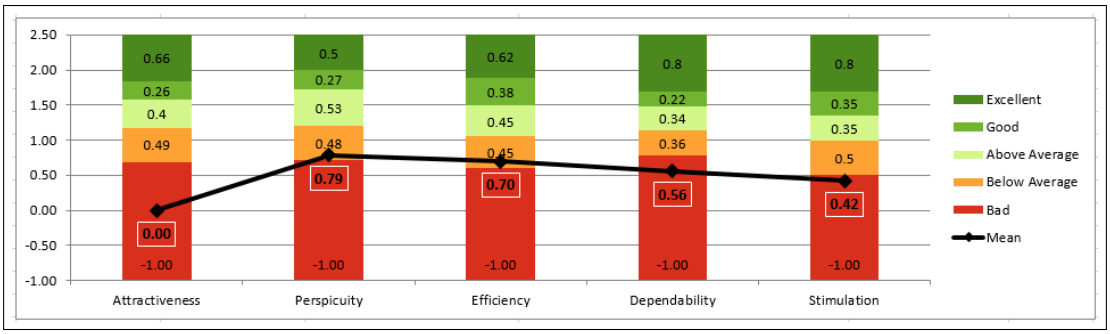


（c）主题强度与批判性问题数量

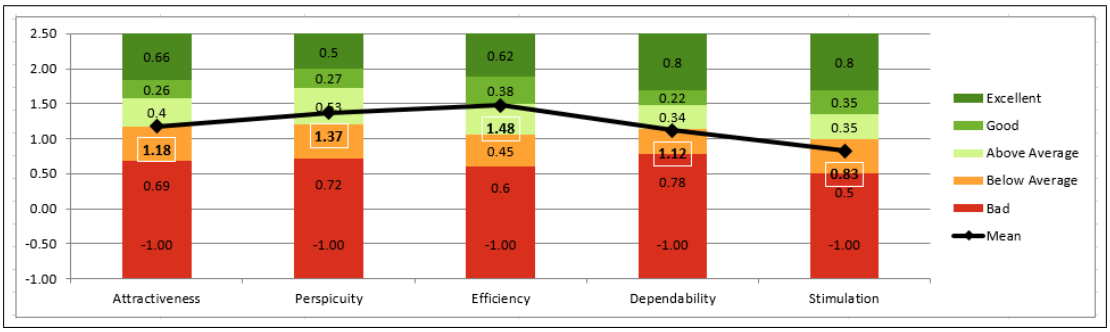


(d) 主题喜好度与批判性问题数量

图8：主题各项量表的线性回归：(a) 熟悉度、(b) 争议性、(c) 强度和 (d) 喜好度与选定的批判性问题数量的关系。每个点代表一位参与者的一条回答。熟悉度与批判性问题数量呈负相关（$p = 0.038$，$r = -0.42$），但其他变量之间均无相关性。



（a）PTM模型的基准



（b）二维AGM模型的基准

图9：两个模型的UEQ基准。差异通过UEQ 分析工具中的（t检验）均值进行测量。二维论证图模型在吸引力（1.18）、可靠性（1.12）和刺激性（0.83）方面低于平均水平；在清晰度（1.37）和效率（1.48）方面高于平均水平。颜色表示均值的基准排名，黑点表示各模型在每个量表上达到的均值。

5.6 用户体验问卷UEQ（与UEQ基准数据集比较）与UEQ基准相比，二维论证图模型处于平均水平，而PTM处于低于平均水平。

图 9 展示了两个受评估模型与基准的结果对比。这有助于得出关于它们相对质量的结论，并将其与其他类似模型进行比较。就基准的均值和中位数（范围为 -3 至 3）而言，UEQ 分析工具（使用 t检验）显示，PTM 在清晰度（0.79，1）和效率（0.70，0.5）方面低于平均水平，在吸引力（0.00，-0.5）、可靠性（0.56，0.0）和刺激性（0.42，0.5）方面表现较差。另一方面，二维论证图模型在吸引力（1.18，1）、可靠性（1.12，1.5）和刺激性（0.83，0.75）方面低于平均水平，但在清晰度（1.37，1.33）和效率（1.48，1.5）方面高于平均水平，见表 A.5 和表 1。因此，总体而言，与UEQ基准数据集相比，二维论证图模型提供了平均水平的用户体验，这与 RQ4 和研究问题5相关。

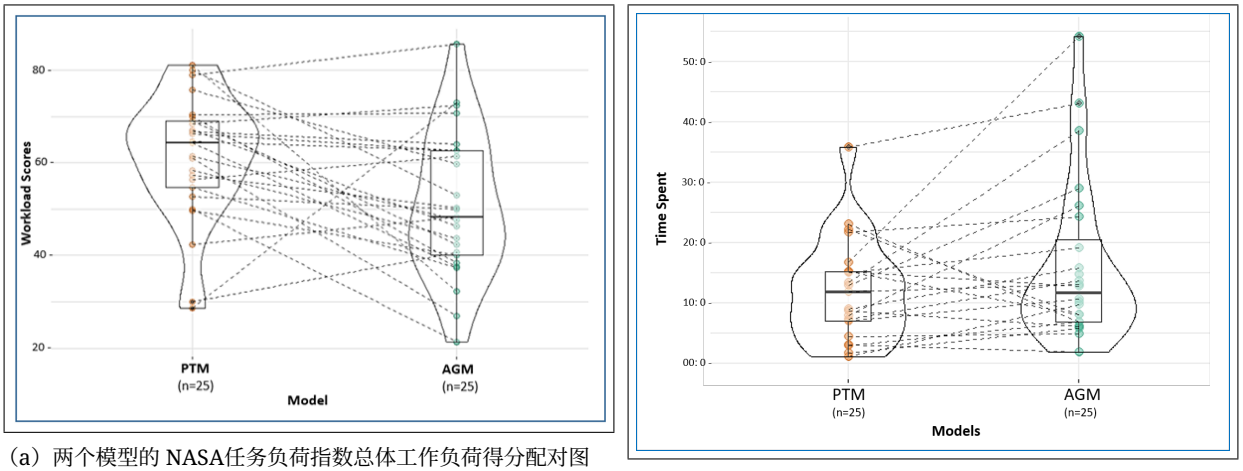
5.7 NASA -TLX问卷（测量工作负荷）：二维论证图模型的工作负荷显著低于 r 工作负荷 PTM。

NASA-TLX 是一种有效的技术，可用于评估相对工作负荷水平，并提供工作负荷的主观评分。工作负荷可定义为工人的能力与工作要求之间的差异，并可能由许多不同因素造成。我们使用 NASA任务负荷指数[60], see附录8.1，分别评估其中的若干因素。

我们观察到，与 PTM 相比，二维论证图模型所需的总体工作负荷显著更低（48.33 对 64.33，威尔科克森符号秩检验：$p = 0.01$*；见表 1、图 10a），这与研究问题4和研究问题5相关。

5.8 完成时间差异：PTM 解读论证所需的时间显著少于二维论证图模型

我们发现，与二维论证图模型相比，PTM 解读一项论证所需的时间显著更短（中位数为 11:13 分钟，对比 13:05 分钟；Wilcoxon 符号秩检验：p=0.015；见图 10b 和表 1）。这与研究问题4相关。不过，二维论证图模型引入了一种新提出的设计与配置（一种解读论证的新机制）。因此，增加交互并弥合对模型的熟悉度差距，或许能够缩小这一绩效差距。这一结果与 [13], 的发现一致：模型性能会受到熟悉度和实际使用的影响。



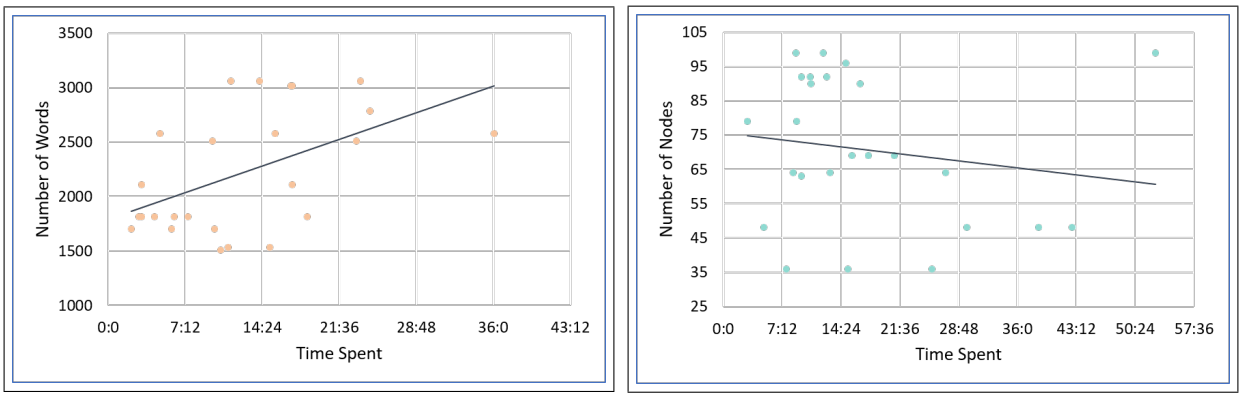
（a）两个模型的 NASA任务负荷指数总体工作负荷得分配对图。

（b）使用两个模型所花时间的配对图。

图10：两个配对图：(a) 两种模型的 NASA任务负荷指数总体工作负荷图。使用二维论证图模型的参与者获得了显著较低的工作负荷评分（$p = 0.01$）。(b) 两种模型的耗时图。使用 PTM模型的参与者耗时显著更短（$p = 0.015$）。黑色虚线连接各参与者的评分。

5.9 耗时与数据量／论证长度：耗时与论证长度成正比

我们对 PTM 和二维论证图模型所花费的时间与论证图表示的长度进行线性回归，其中长度分别为 PTM 中的词数和二维论证图模型中的节点数（见图 11）。我们观察到，PTM 所花费的时间与论证中的词数成正比（p=0.01, $R^2 = 0.258$），见图 11a。另一方面，我们观察到，二维论证图模型所花费的时间与图表示中的节点数之间没有关系（p=0.44, $R^2 = 0.0263$），见图 11b。这与 RQ3 和 RQ4 有关。



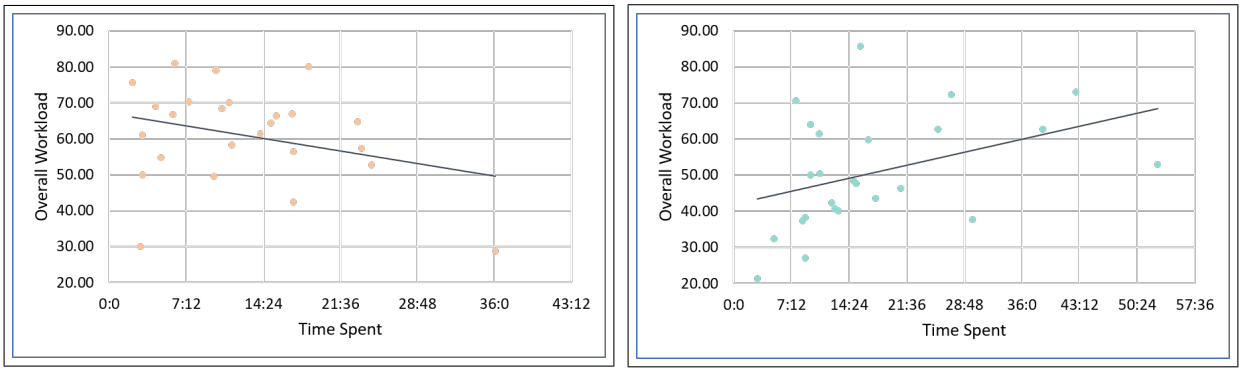
(a) PTM模型

(b) 二维AGM模型

图11：两种模型中耗时与数据量／论证长度的线性回归。PTM中，耗时与论证中的词数存在相关性（p=0.01*）；而二维论证图模型中则不存在相关性（p=0.44）。

5.10 Ti耗时与工作负荷：两种模型的工作负荷与耗时均无相关性

我们对 PTM 和二维论证图模型所花费的时间与总体工作负荷进行线性回归。我们发现，工作负荷与 PTM（p=0.15, $R^2 = 0.089$）或二维论证图模型（p=0.05, $R^2 = 0.157$）所花费的时间之间均无显著相关性，见图 12。这与 RQ3、RQ4 和研究问题5有关。



（a）PTM模型 （b）二维AGM模型

图12：两种模型的耗时与工作负荷的线性回归。工作负荷与PTM（p=0.15）或二维AGM（p=0.05）的耗时之间均无显著相关性。

5.11 对实验的反馈：

参与者认为实验很有帮助[17] 且很有价值 [18]T
able参与者认为实验很有价值（N=23, 92%）。 ment 很有帮助 N=23, 92%），并且喜欢该实验（N=21, 84%）；这些评价依据李克特量表值（>4）得出。上述结果与 RQ1 - RQ6 相关。参见表6，其中包含部分反馈意见。

表2：反馈意见 来自完成实验后的25位参与者。李克特量表的答案为 [1-7], 1 - 强烈不同意，7 - 强烈同意 y

	反馈，中位数 [第一四分位数,Q3]
我喜欢这个实验	6.0 [5, 7]
实验很有帮助	6.0 [5, 6]
实验探讨了一个有价值的主题	6.0 [5, 6]

6 讨论

6.1 总体结果讨论与综合

我们旨在了解丰富的论证表征对批判性解读论证与辩论过程的影响。作为一种表征范式，我们聚焦于论证表征中常见的关系和类别；随着论证挖掘方法近期的发展，这些关系和类别的提取正逐渐成为一种“商品”。正如预期，引入一种与当前基于文本的表征相竞争的新表征模态，意味着终端用户需要经历初始学习，并承担额外的解读负担。参与者在每种模型上花费的时间差异被视为时间测量指标，反映了这种负担（p = 0.015, 5.8）。相比之下，结构化模态（二维论证图模型）在吸引力、刺激性

[17]它有助于并支持解释过程，并从两个不同的视角展示各种论据。[18]其中包含一些重要且尚未尝试过的信息和方法，值得了解。

和效率方面显著优于 PTM，同时也具有更好的可靠性和清晰度（研究问题5——用户如何看待自己对所提出方案的偏好，以及该方案的易用性、有效性和实用性？）。此外，在基线用户体验基准方面，2DAGM 显著优于 PTM，尤其是在清晰度和效率方面。这些结果表明，能够呈现论据结构的可视化模型可以支持对复杂论证和辩论的解读（回应 RQ1——话语层图表示能否支持对复杂论证和辩论的解读？以及 RQ3——论证图式和论证类型能否支持构建这种表征？）。

我们还证实了解读论证所需时间与批判性解读之间固有的权衡，并提出假设：额外的结构虽然会带来时间开销，却会促使用户更加批判且系统地评估论证的质量（主张—前提结构、论证图式和支持性批判性问题）。为回答 RQ4（与纯文本解读相比，所提出的模型在回忆和工作负荷指标方面表现如何？），我们发现，与 PTM 相比，二维论证图模型所需的用户工作负荷也显著较低（$p = 0.01$，5.7）。此外，有一定证据（但未达到统计显著性）表明，与 PTM 相比，基于图的表征在处理较长论证时可能具有更好的扩展性（$p = 0.01$，5.9）。根据回忆任务的测量结果，两种模型对工作记忆的影响相似（$p = 0.247$，5.3）。

能够对论证进行批判，意味着参与者可以直接对话，并能够针对论证的各个部分提出批判性问题。我们还控制了在论证情境中提出的批判性问题数量与潜在偏见和初始立场之间的关系，例如对主题的感知熟悉度以及争议性程度。感知熟悉度越高，提出的批判性问题数量越少，不过这并不会明确影响解释过程（回应 RQ2——哪些类别和关系会对解释过程产生积极影响？）。

我们的研究结果表明，模型配置在影响参与者对论证的解释方面发挥着不同作用，尤其是在参与者对主题立场的改变方面；而且，对于不熟悉的论证主题，模型配置的影响更大。因此，我们首先讨论模型设计的启示与局限、不同模型之间相似的结果，以及二维论证图模型相较于 PTM 所产生的轻微影响。最后，我们讨论二维论证图带来的积极影响以及实验设置的益处。

6.2 二维论证图模型的设计启示与局限

我们的研究结果显示，模型的绩效（耗时）实际上取决于参与者的理解程度以及论证主题的复杂性，却几乎不受模型设计的影响。这表明，可能需要为参与者提供一些支持，帮助他们更好地理解设计及其相关组成部分（节点、分类方案和关系类型）。此外，向参与者传达的关键信息是：阅读更复杂的主题所需时间更短。例如，在复杂度最低的主题中，任务耗时最短的参与者由于已经进行了预先同意评估，因此没有花时间阅读所有前提。另一方面，在复杂度最高的主题中，耗时最长的参与者花了很长时间阅读所有前提，尽管已经进行了预先同意评估，见补充图 A.8。显然，参与者在两个复杂程度不同的主题上花费的时间有显著差异。换言之，参与者准确评估自己阅读论证的能力并理解模型设计至关重要，这样才能对模型进行客观评估。实际上，参与者在二维论证图模型上耗时较长还有其他原因。根据 Walton 分类方案，这些结果是参与者初次与不熟悉的图进行交互时正常且可预期的反应。该图具有表示设计中的一些特定规格和新的配置。模型的不熟悉程度及其配置带来的两种难度，导致时间需求和身体需求增加，见 5.8。值得庆幸的是，工作负荷和数据均未受到时间的显著影响，见 5.9 和 5.10。

6.3 与 PTM 类似的结果

根据实验中的多项指标（如 A-BLEU分数），我们还发现，二维论证图模型的效应与 PTM 的效应相当。工作记忆是主要研究发现之一，A-BLEU分数用于衡量工作记忆（见结果6.3）。由于 A-BLEU是根据参与者回忆句子的能力，衡量候选句（回忆句子）与参考文本之间的相似度，因此很难看出这两个模型之间存在显著差异（见 A.1）。尽管二维论证图模型的配置有助于将文本拆分成句子、将句子归为不同类型、以关系树的形式将它们连接起来，并为前提附上图式类型，但回忆句子取决于参与者的回忆能力，不受模型设计影响。例如，在最复杂的主题“堕胎”中，二维论证图模型的 BLEU分数为 0.05，远低于 PTM（0.12）。在另一个复杂度较低的主题“平板电脑和教科书”中，二维论证图模型（0.2）则远高于 PTM（0.02）。

在复杂度较低或简单的主题“动物园”中，二维论证图模型（0.09）与 PTM（0.095）相当接近（见 A.7）。

6.4 2DAGM 的轻微影响

在其他调查结果中，研究要求参与者分别在阅读每个模型中的主题之前和之后表达自己的看法。这项评估有助于我们了解模型是否会影响参与者的判定。例如，二维论证图以不同于 PTM 的方式呈现论据，使参与者能够看到前提的来源，以及其主张如何削弱或反驳其他主张。参与者可以通过提出附带的批判性问题或类似问题来检验前提陈述的可接受性，从而增强对前提的置信度。此外，二维论证图还显示了论证中哪一方包含更多支持节点。此外，一致性评估以双方各自的主要主张为依据，再由总分得出最终判定；与一次性判定相比，这更接近参与者本人的判定。在二维论证图中，判定是基于论证各部分逐步累积的过程得出的，而非像 PTM 中那样一次性作出，见（ 5.2 ）。这使判定更加合理、公平，也更贴近辩论过程的本质。综合来看，这些原因意味着二维论证图比 PTM 提供了更多可供选择的判定。实际上，在这一点上，我们仍能看到二维论证图的设计对 PTM 有轻微影响。

6.5 AGM 与 PTM 的绩效比较

我们在用户体验、工作负荷和用户感知（偏好）方面的研究结果表明，二维论证图模型的表现优于 PTM，两个模型之间存在显著差异，详见 5.1、5.5、5.6 和 5.7 节。这些发现可能包含主要影响测量结果的偏差，但不太可能成为影响我们主要结果的混杂变量，即模型是否会影响用户体验或工作负荷。这是因为我们的所有分析仅使用了两份标准问卷，这些问卷基于若干定量条目和量表，详见第 9 节和 A.5 节。此外，我们的实验任务、一致性评估及其他问题均以相同方式设计，没有偏向任一模型。因此，我们认为我们的结果具有较高的真实性和可靠性，详见 A.9 和 A.10 节。

6.6 仅针对 2DAGM：选定的批判性问题数量与主题熟悉度相关

最重要的是读者做出的决定，以及该决定是否基于置信度和可接受性的理由。因此，我们发现，可以使用其他衡量指标来评估论证或辩论，这些指标是评估论证和辩论的自然标准，例如熟悉度和论证的呈现方式。例如，如果你面对两个熟悉程度不同的论证，那么对于较为熟悉的论证主题，读者会更有信心，也更确信自己所做的决定。同样，阅读论证的方式以及提供支持的材料或工具，都可能对读者针对该论证做出的决定产生积极或消极的影响。比如，了解前提陈述的来源，或了解双方的大多数理由，可以增强读者的信心。在本研究中，我们可以看到论证呈现方式的影响。一个有趣的结果是，呈现方式能够增强读者的信心，即使面对不熟悉的主题也是如此，因此读者无需提出许多批判性问题，见 5.4。

6.7 使用“Uni-Excel”插件包的优势

该项目在桌面设备上向参与者展示，并通过 Microsoft Teams 共享访问权限；他们必须拥有一个匿名账户。该项目嵌入了“Uni-Excel”插件包，在实验期间收集、整理和保存结果方面具有显著优势且非常实用。其主要优势在于，如果发生技术问题导致实验中断，参与者无需从头开始，只需从中断处继续即可，因为结果会在每一帧之后收集，这样可以节省时间。三四名参与者遇到了这种情况，原因可能是网络连接问题或其他意外断连问题。

7 未来工作

基于本研究获得的实证证据，我们希望未来开展更多验证性研究，以验证这些结果，并考察它们在不同情境和表征形式、不同人群以及不同任务中的影响。事实上，我们将采用更严谨的论证表征，研究明确呈现的论证图式类型将如何影响批判性解释过程。本研究将依据对人类互动中细微性、科学性、分析性和自发性的测量结果得出结论。最终，

我们承认，解释是一个复杂的概念，也是一种多维构念。尽管我们尝试使用一组指标来衡量解释，包括自然语言处理功能、推理任务和调查方法，但仍需开展未来研究，以理解如何对复杂论证或辩论进行批判性解读。

8 结论

随着日常生活中的论证变得越来越复杂、分叉并不断扩展，研究人们如何批判性地解读论证变得愈发重要，也需要采用文本之外的有效解读方式。本研究探讨了与文本形式的论据相比，论证的结构化表示（二维护论证图模型）能否改善论辩话语情境中的解读。本研究开展了一项实验，以两种模型呈现了10个论证主题：PTM和二维论证图模型。模型的评估基于任务相关构念（耗时、工作记忆、努力程度和工作负荷）以及用户体验（25名参与者）。在吸引力、清晰度和效率方面，结构化模态（二维护论证图模型）显著优于文本模态（PTM），在可靠性和刺激性方面也表现更佳。这些结果表明，能够揭示论据结构的可视化模型有助于解读复杂论证和辩论。我们还证实了解读论证所花时间与批判性解读之间固有的权衡关系：与文本模态相比，二维论证图模型所需时间更长。我们推测，额外的结构虽然带来了时间开销，却会促使用户在判断论证质量时更加批判、更加系统。我们还发现，与PTM相比，二维论证图模型显著降低了用户的工作负荷。此外，也有一定证据表明，与PTM相比，结构化模型在处理较长论证时可能更具可扩展性。两种模型对工作记忆的影响相似。未来的研究可进一步探明，论证图式中的哪些类别与批判性解读最为相关。

致谢

我们感谢所有志愿者、参与者以及全体工作人员，他们为本研究和实验提供了有益的评论。作者谨向所有与他们共事的人表示诚挚谢意。

参考文本

[1] Bruno Yun 和 Pierre Bisquert。“DAGGER：Datalog + / - 论证图生成器”。载于：（2018），第 1841–1843 页。[2] Douglas Walton 和 Fabrizio Macagno。“ 论证图式分类系统 ”。载于：《论证与计算》6.3（2015），第 219–245 页。ISSN：19462174。DOI：10.1080/19462166.2015.1123772。[3] Javier Tamames 和 Victor de Lorenzo。“EnvMine：用于自动提取上下文信息的文本挖掘系统”。载于：《*BMC* 生物信息学》11.1（2010 年 6 月），第 1–10 页。ISSN：14712105。DOI：10.1186/1471-2105-11-294。[4] Manfred Stede。“从论证图示到文本中的论证挖掘：”载于：《国际认知信息学与自然智能杂志》7（2013 年 1 月），第 1–31 页。DOI：10.4018/jcini.2013010101。[5] *Course ImprovingReasoning*。网址：https://sites.google.com/site/courseimprovingreasoning/ home。[6] Markus Brenneis 和 Martin Mauve。“结合协同过滤的在线审议式论证”。载于：《人工智能与应用前沿》。第 326 卷。IOS Press BV，2020 年 8 月，第 453–454 页。ISBN：9781643681061。DOI：10.3233/FAIA200530。[7] Nancy L Green、Michael Branon 和 Luke Roosje。“用于批判性思考国际政治的论证图式与可视化软件”。载于：《论证与计算》10（2019），第 41–53 页。DOI：10.3233/AAC-181003。[8] Simon Cullen 等。“提升分析推理与论证理解：一项关于论证可视化的准实验实地研究”。载于：《*npj* 学习科学》3（2018 年 12 月）。DOI：10.1038/s41539-018-0038-5。[9] Mathias Rieche。多利益相关方开发流程中的论证可视化改进——一个原型设计案例。技术报告。[10] Pietro Baroni 等。“通过定量论证自动评估设计方案”。载于：《论证与计算》6.1（2015），第 24–49 页。DOI：10.1080/19462166.2014.1001791。网址：http://dx.doi. org/10.1080/19462166.2014.1001791。

[11] Aamna Al-Shehhi、Zeyar Aung 和 Wei Lee Woon。“用于协作式空间决策与知识建构的论证可视化和叙事方法：海上风电场项目案例研究”。载于：《计算机科学讲义》（包括《人工智能讲义》和《生物信息学讲义》子系列）。第 9518 卷。施普林格出版社，2015 年，第 135–154 页。ISBN：9783319274294。DOI：10.1007/978-3-319-27430-0_10。网址：https://link.springer.com/chapter/10.1007/978-3-319-27430-0_10。[12] Robert Bembenik 和 Piotr Andruszkiewicz。“面向地方政府在线咨询审议模型的自动论证抽取与可视化”。载于：《计算机科学讲义》（包括《人工智能讲义》和《生物信息学讲义》子系列）。第 9809 卷 LNCS。施普林格出版社，2016 年，第 74–86 页。ISBN：9783319440385。DOI：10.1007/978-3-319-44039-2_6。网址：https://link.springer.com/chapter/10.1007/978-3-319-44039-2%7B%5C_%7D6。[13] Mathias Riechert。多利益相关方开发流程中的论证可视化改进——一个原型设计案例。第 18 卷。信息系统管理研究论文。柏林：柏林工业大学出版社，2018 年。DOI：10 . 14279 / depositonce - 6743。网址：http : / / dx . doi . org / 10 . 14279/ depositonce-6743。[14] Marco Lippi 等。“论证挖掘：现状与新兴趋势”。载于：《ACM互联网技术汇刊》16.10（2016），第 10 页。DOI：10.1145/2850417。网址：http://dx.doi.org/10.1145/ 2850417。[15] Fabrizio Macagno、Douglas Walton 和 Chris Reed。“论证图式：历史、分类与计算应用”。载于：《IFCoLog逻辑及其应用期刊》4.8（2017）。网址：http: //dougwalton.ca/papers%20in%20pdf/17IFColog%20SCHEMES.pdf。[16] Arthur C. Hastings。“论证中推理模式的重新表述”。博士论文。伊利诺伊州埃文斯顿，1963 年。[17] Manfred Kienpointner。《日常逻辑：论证模式的结构与功能》。技术报告。宾夕法尼亚州立大学，1994 年。网址：https://www.jstor.org/stable/40237812。[18] Douglas N. Walton。可推定推理的论证图式。L. Erlbaum Associates，1996 年，第 218 页。ISBN：9780805820720。[19] Wayne Grennan。非形式逻辑：议题与技巧。麦吉尔女王大学出版社，1997 年，第 309 页。ISBN：9780773566453。[20] John；Lawrence、Chris Reed 和 John Lawrence。“利用论证图式结构进行论证挖掘”。载于：《人工智能与应用前沿》287（2016），第 379–390 页。DOI：10.3233/978-1-61499- 686-6-379。[21] Douglas Walton、Chris Reed 和 Fabrizio Macagno。论证图式。剑桥大学出版社，2008 年 1 月，第 1–443 页。ISBN：9780511802034。DOI：10.1017/CBO9780511802034。[22] John Lawrence 和 Chris Reed。“论证挖掘：综述”。载于：（2019）。DOI：10 . 1162 / COLI。网址：https://doi.org/10.1162/COLI。[23] S. Nazli Can 和 Deniz Saribas。“促进批判性评价的论证工具：通过 MEL 图示探究职前教师对社会科学议题的评价水平”。载于：《科学与教育》28.6-7（2019 年 9 月），第 669–687 页。ISSN：15731901。DOI：10.1007/s11191-019-00068-8。网址：https://doi.org/10.1007/s11191-019-00068-8。[24] Tobias Milz。“硕士论文 ArguE——论证挖掘与建模”。载于：（2017）。[25] Ramzan Talib 等。文本挖掘：技术、应用与问题。技术报告 11。2016 年。网址：www.ijacsa. thesai.org。[26] Elena Cabrio 和 Serena Villata。“论证挖掘五年：数据驱动分析”。载于：IJCAI 人工智能国际联合会议。第 2018-July 卷。人工智能国际联合会议组织，2018 年，第 5427–5433 页。ISBN：9780999241127。DOI：10.24963/ijcai.2018/766。[27] Dietrich Trautmann。基于方面的论证挖掘。技术报告。arXiv：2011.00633v1。网址：https://github. com/trtm/ABAM。。[28] Reid Swanson、Brian Ecker 和 Marilyn Walker。论证挖掘：从在线对话中提取论据。技术报告，第 217–226 页。网址：http://www.createdebate.com/。[29] Jie Ma 等。“用于事件论元抽取的资源增强型神经模型”。载于：ArXiv abs/2010.03022（2020）。[30] Christian Stab 等。ArgumenText：从异构来源中搜索论据。技术报告，第 21–25 页。网址：https://www.ukp.tu-darmstadt.de/。[31] Marco Lippi 和 Paolo Torroni。“MARGOT：一个论证挖掘网络服务器”。载于：《应用专家系统》65（2016 年 12 月），第 292–303 页。ISSN：09574174。DOI：10.1016/j.eswa.2016.08.050。

[32] Walton Douglas. “Douglas Walton 运用论证图式进行论证挖掘 :: SSRN”。发表于：(2011年4月)，第38–64页。URL：https://papers.ssrn.com/sol3/papers.cfm?abstract%7B%5C_%7Did=2034890.[33] Kawsar Noor、Anthony Hunter 和 Astrid Mayer. 基于社交网络上表达的患者经历分析医学论证。技术报告。URL：https://github.com/robienoor/NLTKForumScraper.[34] Anthony Hunter. “从概率信息生成实例化论证图”。发表于：(2020)。DOI：10.3233/FAIA200165.[35] Artem Chernodub 等。TARGER：触手可及的神经论证挖掘。技术报告。[36] Henning Wachsmuth 等。为万维网构建论证搜索引擎。技术报告。2017年，第49–59页。URL：http://www.arguana.com/software.html.[37] Mennatallah El-Assady 等。“ConToVi：运用主题空间视图探索多方对话”。发表于：Computer Graphics Forum 35.3（2016年6月），第431–440页。ISSN：14678659。DOI：10.1111/cgf.12919.[38] Mirko Lenz 等。迈向将文本转换为论证图的论证挖掘流水线。技术报告。2020年。arXiv：2006.04562v2.[39] John Lawrence 和 Chris Reed. 结合论证挖掘技术。技术报告。2015年，第127–136页。URL：http://wordnet.princeton.edu/.[40] Marco Lippi 和 Paolo Torroni。“论证挖掘：机器学习视角”。发表于：形式论证理论与应用。Elizabeth Black、Sanjay Modgil 和 Nir Oren 编。施普林格国际出版，Cham，2015年，第163–176页。ISBN：978-3-319-28460-6.[41] Fabian Sperrle 等。VIANA：论证的可视化交互式标注。技术报告。arXiv：1907.12413v1.[42] 主页 - ProCon.org。URL：https://www.procon.org/（访问于2021年3月8日）。[43] 标注示例 - brat 快速标注工具。URL：https://brat.nlplab.org/examples.html（访问于2021年3月1日）。[44] K. Kucher 和 A. Kerren。“文本可视化技术：分类体系、可视化综述与社区见解”。发表于：2015年 IEEE 太平洋可视化研讨会（PacificVis）。2015年4月，第117–121页。DOI：10.1109/PACIFICVIS. 2015.7156366.[45] Akrivi Katifori 等。本体可视化方法——调查。2007年11月。DOI：10.1145/1287620.1287621.[46] Unity实时开发平台 | 3D、2D、虚拟现实与增强现实引擎。2023年。URL：https://unity.com/.[47] Mei Yang、Sarah Alice Gaggl 和 Sebastian Rudolph。“Neva——论证框架的扩展可视化”。发表于：第八届国际论证计算模型会议（COMMA 2020）论文集。第326卷。《人工智能与应用前沿》。2020年9月，第477–478页。DOI：10.3233/FAIA200542.[48] Douglas Walton。“构建论证反驳意见查找系统”。发表于：()。DOI：10.1007/s10503- 012-9261-z。URL：http://www.dougwalton.ca.[49] Esin Durmus、Faisal Ladhak 和 Claire Cardie。“语用与话语语境在判定论证影响中的作用”。发表于：2019年自然语言处理实证方法会议与第九届国际自然语言处理联合会议（EMNLP-IJCNLP）论文集。中国香港：计算语言学协会，2019年11月，第5668–5678页。DOI：10.18653/v1/D19- 1568。URL：https://www.aclweb.org/anthology/D19-1568.[50] D. Kiesel 等。“论文中的论证可视化分析”。发表于：IEEE可视化与计算机图形学汇刊 27.2（2021年2月），第1139–1148页。ISSN：1941-0506。DOI：10.1109/TVCG.2020.3030425.[51] Jordan Beck、Bikalpa Neupane 和 John Carroll。“管理在线辩论社区中的冲突：凸显 Kialo 版主的信念与价值观”。发表于：(2018)。DOI：10 . 31219 /osf . io /cdfq7。URL：https://osf.io/cdfq7/.[52] Iwan Ittermann 和 Brian Plüss。“PEOPLES：从消息私人回复到双边对抗性在线对话中的去极化助推”。英文。发表于：论证计算模型。Henry Prakken 等编。《人工智能与应用前沿》。第八届国际论证计算模型会议，COMMA 2020；会议日期：2020年9月8日至2020年9月11日。荷兰：IOS出版社，2020年，第467–468页。ISBN：9781643681061。DOI：10.3233/FAIA200537.[53] Michael Hoffmann 和 Jason Borenstein。“通过协作学习与论证可视化方法理解结构不良的工程伦理问题”。发表于：《科学与工程伦理》20.1（2014年3月），第261–276页。ISSN：13533452。DOI：10.1007/s11948-013-9430-y。URL：https://pubmed.ncbi. nlm.nih.gov/23420467/。

[54] Jeongyun Han 等。“面向面对面协作论证自适应支持的学习分析仪表板”。发表于：《计算机与教育》163（2021），第104041页。ISSN：0360-1315。DOI：https://doi.org/ 10.1016/j.compedu.2020.104041。URL：https://www.sciencedirect.com/science/article/pii/ S0360131520302396.[55] Michael H.G. Hoffmann。“通过论证映射激发反思与自我纠正推理：三种方法”。发表于：《*Topoi*》37.1（2018年3月），第185–199页。ISSN：15728749。DOI：10.1007/s11245-016- 9408-x.[56] Thomas F. Gordon、Henry Prakken 和 Douglas Walton。“论证的 Carneades 模型与举证责任”。英文。发表于：《人工智能》171.10-15（2007）。关系：http://www.rug.nl/ 权利：格罗宁根大学，第875–896页。ISSN：0004-3702。DOI：10.1016/j.artint.2007.04.010.[57] Dan Cartwright 和 Katie Atkinson。“运用计算论证支持电子参与”。发表于：《*IEEE*智能系统》24.5（2009），第42–52页。ISSN：15411672。DOI：10.1109/MIS.2009.104.[58] M. Riechert 等。“运用论证可视化促进开发流程的透明度”。发表于：2018年。[59] 用户体验问卷（*UEQ*）。URL：https://www.ueq-online.org/（访问于2022年2月22日）。[60] *NASA*艾姆斯研究中心 *TLX - NASA*任务负荷指数应用。URL：https://humansystems.arc.nasa.gov/groups/tlx/tlxapp. php（访问于2021年6月29日）。

附录

本节将介绍有关论证模型的补充背景信息和详细内容。

论证结构：随机主题

不同的辩论主题收集自“procon.org”网站。随后，选取了 10 个不同主题，分别列于以下不同领域：

1. **教育：** 禁书（家长或其他成年人应当能够禁止学校和图书馆收藏某些书籍。）以及平板电脑与教科书之争（高中应以平板电脑取代教科书。） 2. **健康与医学：** 肥胖（肥胖是一种疾病。）、堕胎（堕胎应当合法。）以及素食主义（人们应当成为素食者）。 3. **科学与技术：** 瓶装水（应当禁止瓶装水。）、动物实验（应当将动物用于科学或商业测试。）、气候变化（人类活动是全球气候变化的主要原因。）、动物园（动物园保护动物。动物园应当存在。）以及能源（替代能源能够有效取代化石燃料。）

论证主题复杂度

使用 Python 代码计算论据的复杂度。我们测量词语总数、句子总数、平均词频和平均词长。这些是一些基本指标，可用于量化文本的特征，并帮助我们了解文本的语言和内容，见图 A.3。我们发现，论据的复杂度可通过词汇多样性[19], 词汇密度[20], 差异度[21] 和连接词比例[22]来衡量。这些指标有助于增强文本的连贯性和衔接性，见图 A.4。

在可读性方面，我们使用弗莱施阅读易读性评分[23], 斯帕奇可读性评分[24], 戴尔-查尔可读性评分[25] 和SMOG可读性分数[26]来衡量论证主题。这是四种不同的可读性指标，用于评估文本的可读性，见图 A.5。

论证结构：沃尔顿批判性问题

- 专家意见论证
  – E 作为专家来源有多可信? – E 是 A 所属领域的专家吗? – E 断言了什么，从而能够推出 A? – E 作为来源是否可靠? – A 是否与其他专家的断言一致? – E 的断言是否基于引用的证据?

[19]“词汇多样性” 是衡量文本中词汇多样性的指标，计算方法是用不同词汇的数量除以 e 词语总数。

[20]“词汇密度” 是衡量文本中实词与虚词比例的指标。实词是指承载句子 y 含义的词 e，而虚词则用于将实词连接在一起。

[21]“差异度”是衡量两个文本之间差异的指标，计算方法是一个文本中未出现在另一个文本中的不同词汇所占的比例。

[22]“连接词比例”是指文本中用于连接观点或句子的词语所占比例，例如连词（如“和”“但是”）、介词（如“在.....中”“在.....上”）以及代词（如“他”“她”）。

[23]弗莱施阅读易读性评分是一个根据文本的平均句子长度和每个词的平均音节数来计算其可读性的公式。

[24]斯帕奇可读性评分是衡量文本简易程度的指标，依据句子长度和所使用的熟悉词汇数量进行评估。它通过一个公式计算，该公式考虑了句子中的音节数以及所使用的简单词汇数量。该评分范围为 0 到 100，分数越高表示可读性越强。

[25]戴尔-查尔可读性评分是一种可读性指标，通过考量文本中生词出现的频率以及所用词语的长度来衡量文本的可读性。戴尔-查尔评分为 0 表示文本非常易读，且几乎不含或完全不含生词。

[26]简明废话指数（SMOG）是另一种估算文本可读性的公式，它通过统计多音节词的数量并应用数学公式进行计算。SMOG 是一种常用于健康素养材料的流行方法。

– A 是否与近期研究一致? – A 是否属于结论的主题领域?

- 基于知情立场的论证
  – “P”是否处于能够判断 A 是否为真的知情立场? – “P”是否诚实（值得信赖、可靠）的来源? – “P”的断言是否意味着 A 为真或为假? – P 的断言是否基于引用的证据?

- 类比论证
  – C1 与 C2 之间是否存在会削弱所引相似性说服力的差异? – A 在 C1 中为真/假吗? – C1 与 C2 之间是哪一类相似性?

- 因果论证
  – 因果概括的强度如何? – 引用的证据是否足够有力，足以支持该因果概括? – 在此情况下，是否存在其他可能干扰该效应产生的理由?

- 基于正面／负面后果的论证
  – 所引用的后果发生（必然发生、可能发生）的可能性有多大? – 有什么证据支持所引用的后果将会（必然、可能）发生这一主张? – 这些证据是否足以充分支持主张的力度? – 是否存在其他应予考虑的相反后果?

- 基于证伪的论证
  – 如果假设为真，那么陈述某一事件的命题是否为真? – 是否观察到陈述某一事件的命题为真（假）? – 除了该命题之所以为真是因为给定假设为真之外，是否还存在其他原因可以解释其为真?

实验模型任务

1. **任务1：回忆主张（BLEU分数）**A-BLEU分数（双语评估替补分数）是一种评估指标方法，用于将参考句与用户生成句子（候选句）进行比较，并表明候选句与参考句的相似程度或接近程度。BLEU指标的取值范围为0到1；它能够识别参考文本与候选句之间在词语选择和词序方面的合理差异，其中越接近1的分数表示相似度越高。从一个熟悉的指标——**精确率**开始，其定义如下：

$$Precision = \frac{No.OfCandidateWordsOccurringInAnyReference}{TotalNo.OfWordsInTheCandidates} \tag{1}$$

BLEU指标的取值范围为0到1；它能够识别参考文本与候选句之间在词语选择和词序方面的合理差异，其中越接近1的分数表示相似度越高。**形式化表示：**

$$BLEU = BP \times \exp(\sum_{n=1}^{N} w_n * \log(p_n)) \tag{2}$$

其中，BP是简短惩罚项；当候选句长度与任一参考文本的长度相同时，BP将为1；N是n 元语法的数量，p_n 是修正精确率，w_n 是相应的权重，实际应用中通常采用均匀权重。

为了衡量参考文本与候选句之间的匹配度，我们比较候选句与参考文本的 n 元语法精确率，并采用修正精确率的形式来比较多个候选句与多个参考文本。

2. **任务2：选择性批判性问题**我们计算了所选批判性问题的总数。形式化地说：

$$Total_Matches = \sum_{n=1}^{N} COUNTIF(R_{1\to m,C}) \tag{3}$$

其中，N为批判性问题总数，C为判定标准（批判性问题已勾选（回答为“TRUE”），R为包含批判性问题的行。

实验问卷

1. **NASA-TLX问卷**NASA-TLX方法的测量分为两个阶段，即对各量表进行比较（成对比较）以及为工作评分（事件评分）。NASA任务负荷指数是一种多维评分方法，根据六个子量表评分的加权平均值计算总体工作负荷评分：心理需求、身体需求、时间需求、自身表现、努力程度和挫折感。这组六个评分量表用于评估用户在不同任务中的体验。根据 NASA-TLX[60], , 有 15 张可能的工作负荷来源比较卡片，其中包含六个量表的两两比较。实施 TLX 包括两个步骤。

   第一步中，参与者会思考自己要执行的任务，并查看六个量表因素的所有成对组合，以确定哪一项与自己对该任务工作负荷的定义更相关。这样，参与者需要进行 15 次成对比较。例如，他们需要判断，对于“你最近执行的特定任务”，身体需求或心理需求哪一项“对工作负荷的影响更重要”（见图 A.9a）。

   第二步中，参与者需要根据从极低到极高的量表，对六个维度分别进行评分。每个量表均通过一个滑块呈现，滑块将 0 到 100 的范围按 5 分的增量划分（见图 A.9b）。要计算某项任务的工作负荷，需将六个量表的分值分别乘以第一步中得出的权重，得到“调整后评分”；将六个量表的调整后评分相加，再除以 15，即可得出参与者在该任务中的总体工作负荷评分。

2. **用户体验问卷 UEQ** 用户体验问卷 UEQ[59], 英文版包含实用质量和享乐质量两个量表，用于衡量用户体验和交互式产品的质量。UEQ 可用于比较用户对两种产品的体验，通常是同一产品的两个版本：旧版或既有版本，以及新版。该问卷用于确定新版是否能提供更好的用户体验。通过主成分分析，从该数据集中提取出代表各量表的条目。这些条目采用语义差异量表的形式，即每个条目由两个含义相反的词语构成。UEQ 使用七级量表，以减少这类条目中众所周知的中心倾向偏差。条目按 1 至 7 计分，其中 1 表示最负面的回答，4 表示中立回答，7 表示最正面的回答。UEQ 包含 26 个条目，分为 6 个主要量表：吸引力、清晰度、效率、可靠性、刺激性和新颖性（见图A.10）。

补充表格

1. **表A.1对应任务1：回忆主张句子**。

表A.1：PTM 和二维论证图模型任务1的主张候选项示例。其中，PTM模型中主题“肥胖”的三个候选陈述，以及二维论证图模型中主题“平板电脑与教科书”的四个候选陈述，获得了最高的 BLEU分数。表中包含参与者项目生成的ID，用于展示各参与者的个人结果。

A- ID T606	PTM - 任务1：回忆主张（候选陈述）
1-	这是其他因素造成的结果，因此并非一种疾病。
2-	这是一种疾病，因为某些激素的产生会导致其他器官功能失常。
3-	这不是一种疾病，因为可以通过向人们普及饮食习惯方面的知识来预防。
B- ID T692	二维论证图模型 - 任务1：回忆主张（候选陈述）
1-	平板电脑会导致学生出现视力问题。
2-	平板电脑的生产更加环保。
3-	平板电脑会让学生更容易分心。
4-	研究表明，使用平板电脑学习相同学习材料的学生考试成绩更高。

2. **表A.2为PTM模型（任务1）：主张参考陈述**。

表A.2：PTM模型中任务1的主张参考示例，其中包含论证主题（肥胖）中每一方（正方与反方）的三条参考文本。

A-	**正方／赞成方的PTM主张句**
1-	参考文本1：它符合疾病的定义。
2-	参考文献2：它会缩短预期寿命，并损害身体的正常机能。
3-	参考文献3：它可能由遗传因素导致。
B-	**反对／赞同方中的 PTM主张句**
1-	参考文献1：它是其他疾病的可预防危险因素。
2-	参考文献2：这是饮食过量的结果。
3-	参考文献3：这是运动不足导致的。

3. **表 A.3 是二维AGM模型（任务1）的论点参考陈述**。

表 A.3：二维AGM模型中任务1的主张参考示例。该示例针对论证主题「平板电脑与教科书」，在正方与反方各自的一侧均包含来自该主题的三条参考文本。

A-	**二维 AGM 论点句（正方／赞成方）**
1-	参考文本1：大多数教师和学生都支持平板电脑，而且它们轻得多，比纸质教科书轻得多，并且能提高标准化考试成绩。
2-	参考文本2：平板电脑可以存储数百本教科书，并通过.....保护环境减少印刷量。
3-	参考文献3：提升学生互动性和创造力，而且数字教科书比印刷版教科书更便宜。
B-	**二维 AGM 主张句（赞成/反对方）**
1-	参考文献1：它们价格昂贵，容易分散学生的注意力、容易损坏，而且维修成本高／耗时长。
2-	参考文献2：他们表示，平板电脑会导致眼疲劳、头痛和视力模糊，这让学生有更多借口不做家庭作业，作业。
3-	参考文献3：需要昂贵的 WiFi 网络，而且会随着新技术的出现而很快过时。

4. **表A.4：二维AGM模型（任务2）：选择性批判性问题示例**

表A.4：二维AGM任务2的一个示例结果，包含11个批判性问题（CQs），这些问题选自沃尔顿批判性问题（见附录8）。该表包含这些批判性问题及所选选项，其中TRUE = 表示选定的批判性问题，FALSE =表示未选定的批判性问题。

No	二维AGM - 任务2：回顾批判性问题（CQs）	已选择
1-	作为专家来源，专家的可信度如何?	TRUE
2-	断言是否与其他专家的断言一致?	FALSE
3-	PTK 是诚实（可信、可靠）的来源吗?	TRUE
4-	PTK 的断言是否基于引用的证据?	FALSE
5-	这些内容是否足以支持主张的力度?	错误
6-	是否存在其他相反后果?	TRUE
7-	因果概括的强度如何?	错误
8-	在此案例中，是否存在导致该效应的其他理由?	TRUE
9-	C1 与 C2 之间属于哪种相似性?	TRUE
10-	断言在 C1 中是真的吗?	错误
11-	如果假设为真，那么报告某个事件的命题是否为真?	错误

5. **表 A.5 co 包含 UEQ 分析工具的结果（通过 t 检验），用于比较两个模型 s**

表 A.5：用户体验问卷（UEQ）展示了两个模型在五个量表上的结果：吸引力、清晰度、效率、可靠性和刺激性。该表列出了 PTM 和二维论证图模型的量表均值、标准差（STD）、置信度以及相应的 5% 置信区间（CI）值。置信区间用于衡量量表均值估计的精确程度。置信区间越小，估计的精确度越高，结果也越可信。

PTM模型				
量表	Mean	STD	置信度	（5%置信区间）
吸引力	0.00	1.63	0.64	[-0.64,0.64]
清晰度	0.79	1.32	0.52	[0.27,1.30]
效率	0.70	1.54	0.60	[0.10,1.30]
可靠性	0.56	1.44	0.56	[0.00,1.12]
刺激性	0.42	0.74	0.29	[0.13,0.71]
二维AGM模型				
量表	Mean	STD	置信度	（5%置信区间）
吸引力	1.18	0.82	0.32	[0.86,1.50]
清晰度	1.37	0.73	0.29	[1.09,1.66]
效率	1.48	0.93	0.36	[1.12,1.84]
可靠性	1.12	0.96	0.38	[0.74,1.50]
刺激性	0.83	0.59	0.23	[0.60,1.06]

6. **表 A.6 包含参与者的所有反馈意见**

表 A.6：该表包含两列：第一列包含项目生成标识符，第二列包含参与者在实验结束时提供的评论。

参与者 ID	评论
T756	Hanadi，感谢你有趣的实—— 实验和热情款待.....祝你好运 :)..
T298	我喜欢这个模型的想法，但图形用户界面可以 改进一下
T606	谢谢，祝你的研- 究顺利 :)
T430	我会向未来的参与者推荐 使用鼠标而不是触控板。
T014	我从未见过这种论证风格 以这种方式呈现。这非常 有助于我对一个 有争议的主题作出自己的判断。通常， 仅使用基于文本的论证很难做到这一点。
T941	挫折感 with 鼠标 控制 and 滚动，很难探索所有内容， 节点布局有些困难 而邻近关系使一些事情更加 比其他项更突出。一些滑块 让人困惑（例如，哪一侧代表完美）。
T966	AG图看起来非常易于阅读。
T062	这很有趣，但有点长。
T809	说明往往非常不清楚，而且 整体体验有些令人沮丧—— 我经常发现，选项的表述方式 并不利于理解。 also.
T115	或许可以在各项主张之间添加箭头， 以表明哪些内容受到了削弱。
T020	我不熟悉批判性问题这一部分， 发现它很难理解。
T498	更好的布局设计或许能提升用户 体验很多。
T177	我很喜欢第二种模型在清晰度和组织性方面的表现 ，我唯一想 改进的是图形可视化 ：）
T375	good

补充图

1. **图 A.1 是论证结构的一个示例。**

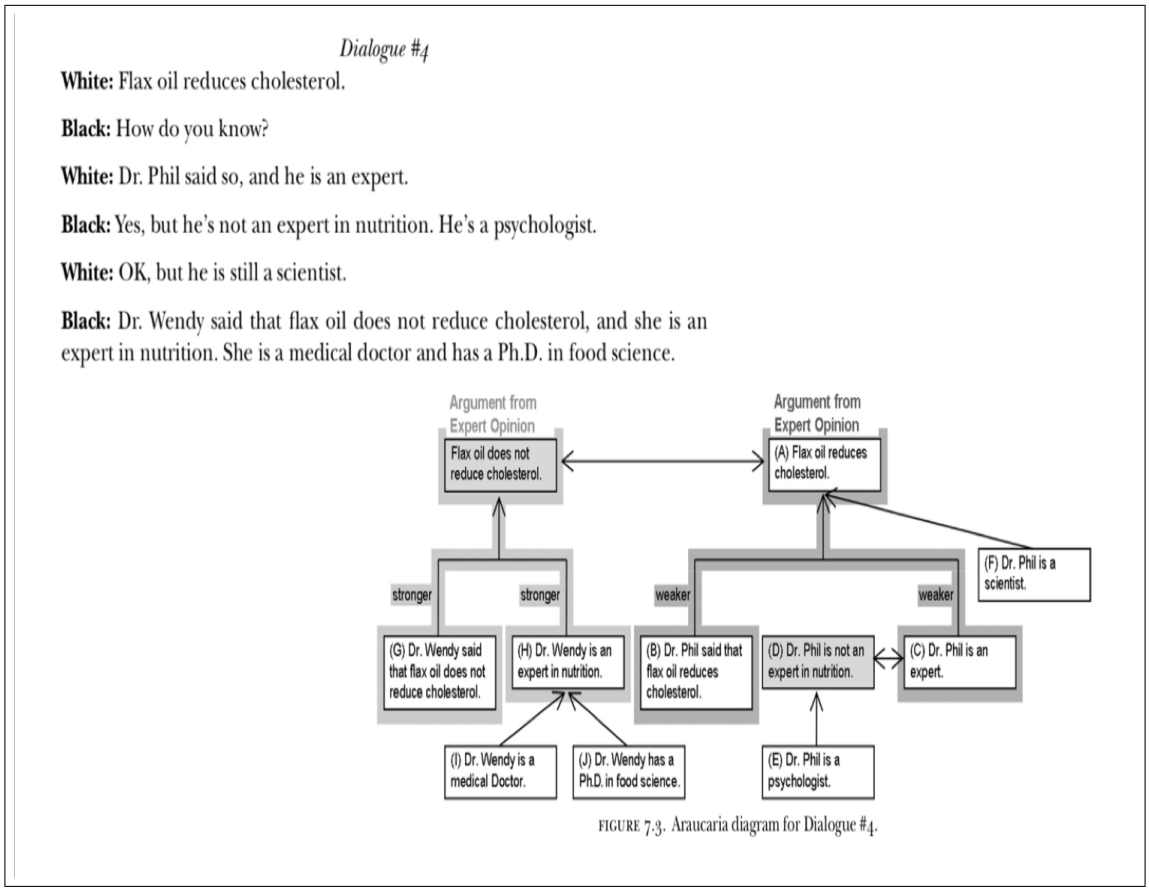


图 A.1：Walton 论证结构的一个示例。它展示了关于“亚麻籽油可降低胆固醇”的论证结构。

2. **图 A.2 展示了 Milz[24] 提出的论证结构，该结构基于 Walton 的论证结构。**3. **图 A.3、A.4 和 A.5 分别展示了论证文本复杂度（包括词语总数、句子总数、平均词频和平均词长）、词汇多样性、词汇密度、词性差异度和连接词比例，以及弗莱施阅读易读性评分、斯帕奇可读性评分、戴尔-查尔可读性评分和 SMOG 可读性分数。**4. **图 A.6 展示了二维AGM模型的评估句**5. **图 A.7 展示了所有参与者的 A-BLEU分数条形图**6. **图 A.8 展示了所有参与者所用时间的条形图**7. **图 A.9 为 NASA-TLX问卷面板**8. **图 A.10 为 UEQ问卷面板**9. **图 A.11 展示了实证实验的工作流程，并包含对图、表和结果的引用。**

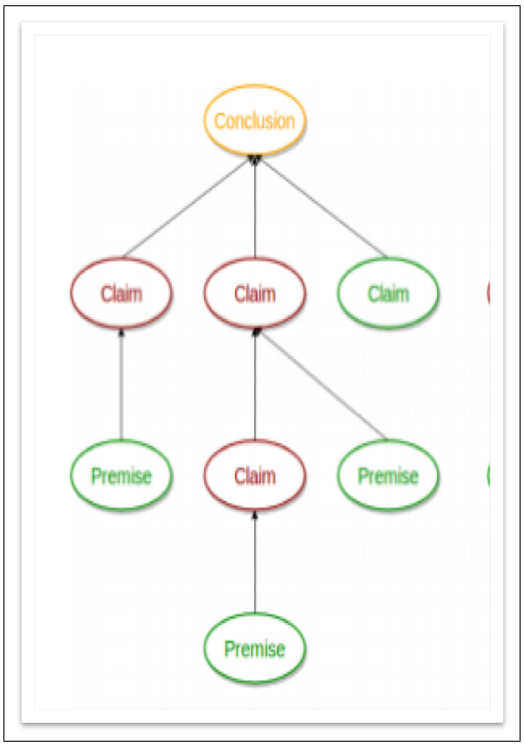


图 A.2：Milz 的[24] 论证结构基本模板示例，其中包含论证中的三个主要单元：结论、主张和前提节点。

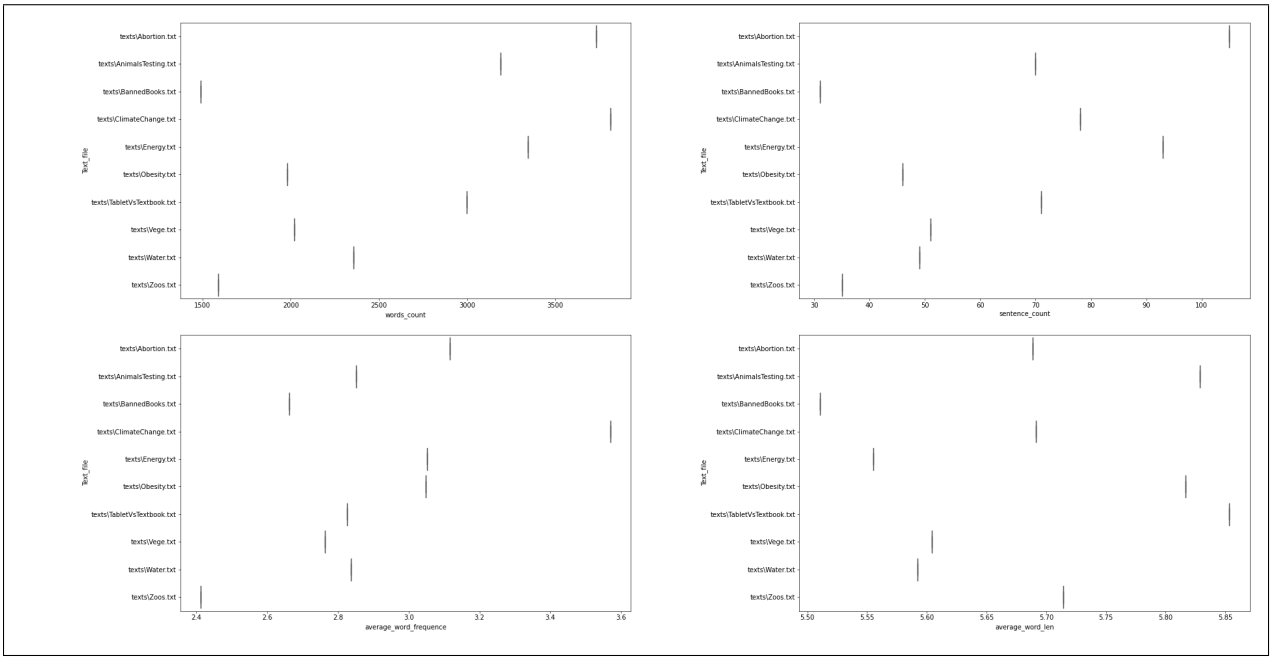


图 A.3：10 个论证主题的复杂性 根据词语总数、句子总数 S,
平均词频 流畅性以及平均词长。

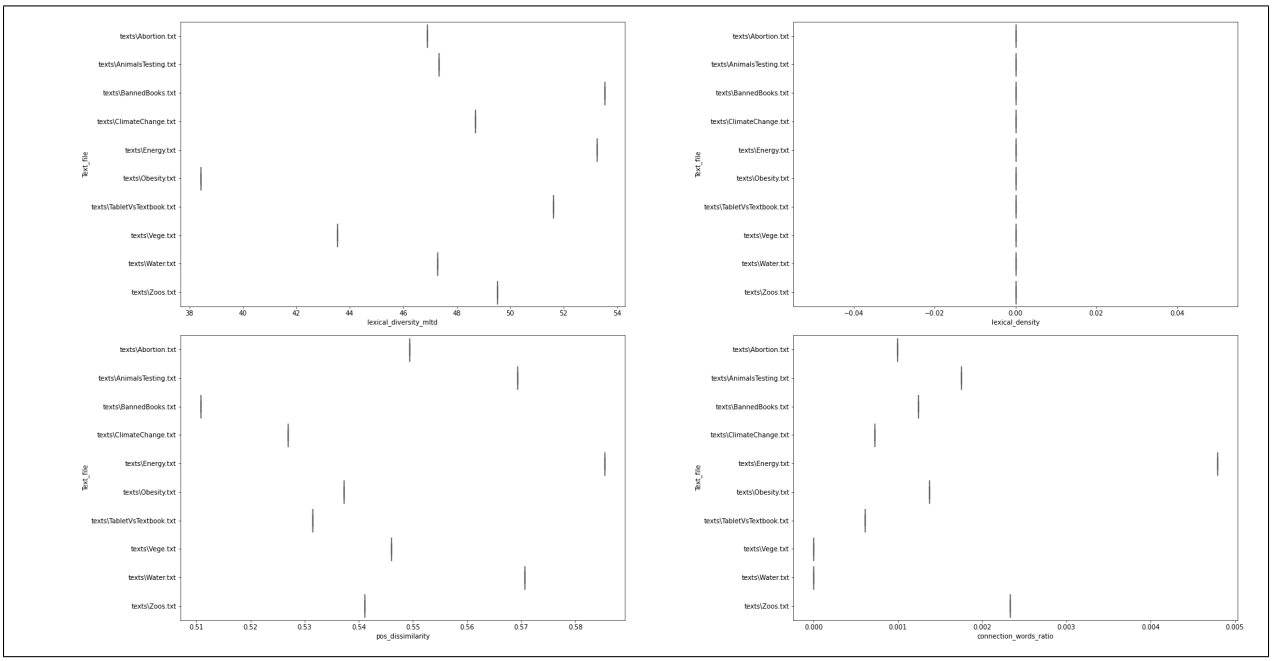


图 A.4：10 个论证主题在词汇多样性、词汇密度、词性差异度和连接词比例方面的复杂度

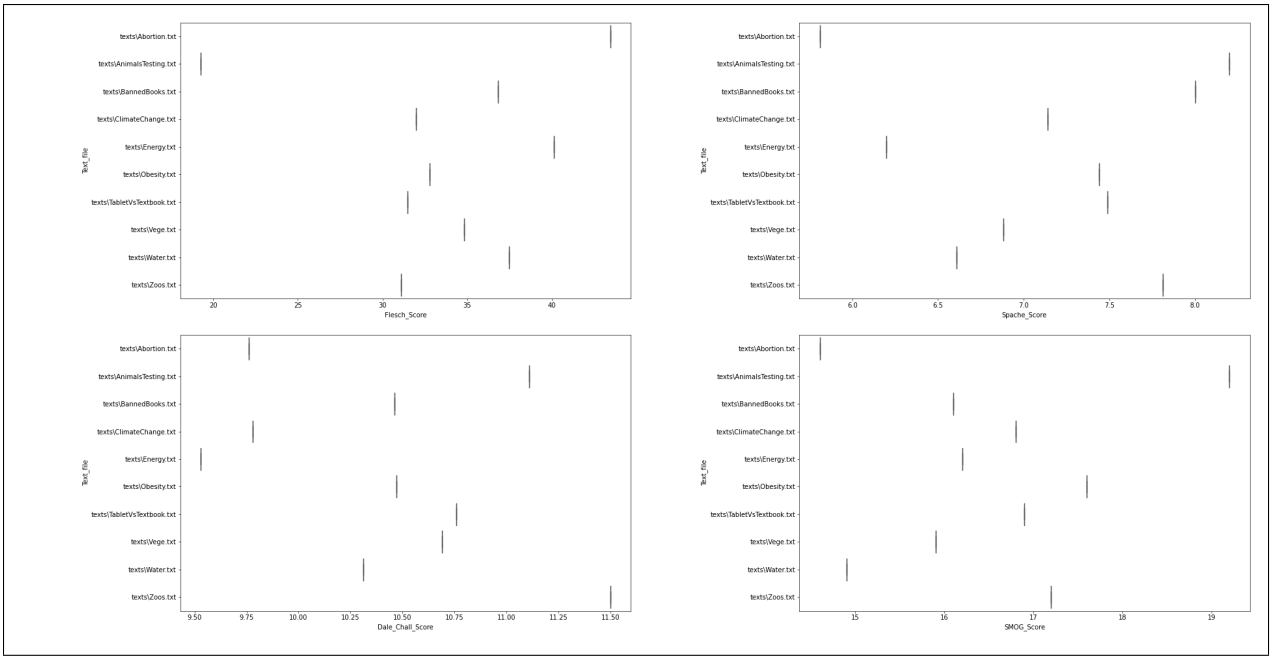


图 A.5：10 个论证主题在弗莱施阅读易读性评分、斯帕奇可读性评分、戴尔-查尔可读性评分和 SMOG可读性分数方面的复杂度。

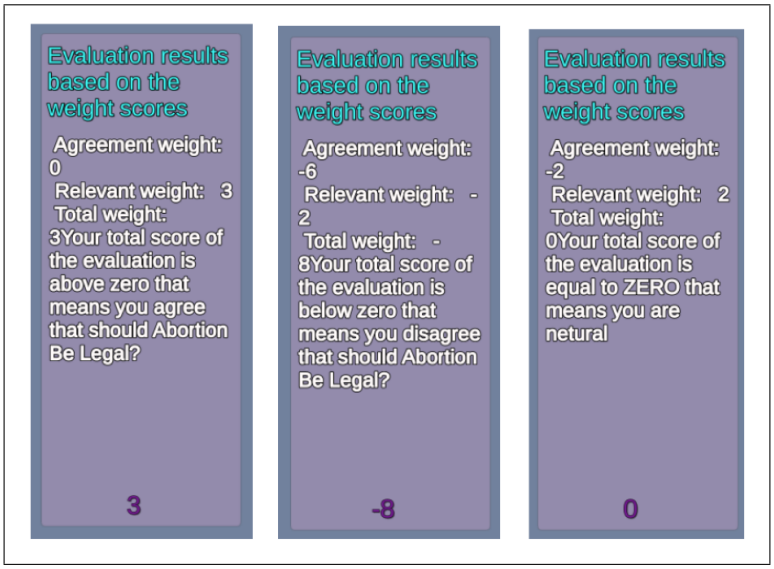


图 A.6：二维 AGM 评估结果消息展示了三个不同的评估句，并根据主要主张的（相关性和一致性）汇总得分显示参与者的决定。详情请参见第 3.1 节。从左至右：参与者对主题表示同意（3）、不同意（-8）和中立。

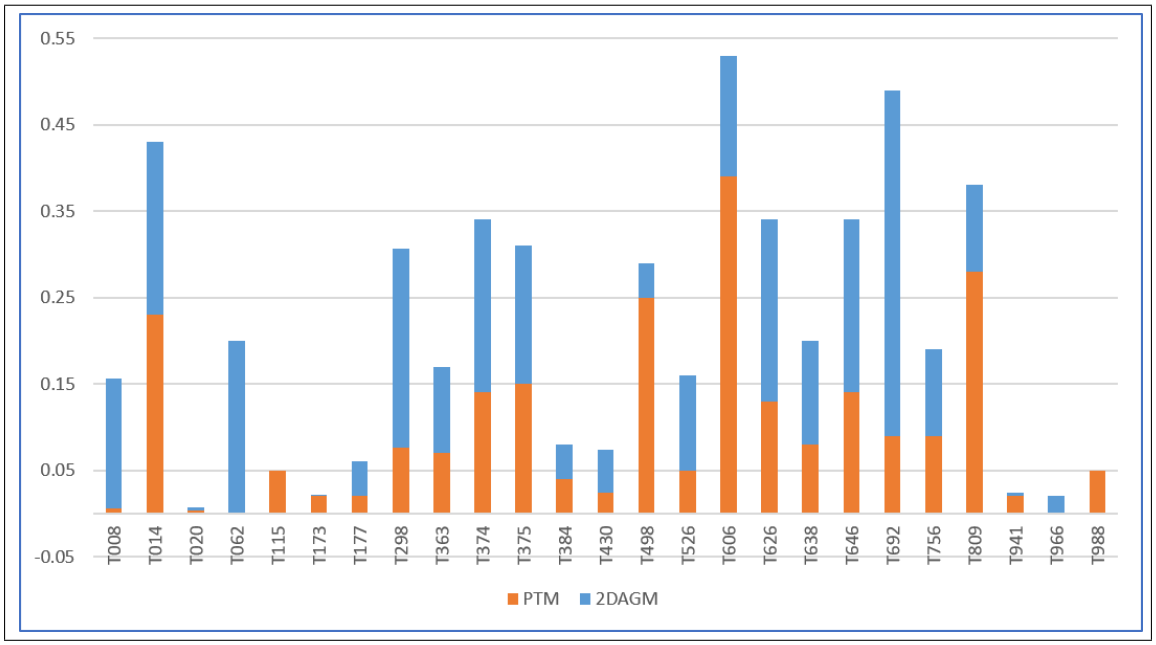


图 A.7：条形图展示了所有参与者的 A-BLEU分数。A-BLEU 分数（聚合双语评估替补分数）表示候选句与参考句的相似程度。A-BLEU 指标范围为 [0, 1], , 分数越接近 1，表示相似度越高。

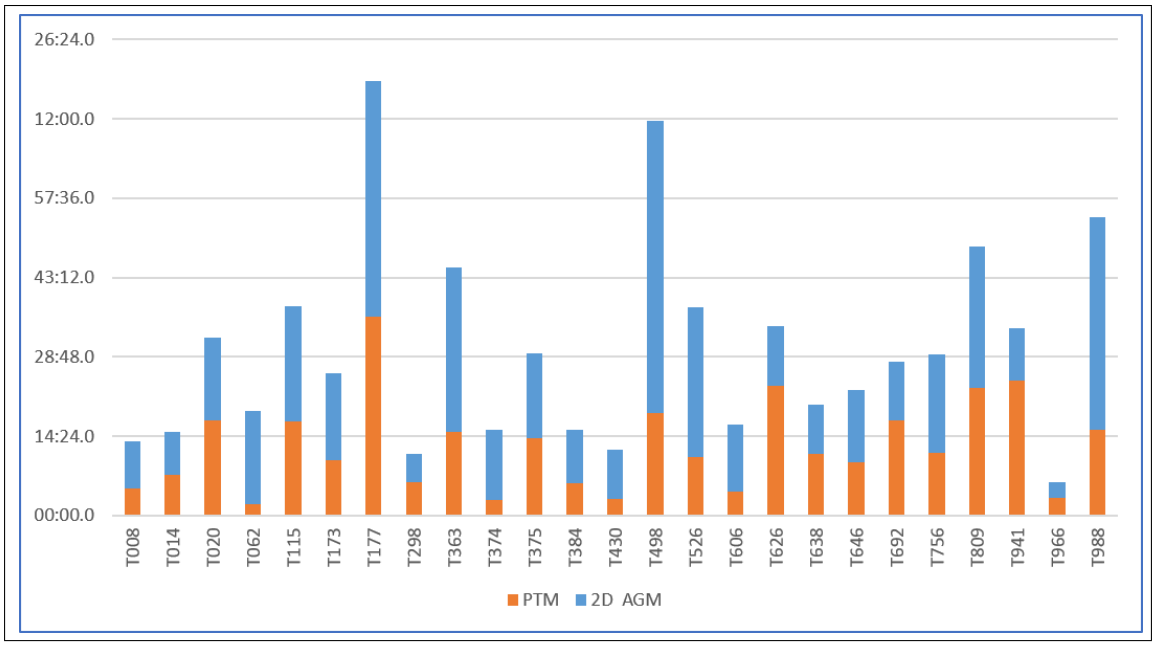
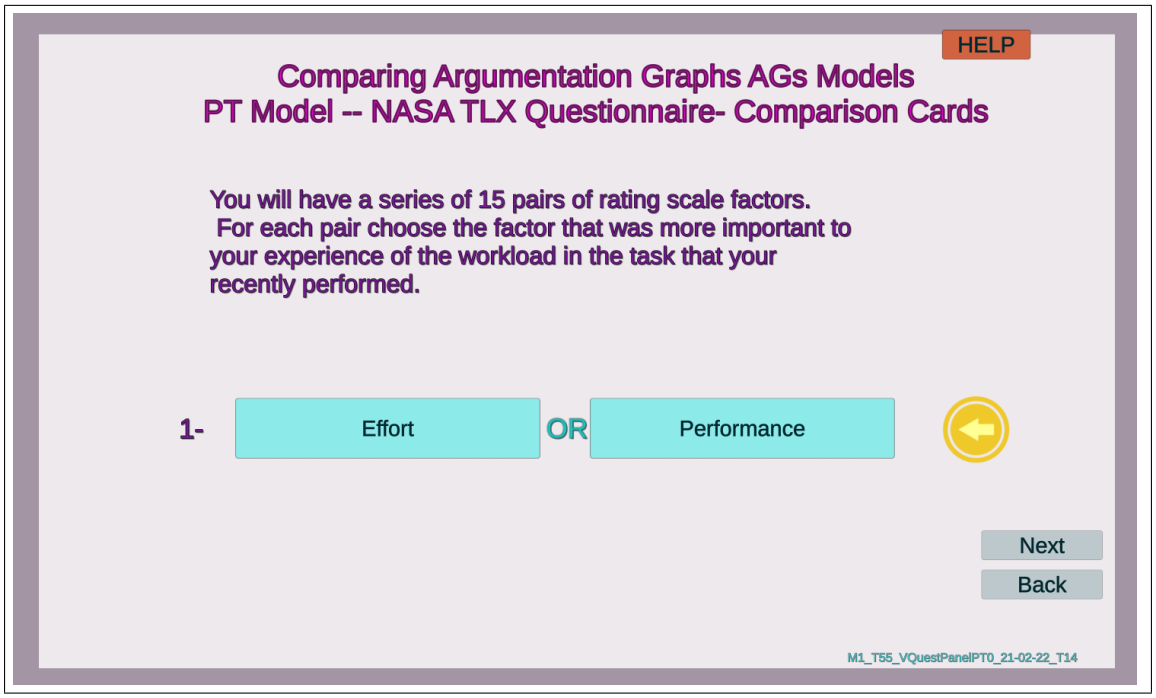
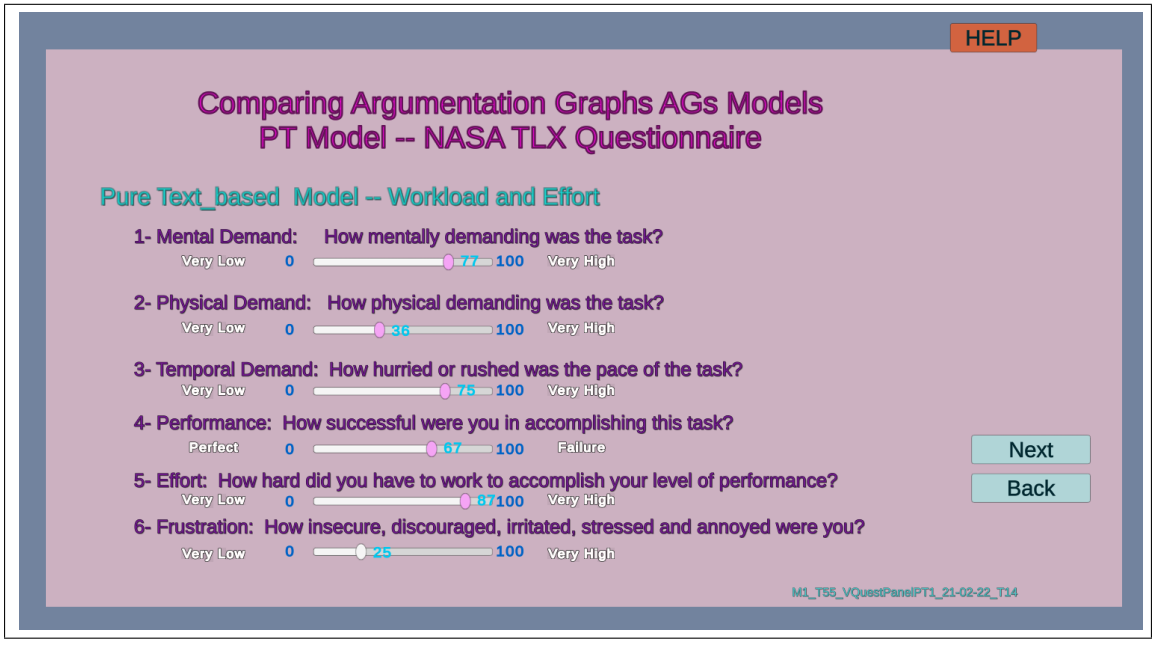


图 A.8：条形图显示参与者所用时间，单位为分和秒。最长和最短用时分别为 PTM（36:04，2:09）和二维论证图模型（52:55，2:56）。每种模型所用的时间均由隐藏计时器通过 C# 代码计算。

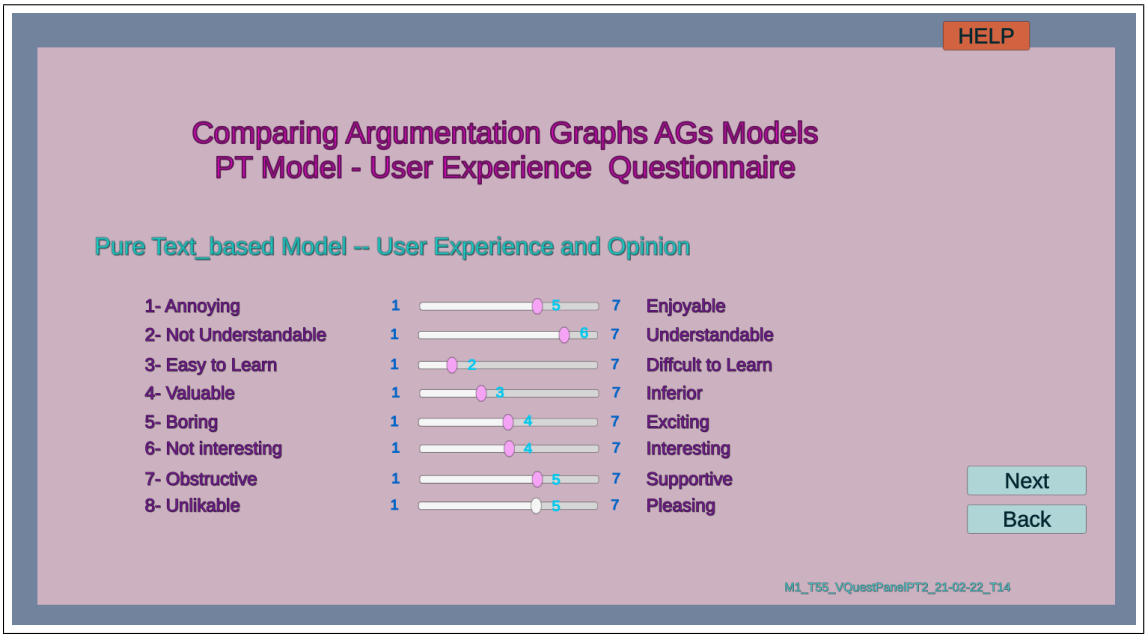


(a) NASA-TLX 问卷第 1 部分

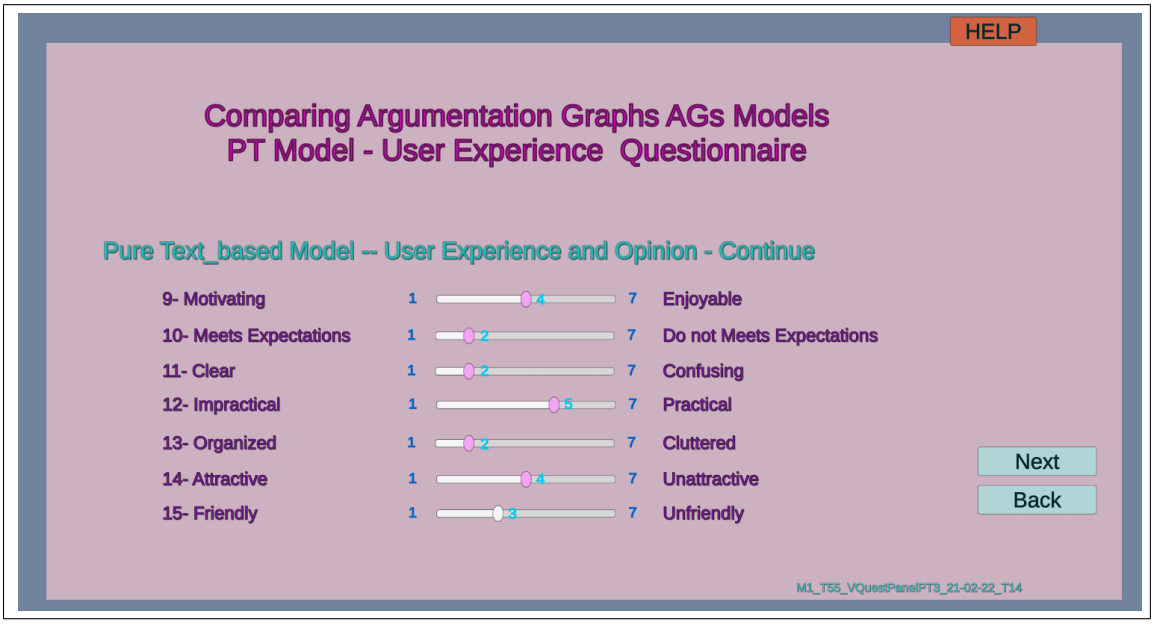


(b) NASA-TLX 问卷第 2 部分

图 A.9：NASA-TLX 问卷的第二部分分为两部分：(a) 考察六个量表因素的每一组成对组合，以确定哪一项与任务中的工作负荷定义更相关。这会让用户进行 15 次成对比较。(b) 按照 5 分的增量，在从极低（0）到极高（100）的量表上为六个因素分别评分。参见附录 1 和第 5.7 节。



(a) UEQ问卷第1–8项



(b) UEQ问卷第9–15项

图A.10：UEQ问卷的两个面板共包含15个项目。这些项目采用1至7的量表，其中1代表最负面的回答，4代表中立回答，7代表最正面的回答。该UEQ包含15个项目，分为5个不同的主要量表：吸引力、清晰度、效率、可靠性和刺激性。详见附录2和第5.5节。

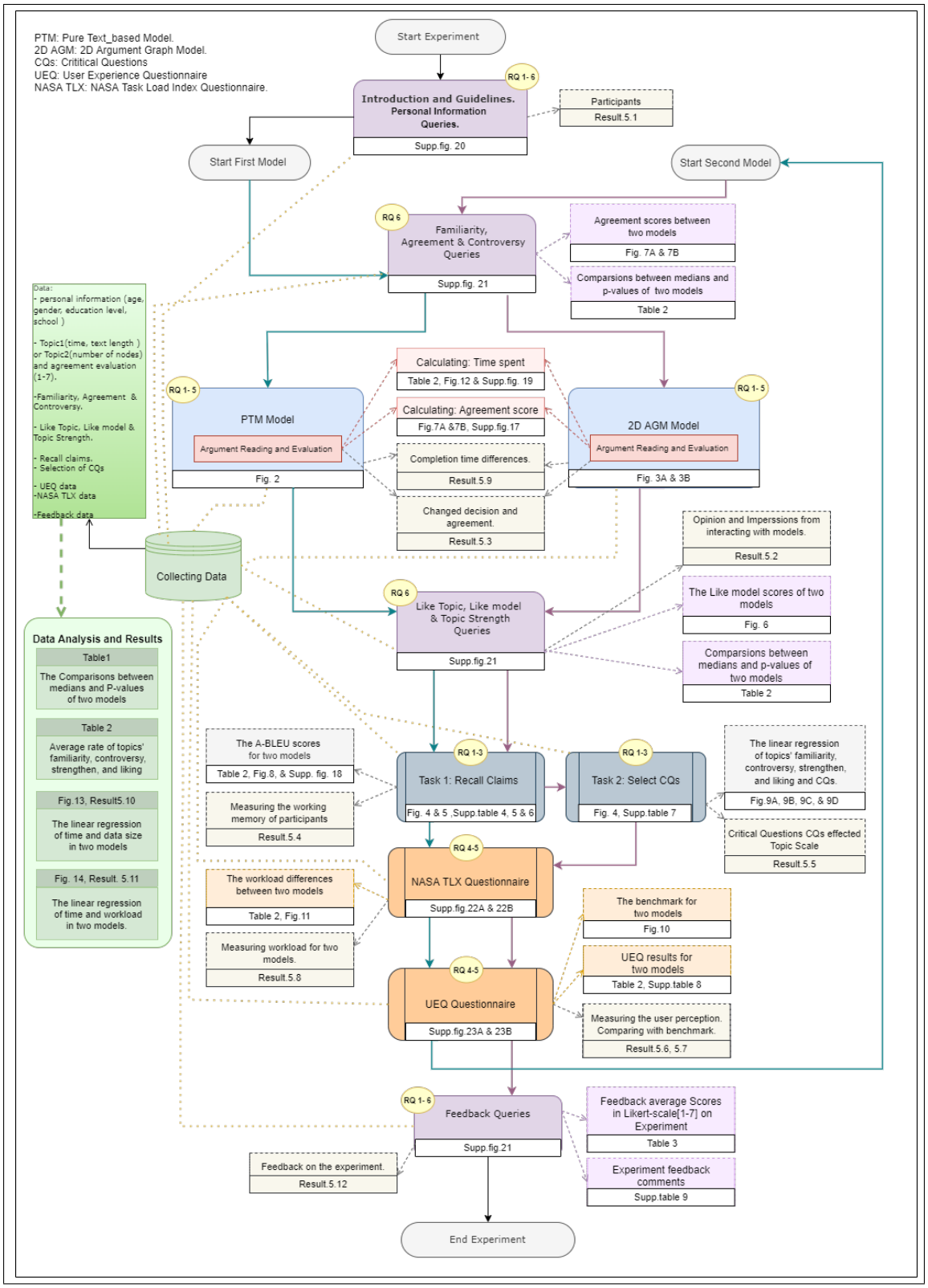


图 A.1 1：实证实验的工作流程包括对图、表和结果的参考 s.