arXiv:2601.05487v1 [cs.MA] 9 Jan 2026

EvidFuse：面向一致性文本–图表数据报告的写作时证据学习

Huanxiang Lin[1]*, QianyueWang[1,2]*,Jinwu Hu[1,2],BailinChen[1,2], Qing Du[1,2]†,Mingkui Tan[1,2]†,

[1]华南理工大学， [2]琶洲实验室，

摘要

数据驱动报告通过将叙述文本与基于底层表格的图表紧密交织，传达有助于决策的洞见。然而，当前基于大语言模型的系统通常采用分阶段流程生成叙述文本和可视化，遵循文本优先、图表其次或图表优先、文本其次的范式。这些设计往往导致图表与文本不一致以及洞见冻结：中间证据空间变得固定，随着叙述文本的发展，模型无法再检索或构建新的视觉证据，最终造成浅层且预先设定的分析。为解决这些局限性，我们提出了**EvidFuse**，一种无需训练的多智能体框架，支持在写作过程中交织生成文本与图表，服务于数据驱动报告。EvidFuse通过两个协作组件将可视化分析与长篇内容起草解耦：一个**数据增强分析智能体**，配备源自探索性数据分析（EDA）的知识并可访问原始数据表；以及一个**实时证据构建写作器**，负责规划大纲并起草报告，同时不时发出细粒度分析请求。这种设计使视觉证据能够恰在叙述文本需要时构建并纳入其中，从而直接约束后续论断，并支持按需拓展证据空间。实验表明，EvidFuse在图表质量、图文对齐和报告层面的实用性方面，在大语言模型评审和人工评估中均排名第一。

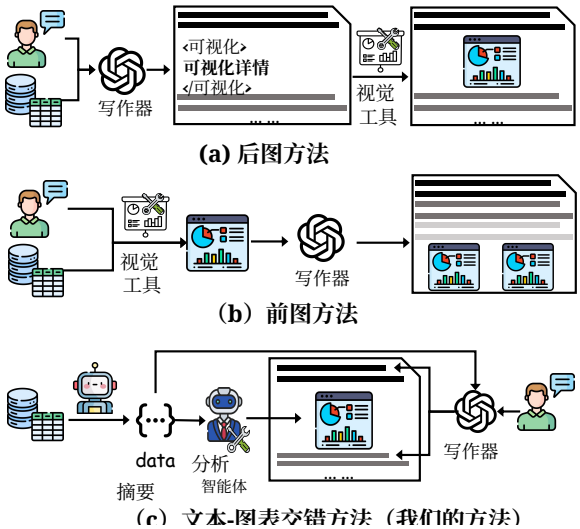


图1：不同图表与文本交错报告生成范式的示意图。

1 引言

数据驱动报告是决策场景中传达复杂数据集的主要媒介，应用范围从公共政策（Stuart，2015年）和科学分析（Erkmen，2023年）到商业智能（Dhanalakshmi 等人，2025年）。然而，手动撰写此类报告成本高昂且耗时，因为分析师必须反复探索数据、设计恰当的可视化，并撰写连贯的叙述，而不同写作器所产出的质量和风格可能存在显著差异（Schneider 等人，2018）。随着大语言模型在长文本生成方面迅速进步（Yang等人，2022年；Wang 等人，2025年），自动化数据驱动报告撰写已日益可行（Jin 等人，2025年；Tan 等人，2025年；Wan 等人，2025年）。

遗憾的是，借助大语言模型生成数据驱动报告仍然面临挑战。这类报告通常是针对涉及多张分散表格的具体分析请求而生成，并以紧密交错的文本与图表形式呈现，其中叙述性主张和可视化相互印证（Yang等人，2025b）。这带来了两项关键挑战：**1）按需生成有依据的可视化：** 报告需要在叙述恰好需要图表时，从多张数据表中检索并构建恰当的图表

*共同贡献：sehuanxianglin@scut.mail.edu.cn, 202420145083@scut.mail.edu.cn
†通讯作者：duqing@scut.edu.cn, mingkui-tan@scut.edu.cn

作为有依据的证据，确保严格的图表–文本一致性（Yang等人，2025b）。**2）面向决策的洞察深度：** 报告需要组织面向决策的多步骤洞见，超越对可视化的表层描述（Islam等人，2024年；Aggarwal等人，2025年）。

现有大多数文本图表报告生成方法采用分阶段流水线，大致可分为两种范式，如图1所示。**文本优先、图表随后**方法首先基于一组固定的推断洞见和用户请求起草叙述，之后才调用外部工具，根据生成文本中嵌入的图表规格创建可视化（Islam等人，2024年；Yang等人，2025b）。由于撰写叙述时无法获取实际的视觉证据，这些方法生成的论断往往缺乏充分依据，或与最终生成的图表不一致。**图表优先、文本随后**方法先根据用户请求预先生成一组候选图表，再以这些图表为条件生成文本（Dibia，2023年；Zhang等人，2025年）。尽管这提升了可视化定制能力，叙述通常缺少具体论断与具体图表之间明确且可验证的对应关系，从而削弱了文本—图表对齐。更根本的是，这两种范式都会冻结中间证据空间：一旦洞见或图表集合固定，模型便无法随着叙述的演进检索或构建新的视觉证据。因此，现有系统往往只能生成表层描述或预定义洞见，难以胜任面向决策的分析。

为解决上述局限性，我们提出 EvidFuse，这是一种免训练多智能体框架，用于在数据驱动报告中进行写作时文本—图表交错生成。具体而言，EvidFuse 由两个协同组件构成，将可视化分析与长篇报告撰写解耦，分别是 **数据增强分析智能体**和**实时证据构建写作器**。数据增强分析智能体具备源自探索性数据分析（EDA）的知识和数据集特定背景信息，包括数据集摘要以及对原始数据表的访问权限。它通过按需构建有依据的可视化及相应说明文字，响应细粒度分析请求。实时证据构建写作器基于多模态大语言模型构建。它首先规划大纲，然后起草报告，并在过程中间歇性地发出分析请求。每当发出请求时，生成就会暂停，并将返回的视觉证据注入上下文，以约束后续论断。随后继续起草，直至生成 <序列结束> 。借助这种写作时证据构建机制，EvidFuse 会在叙述需要时准确插入视觉证据。这一设计提升了图表–文本一致性，并能开展更深入、与决策相关的分析，而不受固定预计算洞察集的限制。我们的主要贡献如下：

- **一种用于交错生成文本与图表的数据报告范式。** 我们指出，文本与图表不一致是分阶段流程造成的，并进一步提出一种交错范式：在撰写上下文时生成可视化证据并将其插入上下文，使后续的上下文论断以有依据的可视化为条件，从而提升文本与图表的一致性。

- **一种支持在写作时构建视觉证据的多智能体协作框架。** 我们提出了 EvidFuse，这是一个由数据增强分析智能体 和实时证据构建写作器组成的协作框架。通过将分析智能体注册为可调用工具，写作器可在生成过程中发出细粒度分析请求，并接收有依据的图表与证据，再将其注入上下文。这种写作时的数据交互将洞察构建与叙事生成紧密结合，从而实现渐进式、面向决策的分析，超越预定义洞察的局限。

- **大语言模型评审和人工评估者在多个质量级别上都更青睐 EvidFuse 生成的报告。** 在三种不同的报告来源以及图表、章节和报告级评估的六项标准中，EvidFuse 在大多数情况下，无论是自动排名还是人工排名，平均表现均为最佳。这表明，与分阶段流程相比，写作时交错生成文本与图表能够产出文本与图表一致性更强、洞见更深入的报告。

2 相关工作

自动化数据驱动报告生成旨在从结构化

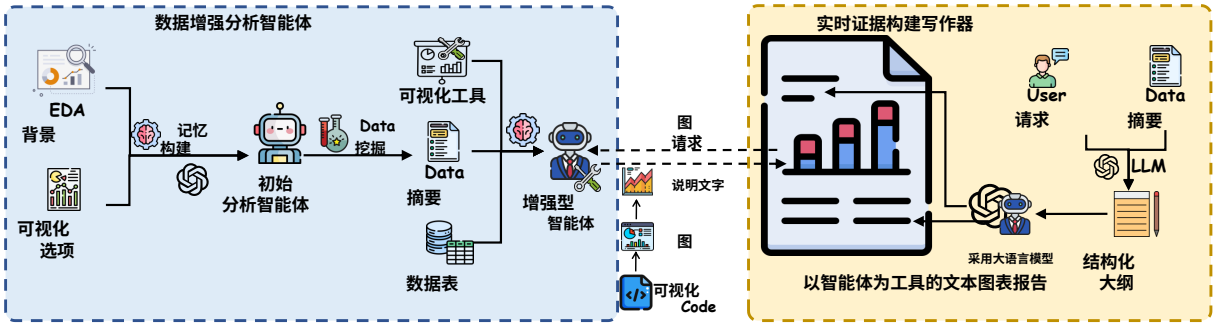


图2：EvidFuse示意图。给定用户分析请求和多个数据表，EvidFuse首先基于EDA衍生的数据集概览和原始数据表，对<style id='1'>数据增强分析智能体进行专门化。随后，**实时证据构建写作器**规划大纲并生成报告。在生成过程中，写作器会因特定的可视化请求而暂停；当其接收到来自分析智能体的可视化结果及基于请求生成的说明文字，并将其作为后续生成的上下文后，便会继续生成，直至 <EOS> 终止生成。

数据集生成分析报告，通过将叙述文本与数据可视化紧密结合起来。现有的基于大语言模型的方法大多遵循分阶段生成流程，根据生成顺序可分为两种范式：**文本优先–图表其次**和**图表优先–文本其次**。

文本优先、图表其次范式。该范式中的方法首先起草叙述（通常以提取出的或推断出的洞见为条件），然后事后插入可视化内容，通常通过占位符或之后再进行渲染的行内可视化说明（Yang等人，2025b；Islam等人，2024年）实现。代表性方法是DataNarrative（Islam等人，2024年），该方法生成完整的叙述，随后将带注释的图表位置替换为可视化结果。该范式有利于长篇叙述，但由于写作时没有实际图表，叙述完成后会导致图表与文本不一致。

图表优先、文本其次范式。图表优先方法根据用户意图和数据预先计算视觉证据（通常是一组图表），然后基于这些图表生成报告（Dibia，2023年；Zhang 等人，2025年；Ma 等人，2023年）。DeepAnalyze（Zhang 等人，2025年）通过在撰写报告之前执行数据分析和可视化， exemplifies 了这一方向。其他用于大语言模型辅助数据探索和可视化的系统，例如InsightPilot（Ma 等人，2023年）和 LIDA（Dibia，2023年），也自然符合这一范式：图表构建阶段将分析意图转换为图表规格，并将生成的图表作为下游叙事生成的证据。这一范式提升了图表的相关性和定制性，但在写作前固定证据空间可能会妨碍迭代式证据寻求，从而限制报告的信息深度。

3 问题设定

我们研究从多个数据表和用户分析请求生成数据驱动的分析报告这一问题。给定输入数据集 D 和人工请求 Request （例如任务描述、分析目标或报告意图），我们的系统会生成文本与图表交错报告。

$$CT = \{t_1, v_1, \ldots, t_n, v_n\},$$

其中，每个 t_i 都是一个文本片段，每个 v_i 都是 D 中的视觉证据。关键挑战在于，如何在特定人工请求下，将可靠的数据分析与连贯的叙事写作相结合。

4 面向一致性文本与图表数据报告的写作时证据构建

我们提出 EvidFuse，这是一种多智能体协作框架，可利用多个数据表生成文本与图表交替呈现的报告。如图 2所示，EvidFuse 由数据增强分析智能体 A' 和报告撰写器 W 组成，通过数据交互实现写作时证据构建。给定数据表 D 和用户请求R_{user}，一个初始的基于多模态大语言模型的分析智能体 A 首先构建数据表概览 DO，然后利用该概览增强 A，使其 A' 能够访问原始数据表。写作器 W 遵循“先规划、后写作”的范式，首先根据 R_{user} 和 DO生成大纲，随后撰写报告的上下文内容。每当需要视觉证据来支持即将提出的论断时， W 便会发出以<可视化>和 </可视化>界定的分析请求，并

暂停生成。作为回应， A' 会查询 D 中的相关数据，并生成有数据依据的可视化结果，再将其注入上下文。随后， W 以返回的可视化结果为条件继续写作。这种交替循环会持续进行，直到 W 生成 <序列结束>。通过在写作时构建视觉证据，并以实际可视化结果为条件生成后续文本，EvidFuse 提升了文本与图表一致性。证据-写作流程的循环能够针对 R_{user}动态生成更深入且与决策相关的洞见，而非停留在表面摘要。

4.1 数据增强分析智能体

根据特定请求，从多个数据表生成可视化是一项并非易事的数据分析任务，需要选择相关变量、确定合适的图表类型和聚合方式，并生成解读（Yang等人，2025b）。虽然将可视化与长篇报告撰写结合在单个模型中会造成上下文冗余并影响任务表现（Chen等人，2025年），但我们将写作与可视化分离，把数据分析和可视化工作交由专门的数据增强分析智能体负责。数据增强分析智能体 A'是一个基于多模态大语言模型的后端，并配备了可调用基于代码的可视化工具的功能， T 该工具结合了特定于数据的背景知识。给定分析请求 q 和数据表 D， A' 会生成可视化结果 v，并 T附上说明文字 g，同时考虑可视化请求。智能体构建和可视化流程的具体步骤如下。

智能体构建。 我们初始化一个基于多模态大语言模型的分析智能体 A，并利用基础探索性数据分析（EDA）知识更新其记忆（Tukey，1977），其中包括分析概念、统计例程、可视化原则以及一组选定的可视化选项。给定一组数据表 $D = \{D_1, \ldots, D_m\}$， A 随后通过针对表格类型进行 5 至 8 项探索性数据分析探查，构建数据概览：

$$DO = A(P_{overview}, D). \tag{1}$$

这种增强通过 DO 为 A' 提供有关数据集的全局上下文，并通过 T赋予其可执行的可视化能力，使其能够基于底层表格处理请求。 $P_{\{\}}$ 是任务提示，与以下提示相同。

可视化流程。 给定任意分析请求 q， A' 会生成可视化 v 和一段面向分析的说明文字 s，通过系统化且迭代的三步流程生成：

1）可视化规范规划。 A' 将分析请求 q 解读为分析意图 i 和全面的可视化规范 vs，包括引用的表子集 D_{ref} （参见附录D中的示例）。此步骤的格式如下：

$$(i, vs) = A'(P_{ana}, q). \tag{2}$$

2）通过 T进行可视化。根据可视化规范 vs， T 将被调用以生成可执行代码，并通过视觉反馈驱动的额外迭代优化进行完善（详情见附录E）。此步骤的格式如下：

$$c = T(D_{\text{ref}},\ vs;\ \text{IterRefine}), \tag{3}$$

其中，IterRefine表示一个迭代循环，该循环根据多模态大语言模型提供的文本反馈评估并更新生成的可视化内容。

3）说明文字生成。 A' 生成面向分析的说明文字 s，直接回应 q，同时严格以 c为依据。此步骤的格式如下：

$$s = A'(P_{caption}, c, i, q) \tag{4}$$

所得的配对结果（c， s）会作为 W全面的视觉证据 v 返回给写作器，以有效地为后续报告生成提供条件。

4.2 实时证据构建写作器

现有方法通常遵循分阶段流水线，其中可视化与叙述文本异步生成，导致文本与图表不一致，并且报告中的洞察深度有限（Zhang 等，2025年；Islam 等，2024年）。为解决这些问题，我们设计了实时数据交互写作者 W，支持在写作时为文本内容构建视觉证据。

W 与数据增强分析智能体 A' 协作，在上下文生成过程中请求、获取并立即纳入视觉证据。具体而言， W 负责大纲规划和长篇叙述，同时将数据交互和可视化工作委托给 A'。每当 W 需要证据来支持后续论断时，就会发出分析请求 q_i，并将其置于<可视化>和 </可视化> 之间，然后暂停生成过程。随后， A' 会返回有依据的可视化图表和说明文字，作为对写作器的上下文增强。 W 则根据返回的视觉证据继续写作。当生成 <序列结束> 时，这一循环便告结束。

通过从异步、分阶段生成转向写作时证据构建，EvidFuse能够依据按需获取的视觉证据生成叙述，从而提升文本与图表一致性，并带来更深入、面向决策的洞见。

具体而言，报告撰写器 W 是一个基于多模态大语言模型的智能体。它遵循规划-撰写范式（Yao 等人，2019年；Wang 等人，2025年），首先根据用户请求和数据概览生成大纲：

$$O = W(P_{outline}, R_{user}, DO) \tag{5}$$

在每一步中，写作器都会生成一段文本，其中包含由<visualization> 和 </visualization>限定的分析请求：

$$x_i = \{t_i,\ q_i\} = W(P_{report}, H_{i-1}, O, DO), \tag{6}$$

其中，t_i 表示生成的文本内容，q_i 则明确由 <visualization>和 </visualization> 标签界定，这会触发写作器暂停生成。随后，我们调用数据增强分析智能体 A' 来构建相应的视觉证据：

$$v_i = (c_i, s_i) = A'(q_i, D, DO;\ T), \tag{7}$$

历史记录通过生成的内容进行更新：

$$H_i = H_{i-1} \oplus v_i. \tag{8}$$

W 写作器会基于更新后的历史记录继续写作 H_i。这一交错过程会重复进行，直到 W 输出 <序列结束>。最终报告 CT 就是终止后的历史记录 H。完整过程如算法1所示。

算法1 EvidFuse 的文本与图表交错报告生成

要求： 用户请求 R_{user}，表格 D，initialized 分析智能体 A，可视化工具 T，写作器 W，停止令牌集合 $Stop$ = {<序列结束>, </可视化>}

初始化文本与图表交错报告 $CT_0 \leftarrow \emptyset$; $i \leftarrow 0$
获取数据概览 DO，通过公式1。
将 A 的记忆更新为 A'，使用 DO 和 D
获取报告大纲 O，通过公式5
已完成 ←**false**
生成步骤 $i = 0$
当finished 为**false 时**
    通过方程 $x_i = t_i,\ q_i$ 获取步骤内容 6。
    $CT_i \leftarrow CT_{i-1} \cup \{t_i\}$
    如果 < /在 x_i 中的视觉可视化 >**则**
        从 x_i获取 q_i
        将图表 vc_i 和 s_i 作为可视化获取证据 v_i 通过方程 7
        $CT_i \leftarrow CT_i \cup \{v_i\}$
    否则如果 <序列结束> 在 t_i **那么**
        已完成 ← **真**
    $i \leftarrow i + 1$
返回 CT

5 实验

5.1 实验设置

数据集。 我们整理了一个基准测试集，包含来自三个主流报告来源的 表格分析请求 配对数据，用于评估 EvidFuse。具体而言，我们从每个来源抽取 20 份报告，分别来自 Tableau Public Stories（Tableau）、Our World in Data（OWID）和 USAFacts（USAFacts），共涵盖 18 个主题。对于每份报告，我们提取其底层数据表，并将原始报告标题作为分析请求。有关所收集实例的更多详细统计数据，请参见附录 B.1。

基线方法。 我们将其与采用不同生成范式的方法进行比较，这些方法仅通过大语言模型生成文本与图表交错的报告。

本身。对于直接生成，我们提示大语言模型通过单次生成完整报告，其中<可视化> 规范会被渲染为图表。此外，我们还纳入了 DataNarrative（Islam 等人，2024年），它是文本优先、图表其次范式的代表性方法；以及 DeepAnalyze（Zhang 等人，2025年），它是先图后文范式的代表性方法。为简洁起见，我们将大语言模型直接生成的方法称为**直接**。将 DataNarrative 称为**DN.**，将 DeepAnalyze 称为**DA.**。更多详情见附录B.2。

指标。 鉴于生成的报告具有多模态特性，我们从三个级别评估其质量：(i) 图表，(ii)章节对，以及 (iii)报告。对于每个级别，我们采用两项标准，更多详情见附录 B.3。概述如下：

(i) 图表层级。

布局合理性（布局）：图表布局是否有效利用空间。

可读性（阅读）：视觉元素是否清晰明确、易于阅读。

（ii）章节级别。

文本与图表一致性（T-C 一致性）：配套文本是否得到图表的支持。

文本信息深度（深度）：是否

表1：Qwen3-235B-A22B-Instruct-2507 和 Qwen2.5-VL-72B-Instruct 下的自动排名结果。

数据集	级别	指标	直接	DN.	DA.	Ours
Tableau	图表	阅读	3.10	1.85	3.60	**1.45**
		布局	2.95	**1.35**	3.90	1.80
	章节	T-C 一致性	1.95	2.35	4.00	**1.70**
		深度	2.15	2.30	4.00	**1.55**
	报告	信息	2.90	1.75	3.75	**1.60**
		视觉一致性	2.55	2.60	3.05	**1.80**
OurWorld InData	图表	阅读。	2.85	2.12	3.70	**1.30**
		布局。	2.85	1.75	3.90	**1.50**
	章节	T-C 一致性	1.95	2.35	4.00	**1.70**
		深度。	2.20	2.40	4.00	**1.40**
	报告	信息	2.75	2.25	3.50	**1.50**
		视觉一致性	2.05	2.85	3.45	**1.65**
USAFact	图表	阅读	2.85	1.90	3.85	**1.40**
		布局	2.80	1.65	3.95	**1.60**
	章节	T-C 一致性	2.10	2.05	4.00	**1.85**
		深度。	2.80	2.10	4.00	**1.10**
	报告	信息。	2.90	2.25	3.45	**1.40**
		视觉一致性。	2.25	2.65	3.20	**1.70**

文本是否超越表面描述，提供有意义且符合意图的分析洞见。

(iii) 报告层级。

信息丰富度（信息）：报告是否以全面且有用的内容充分回应了所要求的分析。

视觉一致性（视觉一致性）：报告中的所有图表在视觉风格上是否一致。

评估策略。 我们采用大语言模型评审（基于 GPT-4.1（OpenAI 等人，2024年））和人工评估者来比较各种方法。我们不进行绝对评分，而是采用 **基于排名的** 策略，以减少校准和稳定性问题（Wang 等人，2025年）。对于每个样本，评估人员根据相对偏好对输出进行排名，我们将**平均排名结果**作为性能衡量指标。更多详情见附录 B.3。

EvidFuse 的实现。 EvidFuse 被实现为一种智能体框架，其中数据增强分析智能体和在写作时构建证据的写作器，基于多模态大语言模型，并各自使用独立记忆。大纲生成等部分任务则由具有相同上下文记忆的大语言模型替代。可视化工具 T 是一个由大语言模型或多模态大语言模型担任代码生成器的程序（详情见附录E）。遵循Yang等人的设置，我们的实验采用两种模型配置：（1）小规模：仅使用 Qwen3-VL-32B-Instruct（Bai等人，2025a）作为基础模型；（2）大型混合规模：使用 Qwen3-235B-A22B-Instruct-2507（Yang等人，2025a）进行文本生成，并使用 Qwen2.5-VL-72B-Instruct（Bai等人，2025b）进行

表 2：Qwen3-VL- 32B-Instruct 下的自动排名结果。

数据集	级别	指标	直接	DN.	DA.	Ours
Tableau	图表	阅读	3.05	**1.45**	3.65	1.85
		布局。	2.80	**1.45**	3.85	1.90
	章节	T-C 一致性	2.05	2.70	4.00	**1.25**
		深度。	2.35	2.45	4.00	**1.20**
	报告	信息。	3.05	1.95	3.55	**1.45**
		视觉一致性	2.70	2.50	3.10	**1.70**
OurWorld InData	图表	阅读	3.30	1.80	3.40	**1.50**
		布局	3.25	**1.40**	3.60	1.70
	章节	T-C 一致性	2.20	2.75	4.00	**1.05**
		深度	2.55	2.20	4.00	**1.25**
	报告	信息	3.10	2.20	3.35	**1.35**
		视觉一致性	2.55	2.85	2.70	**1.50**
USAFact	图表	阅读。	3.30	1.80	3.40	**1.50**
		布局。	2.95	**1.30**	3.65	2.10
	章节	T-C 一致性	2.65	2.20	4.00	**1.15**
		深度。	2.50	2.40	4.00	**1.10**
	报告	信息。	3.50	1.75	3.25	**1.50**
		视觉一致性。	2.50	2.50	3.15	**1.65**

多模态理解。为确保公平比较，EvidFuse 和基线方法的所有设置均保持一致。更多详情见附录B.2。

5.2 自动报告评估

EvidFuse 在从图表质量到整体报告质量的各个层面均展现出一致且多层次的优势。我们在表1中报告了 GPT-4.1 在所有指标上的排名结果。EvidFuse 在所有数据集和所有评估级别上均优于所有基线方法。分阶段的基线方法 **DN.** 和**DA.** 在图表-文本一致性（**T-C 一致性**）和信息深度（**深度。**）方面的排名较差，表明它们生成的叙述文本通常缺乏最终视觉证据的充分支撑，并且往往停留在表层描述层面。我们将这些优势归因于以下因素：(1) 将复杂的可视化/分析工作负载解耦至两个智能体中，避免生成过程中的冗余提示和支撑不足；(2) 采用大纲引导的写作时交错流程，按需发出可视化请求，并将返回的图表注入生成的上下文中，从而确保后续论断受到具体视觉证据的直接约束。

EvidFuse 在不同基础模型上始终表现出稳定的效果，表明其具备跨模型配置的可扩展性。如表1 和表2所示，EvidFuse 在两种模型配置和多层级指标下均持续优于所有基线方法，这表明其性能提升并不依赖于特定的骨干网络选择。我们将这种跨模型稳健性归因于我们的无需训练的框架设计，以及各模块中所采用的

表3：使用 Qwen3-235B-A22B-Instruct-2507 和 Qwen2.5-VL-72B-Instruct 作为基础模型时的人工排名结果。

数据集	级别	指标	直接	DN.	DA.	Ours
Tableau	图表	阅读。	3.10	2.27	3.40	**1.23**
		布局。	2.93	2.23	3.53	**1.30**
	章节	T-C 一致性	2.23	2.70	4.00	**1.07**
		深度。	3.20	2.30	3.37	**1.13**
	报告	信息	3.60	2.10	3.07	**1.23**
		视觉一致性	2.70	2.70	3.20	**1.40**
OurWorld InData	图表	阅读。	3.23	2.27	3.17	**1.33**
		布局。	3.07	2.20	3.50	**1.23**
	章节	T-C 一致性	2.40	2.60	4.00	**1.00**
		深度。	2.80	2.60	3.57	**1.03**
	报告	信息	3.33	2.23	3.20	**1.23**
		视觉一致性	2.77	2.53	3.23	**1.47**
USAFacts	图表	阅读	3.20	2.17	3.47	**1.17**
		布局	2.83	2.23	3.67	**1.27**
	章节	T-C 一致性	2.37	2.70	3.87	**1.07**
		深度。	2.90	2.33	3.70	**1.07**
	报告	信息。	3.40	2.07	3.33	**1.20**
		视觉一致性。	2.57	2.80	3.43	**1.20**

模型无关提示词，这使得整个流程对骨干网络容量不那么敏感，同时在不同模型间保留相同的分解与 grounding 行为。

5.3 人工评估

人类偏好与大语言模型评审的排名结果一致。比较表格1 与表格 3以及表格 2与表格 4，在相同模型设置下，大语言模型评审与人工评估人员得出的排名结果显示，EvidFuse 在大多数情况下于两种模型配置下均排名第一，而 DataNarrative 和 DeepAnalyze 紧随其后，唯一例外是 Tableau 上的报告层级结果。在较大的混合设置下（表格 3），EvidFuse 在三个数据源和六项指标中均取得接近顶尖的排名，其中包括较强的章节级文本-图表一致性与深度，以及报告层级的信息丰富度和视觉一致性。在较小的模型设置下（表格 4），也呈现出类似模式，EvidFuse 再次在大多数指标上排名第一。唯一值得注意的偏差是 报告层级 的 Tableau 结果：在信息丰富度和视觉一致性方面，DataNarrative 的排名高于 EvidFuse。总体而言，高度的定性一致性支持将大语言模型评审作为人工偏好可扩展代理的合理性，可用于多层级报告评估；同时，这也进一步增强了 EvidFuse 所报告的多层级优势的可信度，使其不再局限于单一评估人员类型。

EvidFuse 在人工评估中生成的报告具有更高的实际效用。从表格3 和表格4可以看出，EvidFuse 在

表4：将 Qwen3-VL-32B-Instruct 用作基础模型时的人工排名结果。

数据集	级别	指标	直接	DN.	DA.	Ours
Tableau	图表	阅读	3.60	2.00	2.50	**1.90**
		布局	3.80	1.90	2.60	**1.70**
	章节	T-C 一致性	3.70	2.10	2.40	**1.80**
		深度	3.90	2.30	2.30	**1.50**
	报告	信息	3.90	**1.60**	2.40	2.10
		视觉一致性	3.60	**1.40**	2.60	2.40
OurWorld InData	图表	阅读。	3.70	2.70	2.10	**1.50**
		布局。	3.90	2.30	2.20	**1.60**
	章节	T-C 一致性	3.00	2.80	2.90	**1.30**
		深度。	3.90	2.30	2.10	**1.70**
	报告	信息	3.80	2.30	2.00	**1.90**
		视觉一致性	3.40	2.30	2.30	**2.00**
USAFacts	图表	阅读	3.50	2.70	2.30	**1.80**
		布局	3.70	2.20	**2.00**	2.30
	章节	T-C 一致性	3.60	2.40	**1.80**	**1.80**
		深度	3.90	2.50	2.20	**1.80**
	报告	信息	3.90	1.90	2.10	**1.50**
		视觉一致性	3.60	1.90	3.00	**1.50**

表5：使用 Qwen3-VL-32B-Instruct 对 EvidFuse 进行的消融研究。

数据集	级别	指标	直接	不使用视觉	不使用写作器	我们的
Tableau	图表	阅读。	3.60	2.00	2.50	**1.90**
		布局。	3.80	1.90	2.60	**1.70**
	章节	T-C 一致性	3.70	2.10	2.40	**1.80**
		深度。	3.90	2.30	2.30	**1.50**
	报告	信息	3.90	2.10	2.40	**1.60**
		视觉一致性	3.60	2.40	2.60	**1.40**
OurWorld InData	图表	可读性。	3.70	2.70	2.10	**1.50**
		布局。	3.90	2.30	2.20	**1.60**
	章节	T-C 一致性。	3.00	2.80	2.90	**1.30**
		深度。	3.90	2.30	2.10	**1.70**
	报告	信息	3.80	2.30	2.00	**1.90**
		视觉一致性	3.40	2.30	2.30	**2.00**
USAFacts	图表	阅读	3.50	2.70	2.00	**1.80**
		布局	3.70	2.20	2.30	**1.80**
	章节	T-CCons.	3.60	2.40	2.20	**1.80**
		深度	3.90	2.50	2.10	**1.50**
	报告	信息	3.90	1.90	2.60	**1.60**
		视觉一致性	3.60	1.90	3.00	**1.50**

在不同人工评估者评估的大多数案例中排名第一。在 OurWorldInData 上，EvidFuse 在文本—图表一致性和信息丰富度方面的排名分别为 1.30 和 1.90，而直接使用基础模型的排名分别为 3.00 和 3.80。这些优势表明，EvidFuse 能够生成信息量更丰富的数据驱动报告，并为实际阅读的读者提供高质量的可视化。

5.4 消融研究

数据增强分析智能体的有效性。我们通过移除 A'，并强制写作器在相同输入下将可视化规划与执行和长文本生成结合起来，来对数据增强分析智能体进行消融（R_{user}和 DO）。如表5所示，该变体使……下降

表6：使用 Qwen3-235B-A22B-Instruct-2507 和 Qwen2.5-VL- 72B-Instruct 作为基础模型时的报告级统计数据。Fig. 表示每份报告中的图表平均数量，Words. 表示每份报告的平均词元数，Content. 衡量报告文本中的信息量。

数据集	指标	直接	DN.	DA.	Ours
Tableau	Fig.	3.85	6.10	4.05	**6.90**
	词数	879.60	1203.00	1852.802	
	内容	16.35	12.60	15.60	**18.85**
OurWorldInData	Fig.	3.70	7.20	4.15	**8.20**
	词数	1037.10	1667.65	1779.902	
	内容	14.95	16.70	15.35	**18.40**
USAFact	Fig.	3.55	5.95	3.10	**6.55**
	词数。	924.30	1695.90	1755.1023	
	内容。	15.65	15.00	15.90	**20.40**

降低了图表级质量和基于图表的推理表现，其中**阅读**和**深度**的降幅明显。例如，在 USAFacts 上，移除 A' 后，**阅读**得分从 1.8 恶化至 2.7，**深度**得分从 1.5 恶化至 2.5。这些结果凸显了将复杂的可视化决策与叙事生成解耦的益处。专门的分析模块对于可靠地选择变量／聚合方式并生成有依据的视觉证据至关重要，这使写作器能够专注于连贯的叙述，并产出更深入、由图表支持的洞见。

写作过程中实时数据交互的有效性。我们通过禁用实时交互来进行写作时证据构建的消融实验，将 EvidFuse 简化为文本优先–图优先的流水线：写作器首先生成完整叙述，并延迟生成可视化占位符；只有在文本定稿后，才调用分析智能体生成相应的图表。如表格5所示，这一消融主要损害了报告的图文对齐和文本分析深度。在 Tableau 上，移除交互机制后，T-C 一致性的排名从 1.8 恶化至 2.4，文本深度也从 1.5 降至 2.3。这表明，写作过程中通过数据交互进行写作时视觉证据构建，对于持续以新生成的可视化图表为论断提供依据至关重要，使写作器能够迭代请求针对性证据、纠正不匹配，并生成信息丰富的内容。

5.5 更多讨论

EvidFuse 生成了视觉内容丰富的长篇报告。根据表格6和表格7，EvidFuse 在所有数据集上生成的可视化内容最多、报告也最长。在 Tableau 上，EvidFuse 生成的报告平

表7：将 Qwen3- VL-32B-Instruct 作为基础模型时的报告级统计数据。其他详细信息与表6相同。

数据集	指标	直接	DN.	DA.	Ours
Tableau	Fig.	3.00	7.60	4.05	**8.30**
	词数	1154.25	1694.95	2004.45	
	内容	11.85	16.45	15.60	**18.85**
OurWorldInData	Fig.	3.05	6.55	4.15	**8.80**
	词数	1539.10	2542.20	1779.90	
	内容	14.40	13.80	15.35	**18.00**
USAFact	Fig.	2.55	5.90	3.10	**6.55**
	词数	1117.35	1593.60	1755.10	
	内容	14.85	16.45	15.90	**16.85**

均包含 8.30 个可视化内容，平均为 4071.05 个词元；而 DataNarrative 生成的报告平均包含 7.60 个可视化内容和 1694.95 个词元。这些结果表明，实时数据交互能够在写作支持过程中持续获取证据，从而支持更全面地生成可视化内容和上下文内容。

EvidFuse 的报告包含的信息最多。从表格6 和表格7可以看出，除可视化和长度外，EvidFuse 的报告在所有数据集上所包含的信息都多于所有基线方法。在采用相同的信息提取设置的情况下（提取详情见附录G ），EvidFuse 的内容得分最高。在 OurWorldInData 上，EvidFuse 包含 18 条信息，而 Direct 包含 14.40 条。这表明，额外的图表和词元转化成了内容更密集、信息更丰富且面向决策的分析。

6 结论

本文将数据驱动的报告生成重新定义为写作过程中的证据构建问题：分阶段的文本–图表流水线（文本优先或图表优先）往往会“冻结”中间洞见和图表，导致叙述展开时难以维持图表–文本的一致性。我们提出 EvidFuse，这是一种无需训练的文本–图表交错框架，将数据增强分析智能体与实时交互式写作器相结合，使模型能够在起草过程中恰逢其需地请求并注入有依据的视觉证据。在多个多表基准测试集上的实验表明，这种交错范式生成的报告具有更高的图表质量、更紧密的图表–文本对齐，以及更高的整体实用性。这表明，对于有依据的长文本生成而言，一种通用范式是将证据视为一等且动态构建的上下文。

参考文献

Aniya Aggarwal、Ankush Gupta、Shivangi Bithel 和 Arvind Agarwal。2025年。目标驱动的数据故事、叙述与解释。载于*2025*年美洲国家计算语言学协会分会会议论文集：人类语言技术（第*3*卷：工业赛道），第684–694页，美国新墨西哥州阿尔伯克基。计算语言学协会。

白帅、蔡宇轩、陈睿哲、陈克勤、陈雄辉、程泽森、邓良浩、丁伟、Chang Gao、葛春江、葛文斌、Zhi- fang Guo、黄启东、黄杰、Fei Huang、Binyuan Hui、姜书彤、李照海、李明生等45人。2025a。Qwen3-vl 技术报告。*arXiv* 预印本 *arXiv:2511.21631*。

白帅、陈克勤、刘雪静、王嘉林、Wen- bin Ge、宋思博、党凯、王鹏、Shi- jie Wang、唐军、钟虎门、朱元志、杨明坤、李照海、万建强、王鹏飞、丁伟、傅哲仁、徐逸恒等8人。2025b。Qwen2.5-vl 技术报告。*arXiv* 预印本 *arXiv:2502.13923*。

陈耀佛、游增、张书海、李浩坤、李奕睿、王耀威和Mingkui Tan。2025年。用于长上下文语言建模的核心上下文感知变换器。载于第四十二届国际机器学习会议。

G. Dhanalakshmi、Chanigaram Kalyani 和 Kareema Shaik。2025年。数据驱动决策：商业智能的力量。《国际工程科学与先进技术期刊》。

Victor Dibia。2023年。LIDA：利用大语言模型自动生成语法无关的可视化和信息图。载于计算语言学协会第*61*届年会论文集（第*3*卷：系统演示），第113–126页，加拿大多伦多。计算语言学协会。

Özlem Erkmen。2023年。数据新闻：系统性文献综述。《新闻学研究》，25：58–79。

Mohammad Javad Hosseini、Yang Gao、Tim Baumgärt- ner、Alex Fabrikant 和 Reinald Ki m Am- playo。2024年。可扩展且领域通用的抽象命题分段。预印本，arXiv:2406.19803。

Mohammed Saidul Islam、Md Tahmid Rahman Laskar、Md Rizwan Parvez、Enamul Hoque 和 Shafiq Joty。2024年。DataNarrative：利用可视化和文本自动生成数据驱动的故事。载于*2024*年自然语言处理实证方法会议论文集，第19253–19286页，迈阿密，佛罗里达州，美国。计算语言学协会。

Song Jin、Shuqi Li、Shukun Zhang 和 Rui Yan。2025年。Finrpt：数据集、评估系统与基于大语言模型的多智能体框架，用于公平研究报告生成。预印本，arXiv:2511.07322。

Pingchuan Ma、Rui Ding、Shuai Wang、Shi Han 和 Dongmei Zhang。2023年。Insightpilot：一种由大语言模型赋能的自动化数据探索系统。载于*EMNLP*（演示），第346–352页。计算语言学协会。

OpenAI，：，Aaron Hurst、Adam Lerer、AdamP. Goucher、Adam Perelman、Aditya Ramesh、Aidan Clark、AJ Ostrow、Akila Welihinda、Alan Hayes、Alec Radford、Aleksander M.adry、Alex Baker-Whitcomb、Alex Beutel、Alex Borzunov、Alex Carney、Alex Chow、Alex Kirillov等401人。2024年。GPT-4o 系统卡片。预印本，arXiv:2410.21276。

OpenRouter。 Openrouter。

OWID。Our World in Data。https://ourworldindata.org。

W Joel Schneider、Elizabeth O Lichtenberger、Nancy Mather和NadeenL Kaufman。2018年。评估报告撰写要点。John Wiley & Sons。

David Stuart。2015年。数据革命：大数据、开放数据、数据基础设施及其影响。在线信息评论，39(2):272。

Tableau。Tableau Public 故事。https://public.tableau.com/app/search/vizzes/Story。

Mingrui Tan、Yang Liu、Kun Gao、Fei Gao 和 Yuting Song。2025年。用于投资指导的多智能体协作：Earnings2Insights 报告生成。载于第十届金融科技与自然语言处理研讨会论文集，第328–334 页，中国苏州。计算语言学协会。

John W. Tukey。1977年。探索性数据分析。Addison-Wesley Publishing Company，雷丁，马萨诸塞州。

USAFacts。USAFacts。https://usafacts.org。

Zhongwei Wan、Che Liu、Xin Wang、Chaofan Tao、Hui Shen、JingXiong、Rossella Arcucci、Huaxiu Yao 和 Mi Zhang。2025年。MEIT：面向报告生成的大语言模型多模态心电图指令微调。载于计算语言学协会论文集：*ACL 2025* 年会论文集，第14510–14527 页，奥地利维也纳。计算语言学协会。

Qianyue Wang、Jinwu Hu、Zhengping Li、Yufeng Wang、Daiyuan Li、Yu Hu 和 Mingkui Tan。2025年。使用记忆增强的动态分层大纲生成长篇故事。载于*2025*年美洲分会计算语言学协会会议论文集：人类语言技术（第*1*卷：长篇论文），第 1352–1391 页，新墨西哥州阿尔伯克基。计算语言学协会。

An Yang、Anfeng Li、Baosong Yang、Beichen Zhang、Binyuan Hui、Bo Zheng、Bowen Yu、Chang Gao、Chengen Huang、Chenxu Lv、Chujie Zheng、Dayiheng Liu、Fan Zhou、Fei Huang、Feng Hu、Hao Ge、Haoran Wei、Huan Lin、Jialong Tang 等 41 人。2025a。Qwen3 技术报告。*arXiv* 预印本 *arXiv:2505.09388*。

Kevin Yang、Yuandong Tian、Nanyun Peng 和 Dan Klein。2022年。Re3：通过递归式重新提示和修订生成更长的故事。载于：*2022*年自然语言处理实证方法会议论文集，第 4393–4479 页，阿布扎比，阿拉伯联合酋长国。计算语言学协会。

Zhaorui Yang、Bo Pan、Han Wang、Yiyao Wang、Xingyu Liu、LuoxuanWeng、YingchaojieFeng、Haozhe Feng、Minfeng Zhu、Bo Zhang 和 Wei Chen。2025b。多模态深度研究员：利用智能体框架从零开始生成文本—图表交错的报告。预印本，arXiv:2506.02454。

Lili Yao、Nanyun Peng、Ralph M. Weischedel、Kevin Knight、Dongyan Zhao 和 Rui Yan。2019年。规划与写作：迈向更好的自动故事生成。载于第三十三届 *AAAI* 人工智能会议、*AAAI 2019*、第三十一届创新人工智能应用会议、*IAAI 2019*、第九届 *AAAI* 人工智能教育进展研讨会、*EAAI 2019*，美国夏威夷檀香山，*2019*年*1*月*27*日至*2*月*1*日，第 7378–7385 页。AAAI出版社。

Shaolei Zhang、Ju Fan、Meihao Fan、Guoliang Li 和 Xiaoyong Du。2025年。Deepanalyze：用于自主数据科学的智能体大型语言模型。预印本，arXiv:2510.16872。

一项局限性

表 8：使用 Qwen3-VL-32B-Instruct 的各方法成本比较。应用程序接口调用次数表示每份报告的平均应用程序接口调用次数，延迟（秒）表示平均生成时间（秒）。

数据集	直接		DN.	
	API.	延迟。（秒）	API.	延迟。（秒）
Tableau	3.95	145.50	14.30	792.19
OurWorldInData	3.40	179.33	12.60	1047.89
USAFact	3.65	130.33	11.90	725.44
数据集	**DA.**		**Ours**	
	API.	延迟。（秒）	API.	延迟。（秒）
Tableau	18.93	79.50	41.85	946.15
OurWorldInData	15.75	78.05	43.20	1511.40
USAFact	21.32	81.80	33.13	1034.60

写入时视觉证据构建所需的额外时间和 API 成本。写入时证据构建会增加工具使用量和端到端延迟，因为写作器在生成过程中可能会触发多次可视化请求。如表格 8所示，EvidFuse 产生的额外 API 调用主要来自：（i）构建数据集概览*DO*；（ii）处理每项可视化请求，包括分析意图和规范、基于工具的渲染（并进行迭代优化），以及基于证据的说明生成。尽管存在这些开销，EvidFuse 仍显著提升了报告质量（见表格 1 和2）。在实践中，缓存或重用中间聚合、批处理兼容的请求以及并行工具执行等工程优化，可以在保持证据保真度的同时进一步降低成本。

稳健的可视化执行与恢复。由于 EvidFuse 需要执行代码来渲染可视化结果，反复执行失败可能会降低长篇报告中视觉证据的数量和质量（参见附录中的案例研究 H）。未来工作可通过经验证的执行和自我修复来提升稳健性，例如采用类型安全的数据适配器、对聚合进行单元检查、使用受限图表模板，以及在发生部分故障时采用优雅的回退策略，仍能提供最低限度且有用的可视化图表。

B 实验设置详情

B.1 数据集详情。

我们通过精选来自三个以可靠性和分析深度著称的权威数据源的报告，构建了一个全面的评估数据集。

表9：三个数据来源的主题分布。

主题	Tableau	OurWorldInData	USAFact
环境	3	2	2
经济	3	2	2
健康	1	2	3
商业	1	-	-
社会	2	2	1
交通	2	-	-
能量	1	-	-
教育	4	2	3
宗教	1	-	-
心理学	1	-	-
娱乐	1	-	-
人口	-	2	2
粮食与农业	-	2	-
AI	-	2	-
人权	-	2	-
War	-	2	3
政府	-	-	2
犯罪	-	-	2
总计	**20**	**20**	**20**

表10：不同数据集的统计数据。

数据集	表格。	Row.	列。
Tableau	113	1034.55	61.65
OurWorldInData	109	458.1	35.3
USAFact	87	295.8	37.1

。我们从这三个来源各收集了 20 份报告：Tableau 的公开数据画廊（Tableau）、Our World in Data（OWID）和 USAFact（US-AFacts），共计 60 份报告用于评估。对于每个来源，我们挑选涵盖多个主题领域的报告，包括环境、经济、健康、教育和社会议题，以确保广泛覆盖不同的分析视角。表 9展示了三个数据来源的详细主题分布，其中每个数字表示该主题收集的报告数量。对于每份报告，我们提取其标题、完整的 PDF 内容以及底层表格数据，并将其标准化后以 CSV 格式存储，以便所有方法都能进行一致的处理。

为全面了解数据规模和复杂性，表格10详细列出了各数据集的统计数据。“表格”列表示每个数据集中所有报告所包含的表格总数。“行”列报告每份报告中所有表格的平均总行数，反映数据量和粒度。“列”列显示每份报告中所有表格的平均总列数，体现数据的维度和特征丰富度。

B.2 实施细节

我们评估了两种模型配置，以考察 EvidFuse 的可扩展性和实用性。第一种配置采用 Qwen3-VL-32B-Instruct，通过 vLLM 部署在两张英伟达 A800 GPU（每张 80GB）上，以实现高效推理。第二种配置结合 Qwen3-235B-A22B-Instruct-2507 进行文本生成，并使用 Qwen2.5-VL-72B-Instruct 进行多模态理解，两者均通过 OpenRouter API 平台调用（OpenRouter）。这种混合配置展现了该框架在不同部署场景中的灵活性，同时保持了稳定的性能。

为进行受控比较，我们对所有生成可视化规范的方法使用相同的绘图后端，并在所有 LLM/MLLM 调用中保持解码设置一致（温度设为 0）。

• 直接提示：该基线方法是一种直接的端到端方案，将用户的意图与完整的表格数据（序列化为文本格式）直接拼接，作为大语言模型（LLM）的输入提示。在不依赖中间推理步骤或专用模块的情况下，指示 LLM 生成一份综合报告，并在适当位置嵌入 <可视化>标签。随后，一个确定性后处理模块会解析这些标签，并采用渲染方法将指定内容转换为实际的图表图像。

• DataNarrative：我们采用Islam 等人提出的成熟架构(2024年)，将其作为另一种基线方法。按照原始实现，该方法根据输入数据和叙述目标生成报告。与直接提示类似，该方法会在最终报告文本中生成 <可视化> 标签。为确保比较公平，并将生成质量与渲染质量分开评估，我们采用与直接提示基线相同的绘图工具，将这些标签内容转换为视觉图表。

• DeepAnalyze：我们遵循Zhang 等人（2025年）提出的图表优先、文本其次范式，引入了首个专为自主数据科学而设计的智能体大语言模型

。DeepAnalyze 通过专用动作令牌（<分析>、 <理解>、<代码>、 <执行>、<回答>）与数据环境交互，在报告生成之前执行全面的数据分析和可视化。它遵循基于课程的训练范式，逐步获取并整合多种能力，包括数据问答、专门的分析任务以及开放式数据研究。为确保公平比较，我们使用相同的数据集和用户意图作为输入来生成分析结果和可视化内容，然后让模型基于这些预先计算的可视化结果生成最终的 Markdown 报告。

• EvidFuse：我们采用两种不同的模型配置实现所提出的报告生成框架，以验证其在不同部署场景中的灵活性。在第一种设置中，我们在整个流程中使用 Qwen3-VL-32B-Instruct 作为统一模型，处理包括数据分析、图表生成、图表优化和最终报告撰写在内的所有任务。这种单模型配置展示了该框架能够借助中等规模的多模态大语言模型生成连贯报告。在第二种设置中，我们采用混合架构，根据任务的模态需求分配任务：基于文本的数据增强分析智能体（A'）使用 Qwen3-235B-A22B-Instruct-2507 进行分析推理和代码生成，而图表优化和多模态报告生成等视觉密集型任务则交由 Qwen2.5-VL-72B-Instruct 处理。这种混合配置发挥了更大型语言模型在复杂推理方面的优势，同时使用专用多模态模型执行视觉理解和生成任务。

B.3 评估详情

基于排序的协议。 我们不采用固定量表来分配绝对分数，而是向评估人员提供生成的报告和详细的评分细则，让他们对不同方法的**输出进行排序** 。这一设计解决了开放式生成中标量评分的常见局限性，包括：

1. 校准不佳，且不同实例或评审者使用评分量表的方式不一致。

2. 难以在异质的分析意图之间比较绝对分数。

3. 对锚定效应和中心趋势效应的敏感性。

相比之下，排序关注相同输入条件下的相对偏好，因此能带来更稳定的比较。我们通过平均排名汇总评估集中的各实例排名，从而得到每种指标和粒度下方法层面的最终表现。

各项指标的提示词。 我们设计了六项评估指标，用于评估报告质量的不同方面。以下展示了这些指标的评估提示词：布局合理性（图 3）、可读性（图 4）、文本与图表一致性（图5）、文本信息深度（图6）、信息丰富度（图 7）和视觉一致性（图8）。每个提示词都提供了详细的评分细则和基于示例的标准，以确保排名一致且有意义。

C 提示

本节详细介绍用于实现 EvidFuse 框架的系统提示和任务专属指令。这些提示旨在协调数据增强分析智能体与实时证据构建写作器之间的协作，从而确保有效的任务分解以及连贯的文本与图表交错生成。

C.1 报告大纲生成

大纲生成提示由报告撰写器使用， W 用于分解用户的请求， R_{user}形成结构化高层大纲， O 并以数据集概览（公式 5）为条件。该提示遵循规划-撰写范式（Wang 等人，2025年），指示写作器首先以粗粒度规划报告结构，然后生成详细内容（完整提示见图 9）。该大纲充当路线图，在保持即时分析和可视化灵活性的同时，指导后续的增量撰写。

C.2 增量报告撰写

增量撰写提示引导报告撰写器 W 逐段生成文本与图表交错的报告片段（公式6）。在每一步中 i，写作器会根据历史记录 t_i 、大纲 H_{i-1} 和数据概览 O生成文本片段 DO。

布局评估提示

任务
你的任务是评估并比较图表布局，确定哪一种最能利用空间布局来讲述引人入胜的数据驱动的故事。

布局评估标准
布局指图表、文本和图形元素如何编排，以引导读者理解。根据提供的示例，理想的布局应发挥数据新闻作品的作用，并优先考虑：

- **叙事整合流程**：优先选择视觉层级与分析逻辑相呼应的布局。选择一种结构，从“背景铺陈”（描述性地图／分布）过渡到“深入分析”（散点图／趋势），并以“综合”（结论）收尾。

- **嵌入式洞察与注释**：优先选择将洞察放在图表边界*内*的布局。使用**直接标注、指向异常值的箭头以及图表内文本框**（例如“显著下降.....”或“反向关系”）的布局得分更高，而不是仅依赖外部说明文字。

- **综合仪表板**：优先选择采用**多面板仪表板**布局（通常置于报告末尾）的报告，将关键指标（地图、趋势和统计数据）汇总到单一的高层视图中，以便进行跨指标比较。

- **问题驱动式框架**：优先选择这样的布局：章节标题提出问题（例如“可及性是否在改善？”），紧接着的图表则给出直观答案。图表标题应以陈述句概括数据。

- **统计与视觉一致性**：对于复杂元素，优先选择保持严格网格布局的方案，例如对齐**发散条形图**或带有 LOESS 曲线的**散点图**，确保参考线（中位数、平均值）和误差线清晰易读，并在不同图表之间保持一致。

输入
我已上传下方的图表图片。它们按报告 ID 分组。

图3：用于评估报告图表布局的提示。

阅读评估提示

任务
你的任务是评估多份报告中图表图像的质量和可读性。

评估标准
你将评估图表传达复杂数据分析结果的有效性。根据所提供的示例，高质量图表应优先考虑**统计深度**、**叙事背景**和**多维综合**。请使用以下具体标准：

- **统计与分析严谨性**：优先选择超越原始数据点、包含统计增强的图表。关注**趋势线（例如 LOESS平滑）**等用于展示相关性的特征、用于展示变异性的**误差线/置信区间**，或用于处理指数数据的**对数刻度**。

- **叙事驱动的注释**：优先选择将“故事”直接融入视觉呈现的图表。图表应使用**描述性标题、标注框和对异常或关键洞见的直接标注**（例如“延迟护理降幅最大”），而不是要求读者自行寻找含义。

-**多维综合（仪表板设计）**：优先选择将多个相关指标组合到单一连贯视图中的可视化（例如，将地图、趋势线和条形图结合起来的仪表板）。能够一目了然地综合呈现**地理、频率和幅度/强度**的布局应获得高分。

- **分布与比较清晰度**：优先选择能够揭示数据*形态*而不仅仅是平均值的图表。应关注展示分布范围的**箱线图**、展示正/负划分的**发散条形图**，或带有参考线（例如“全市中位数”）的**直方图**，从而支持即时基准比较。

- **地理空间与时间背景**：当位置或时间具有相关性时，优先选择能够有效将数据映射到**地理集群**（例如气泡地图、分级统计地图），或清晰展示**时间演变**（例如明确区分前后时期或呈现长期趋势）且不会造成视觉杂乱的图表。

输入
我已上传以下图表图片。它们按报告 ID 分组。

图4：评估图表可读性的提示。

T-C 一致性评估提示

任务
你的任务是根据**文本-图表一致性**，评估并比较多份报告中文本-图像对的质量。

文本-图表一致性评估标准

文本与图表一致性是指书面分析与特定图表和表格的关联紧密程度。比较报告时，应优先选择符合以下特点的报告：

* **统一的概念和指标**：关键术语和指标只定义一次，随后在正文、表格和图表中始终使用相同的名称、单位和阈值。 * **文本与图形一一对应**：正文中的每项主要结论都能直接追溯到某个时间范围、变量和比较组相匹配的特定图表或表格。 * **明确的数据范围和限制**：图表清楚标明数据范围、假设以及缺失或不完整的数据，正文在解读结果时也会重申这些限制。 * **综合摘要可视化图**：使用仪表板或综合图来概述关键模式，结论部分明确结合这些可视化图表进行阐述，从而形成完整闭环。

输入
我已上传以下文本-图像对，并按报告 ID 和生成方法对应的简称分组。

图 5：用于评估文本与图表一致性的提示。

提示会指示写作器在需要分析证据或视觉支持时，通过发出 <可视化> 标签来触发可视化请求（完整提示见图 10）。这一设计使写作过程中的实时数据交互成为可能，让写作器能够随着叙述的展开，按需请求分析智能体生成图表。

C.3 可视化规范规划

可视化规范规划提示P_{ana} 由数据增强分析智能体 A' 用于解读并回应作者提出的可视化请求。如式2所示，给定一个可视化请求 q （置于<可视化> 标签中）， A' 会应用 P_{ana} 来提取分析意图 i，并生成一份详细的可视化规范 vs，其中包含所引用的表格子集 D_{ref}。该提示（见图11）引导智能体利用其数据集特定知识（来自注入的数据概览）来选择相关变量、确定恰当的图表类型和聚合方式，并生成基于上下文、符合作者叙述要求的可视化规范。

C.4 标题生成

图注生成提示 $P_{caption}$（见图12）由数据增强分析智能体 A' 用于生成面向分析的见解，直接回应作者的可视化请求。如式4所示，给定生成的图表 c、原始分析意图 i以及作者的查询 q， A' 会应用 $P_{caption}$ 来生成简洁且依据充分的说明文字 s，该说明文字：（1）直接回应 q中提出的分析问题；（2）严格依据 c中呈现的视觉证据；（3）符合分析意图 i。

D 可视化规范的详细内容

可视化规范（vs）是一种结构化的中间表示，用于连接自然语言可视化请求与可执行的绘图代码。它由分析智能体 A', vs 制定，以机器可读的 YAML 格式编码图表生成所需的所有参数，确保视觉输出精准且可复现。这种结构化方法缓解了自然语言请求固有的歧义，同时维持

深度提示。评估

任务
你的任务是根据**信息深度**，评估并对多份报告中的文本-图像对质量进行排序。

文本–深度评估标准
文本深度是指报告将数据和可视化图表转化为结构化、系统性叙述的有效程度，而非罗列相互孤立的评论。比较报告时，应优先选择符合以下特点的报告：

* **多维度覆盖**：图表共同涵盖时间、空间、幅度和能量（或同等重要的维度），形成综合仪表板或概览表，而非仅呈现单一角度。

* **深度定量解读**：文本始终对完整分布和关键统计量（例如典型水平、变异性、偏度和异常）进行解读，并将其联系到核心问题，例如模式是否正在上升或是否存在异常。

* **视觉-文本闭环逻辑**：图表与正文围绕少数反复出现的主题组织，后续章节则整合前文发现，形成简洁、可用于决策的结论。

输入
我已上传以下文本-图像对，并按报告 ID 分组。

图6：用于评估文本信息深度的提示词。

同时保持灵活性，以适应多种可视化类型。

一份完整的可视化规范至少包含四个核心组成部分：（1）**图表_类型**：可视化类别（例如折线图、条形图、散点图），它决定了渲染范式；（2）**标题**：概述可视化洞见的描述性图表标题；（3）**数据**：从原始表格中提取的预处理数据集，以键值对形式呈现，每个条目均代表一个数据点，并包含相应的维度和指标；（4）**标签**：坐标轴标签、注释和辅助文本元素，以提升可解释性。图13 和图14 展示了可视化规范在实际应用中的两个具体示例。

E 可视化工具的详细信息 T

为确保 EvidFuse 中生成的可视化结果具备高质量和可靠性，可视化工具T实现了一套稳健的三阶段优化机制，用以应对自动化图表生成中的关键挑战：代码执行失败、视觉设计欠佳以及候选项选择。该机制结合了：(1) 通过重试逻辑实现错误恢复的迭代式代码生成；(2) 通过多模态反馈提升视觉质量；以及 (3) 系统化的最佳候选项选择，确保只有高质量图表才会整合到最终报告中。

阶段 1：初始图表生成。系统会反复生成并执行图表代码，直到生成有效图表或达到重试上限。每次重试时，基于文本的大语言模型 M_t 都会生成新的可视化代码，然后在环境 E 中执行。此过程形式化如下：

$$c_0 = M_t(d), \tag{9}$$

$$(success, i_0) = E(c_0), \tag{10}$$

其中，c_0 是生成的代码，d 是可视化描述，$success_0$ 表示执行是否成功，i_0 是在执行成功时渲染后的图表图像。重试循环会在首次成功执行（$success = \text{True}$）时终止，并生成初始图表集 $C = \{c_0\}$。如果所有尝试均失败，系统将报告失败。

阶段 2：视觉优化。成功生成初始图表 c_0 后，系统会通过多模态大语言模型评审器M_v启动视觉优化。在每次优化迭代 j ($1 \leq j \leq N_{\text{refine}}$)中，评审器首先评估渲染后的图表

信息评估提示

任务
你的任务是评估并排列多份报告的信息丰富度。

报告信息丰富度
报告的信息丰富度，是指报告通过紧密结合可视化图表与正文，围绕标题中的核心问题涵盖多个关键维度，而非仅仅罗列孤立的结果。比较报告时，应优先考虑符合以下特点的报告：

* **多维度可视化图表**：同一份报告中的图表以一致的布局共同呈现时间趋势、空间模式和震级–能量关系，使每幅图及整体可视化图表集都具有较高的信息密度。

* **以解释性正文为骨干**：正文围绕核心问题组织，解释关键概念和指标，并将观测结果、可能原因与结论串联成连贯的叙述，持续丰富每幅可视化图表所传达的信息。

* **综合综合**：最终的综合总结或仪表板将频率、强度和能量汇总为简洁的全局视图，把详细分析转化为可重复使用的分析概览。

***内容体量与分布**：该报告通过可观的**总词数**体现深度，并确保每个**章节中的文本和图像数量**都很充足，从而在所有部分保持连贯、详尽的内容覆盖。

输入
我已在下方上传报告内容，并按报告 ID 分组。

图 7：用于评估报告信息丰富度的提示。

图像：

$$feedback_j = M_v(i_{j-1}), \tag{11}$$

其中 $feedback_j$ 包含详细的文本批评，指出视觉质量问题。随后，基于文本的大语言模型执行器 M_t 会结合批评者反馈来优化先前的代码：

$$c_j = M_t(c_{j-1}, feedback_j, d). \tag{12}$$

使用公式（10）执行优化后的代码；若执行成功，则将新图表添加到候选集：$C \leftarrow C \cup \{c_j\}$。此优化过程最多持续 N_{refine} 次迭代，并保留所有成功生成的版本。

阶段 3：最佳图表选择。当存在多个图表版本时（$|C| > 1$），多模态批评者 M_v 会将所有候选图表与原始可视化请求r进行比较，从而选出最佳图表 $\tilde{c}$：

$$\tilde{c} = \underset{c \in C}{\text{argmax}}\ M_v(\text{“evaluate quality”}, r, c). \tag{13}$$

如果只存在一张图表，则直接将其选为$\tilde{c} = c_0$。这种三阶段机制显著通过将错误处理与美学优化分开，并通过系统比较确保选出最佳候选项，提升了图表可靠性和视觉质量。

F 人工评估详情。

F.1 评估者配置与抽样策略

为确保评估稳健且公正，我们招募了六位具有数据分析背景的评估人员。在评估过程开始前，所有评估人员均接受了有关评估标准和评估规程的培训。

针对每种模型设置和每个评估维度，我们从六名评估人员组成的人员库中随机选取三人，对相应维度进行评估。这种随机分配策略有助于减轻个体偏差，并确保不同评估方面包含多样化的视角。由于我们在两种模型设置下评估六个维度，每位评估人员都会参与多个维度的评估，但没有任何一位评估人员评估所有维度

视觉一致性评估提示词

任务
你的任务是评估多个报告的可视化一致性并进行排名。

可视化一致性
可视化一致性意味着报告遵循适合专业出版的严谨、统一的设计语言，确保不同部分中的各类视觉元素都像是同一套连贯系统的组成部分。在比较报告时，应优先选择具备以下特征的报告：

* **统一设计系统**：所有图表都严格遵循特定的配色方案和字体层级。无论查看的是折线图、条形图还是复杂的密度图，其视觉识别（例如主色和辅助色）都始终清晰一致。
* **语义化视觉逻辑**：颜色和样式的使用具有逻辑性和语义性，而非随机选择。例如，特定颜色在不同图表类型中始终代表特定的数据集或变量，使读者无需重新查阅图例，就能在整份报告中直观地追踪某个变量。

* **标准化结构元素**：对非数据墨水的处理保持细致而统一，包括网格线不透明度、坐标轴标签格式、图例位置和注释样式。

* **出版级润色**：可视化呈现堪称完美，没有任何风格冲突，确保即使图表类型差异显著，整体视觉呈现依然精致、专业。

输入
我已在下方上传报告内容，并按报告 ID 分组。

图 8：用于评估可视化一致性的提示。

适用于任何给定设置。

我们在两种开源模型设置下评估了报告生成：（1）**Qwen3-VL-32B-Instruct** ，作为具有代表性的小规模开源模型；（2）**Qwen3-235B-A22B-Instruct-2507 & Qwen2.5-VL-72B-Instruct**，作为大规模开源模型。我们从三个数据集中各随机抽取 10 份生成的报告，因此，每种模型设置在每个维度下共评估 30 份报告。这一抽样策略确保了对不同领域的全面覆盖，同时将评估工作量控制在可管理范围内。

F.2 评估流程与问卷设计

人工评估协议的设计严格遵循我们的自动化评估框架。对于每个评估实例，评估人员都会看到针对相同输入数据、由不同方法（我们的方法和三种基线方法）生成的报告。为消除偏见，这些报告被匿名处理，并随机标记为方法 A、B、C 和 D，且标签与方法的对应关系在不同评估实例之间随机打乱。

评估人员依据与自动评估指标相同的六个维度对每份报告进行评估。对于每个维度，评估人员都会收到：

- **维度定义**：清楚说明该维度衡量的内容
- **评估标准**：关于如何判断报告质量的详细指南
- **参考示例**：展示不同质量等级的报告样例

评估人员并非按照固定量表给出绝对分数，而是根据相同的评估标准，对候选报告进行**排序**，从**1（最佳）**到**4（最差）**。

G 生成的报告的信息丰富度测量详情。

为了定量评估生成的报告的信息丰富度，我们采用命题级分析，将报告内容分解为细粒度的语义单元。我们的方法包括两个主要阶段：**(1)抽象式命题分割（APS）**，从报告中提取原子命题；**(2)信息过滤与聚类**，根据命题的信息价值对其进行分类。

G.1 抽象命题分割

我们在本地部署 Gemma-APS-7B（Hosseini 等人，2024年），从每份生成的报告中提取命题。对于包含 N 个句子的报告，该模型会生成 M 个命题，其中通常 $M \geq N$，因为复杂句子会被分解为多个原子单位。

G.2 信息过滤与聚类

提取命题后，我们应用一种信息过滤机制，根据命题的信息价值对其进行分类。该过程应对了这样一个挑战：并非所有命题都对报告质量作出同等贡献——有些命题可能是冗余的、无效的，或过于冗长。我们使用大语言模型将每个命题分为三类：**无效信息**，包括事实错误、模型幻觉或未以源数据为依据的命题；**重复信息**，指在报告其他位置已经表达过、重复语义内容的命题；以及**简化信息**，由有效且不重复、能够贡献独特语义内容的命题组成。通过利用 APS 进行细粒度语义分解，并利用基于大语言模型的信息过滤进行质量感知分类，我们的评估框架超越了表层文本统计数据，转而评估生成的报告所包含的实质性信息价值。

H 案例研究

为全面了解 EvidFuse 的能力和局限性，我们展示了采用我们的方法在不同场景下生成的两个案例研究。

H.1 良好案例示例

以下是一个成功案例，展示了 Evid-Fuse 的有效性。该报告呈现了数量可观且复杂的可视化，并实现了紧密的文本-视觉连贯性，文字发挥着视觉引导作用——例如，明确描述“绿色条形（女性）比橙色条形（男性）延伸得更远”，以解读区域差异。此外，文字描述与复杂的视觉编码直接对应，例如指示读者观察“蓝色小提琴图（女性）的位置低于红色小提琴图（男性）”，以确认各地区死亡率的一致性。

为什么女性比男性长寿？全球与历史视角的探索

背景铺陈：全球寿命性别差距

现象简介

在几乎所有国家和地区，女性的寿命都稳定地长于男性；多方来源的全面数据印证了这一现象。根据数据表1，2023年全球女性的平均预期寿命约为78.07岁，男性则为73.06岁，两者相差约5.01岁。数据表2进一步证实了这一趋势：在28个国家中，女性与男性的平均预期寿命差为4.89岁。这种持续存在的寿命性别差距不只是统计数字; ，它还反映了根深蒂固的生物、社会和环境因素，这些因素从婴儿期到老年都会影响生存率。

这一差距的幅度因地区而异，差异显著。欧洲和大洋洲的差距最大，通常超过5至6年；非洲的差距则较小，通常约为2.5至3年。地区间的这种差异表明，尽管生物学因素可能使女性更容易长寿，但社会条件在扩大或缩小这一差距方面也起着关键作用。例如，白俄罗斯的差距极大，接近9.5年；而巴林的差距则是最小的之一，仅为1.3年。这些异常值凸显了国家政策、医疗服务可及性和文化规范对死亡率结果的深远影响。

世界地图生动地展示了预期寿命差距在全球的分布，凸显出鲜明的地区差异。东欧，尤其是白俄罗斯，以深蓝色标示，十分醒目，表明其差距高达 9.33 年。相比之下，非洲大部分地区和中东部分地区以红色标示，反映出较小的差距。这一视觉呈现强调，尽管长寿方面的性别差距几乎普遍存在，其幅度却深受地理和社会政治背景的影响。这幅地图是识别需要采取干预措施的地区的有力工具

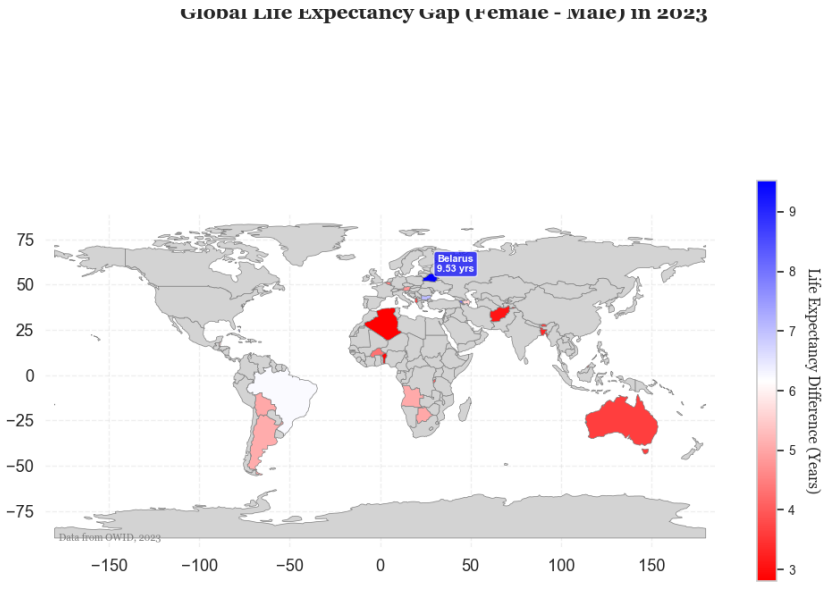


图1：2023年全球预期寿命差距（女性－男性）。2023年全球预期寿命差距（女性－男性）：一幅按女性与男性预期寿命差异进行颜色编码的世界地图，采用发散色阶，其中蓝色表示差距较大（白俄罗斯最高可达9.33年），红色表示差距较小。数据来自OWID，2023年。

可能对减少男性可预防死亡产生最大影响。

长寿的区域模式

长寿的区域模式呈现出从高预期寿命到低预期寿命的明显梯度，其中欧洲和大洋洲位居前列。数据表1显示，欧洲国家男女的预期寿命始终最高，女性平均约为83—84岁，男性约为79—80岁。大洋洲紧随其后，女性预期寿命接近81—82岁，男性接近78—79岁。相比之下，非洲的预期寿命全球最低，女性中位数接近69—70岁，男性中位数接近64—65岁。亚洲、北美洲和南美洲处于中间位置，女性预期寿命通常为70至80岁，男性为65至75岁。

这种区域梯度不仅体现在预期寿命的绝对水平上，也体现在性别差距的大小上。欧洲和大洋洲的性别差距最大（约为5—6年），而非洲的差距较小（约为2.5—3年）。这一模式表明，在整体健康状况较差的地区（如撒哈拉以南非洲），由于传染病、冲突以及医疗服务可及性有限等因素，男女死亡率都较高，因此女性所享有的相对优势有所缩小。相反，在公共卫生系统完善的发达地区，女性的生理优势得以更充分地发挥，性别差距因而更大。

这张水平条形图清晰地展示了六个世界区域中一致存在的预期寿命性别差距。在每个地区，绿色条形（女性）都比橙色条形（男性）更长，证实无论地理位置如何，女性的寿命都比男性长。欧洲和大洋洲的差距最为显著，超过了 5 年。相比之下，非洲的差距较小，但两性的预期寿命总体上都显著偏低。这一可视化进一步印证了核心观察：尽管这一现象遍及全球，但其表现因上下文而异；发达地区的绝对预期寿命更高，性别差距相对也更大。

值得注意的例外

尽管全球趋势显示女性更长寿，但一些显著的例外凸显了特定国家上下文的作用。白俄罗斯是一个异常值，其报告的预期寿命差距最大，约为 9.5 年，这表明可能存在独特的社会或健康因素。可能的影响因素包括男性心血管疾病死亡率的历史模式、职业危害，或吸烟、饮酒等生活方式因素；这些因素在该地区的男性中可能更为普遍。相反，巴林的差距则是最小的之一，仅为 1.3 年，这表明两性的死亡率可能都极低，或者与其他国家相比，男性的健康状况相对更好。

这些极端情况凸显出，尽管生物学为女性提供了基础优势，社会条件仍可能显著改变差距的幅度。在 con-

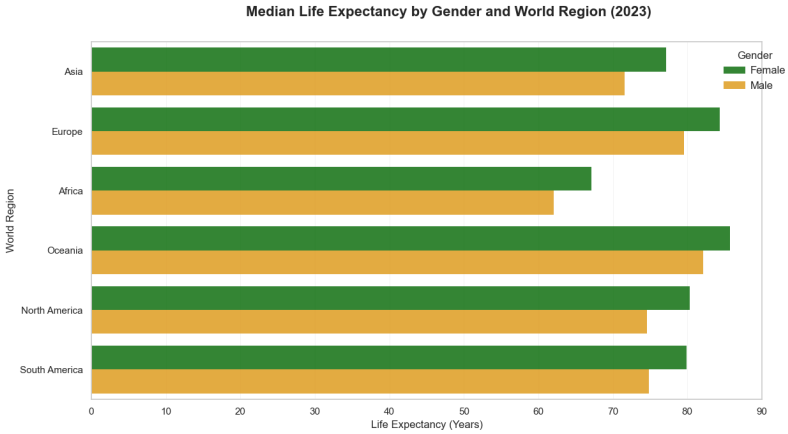


图2：2023年按性别和世界区域划分的预期寿命中位数。该水平条形图展示了2023年六个世界区域按性别划分的预期寿命中位数，显示女性在每个地区的寿命都持续高于男性。在亚洲、欧洲、非洲、大洋洲、北美洲和南美洲，女性（绿色条形）的预期寿命均高于男性（橙色条形），其中欧洲和大洋洲的差距最大。这一视觉结果凸显了女性长寿优势的全球性模式，引发了人们对造成这种差异的生物、行为和社会经济因素的探究。

在冲突地区或医疗保健基础设施薄弱的地区，例如非洲部分地区，由于高死亡率同时影响两个性别，差距会缩小。相比之下，拥有先进医疗保健系统和完善社会安全网的国家，例如北欧国家，能够让女性的生理优势得到更充分的体现，从而形成更大的差距。理解这些异常值对于制定有针对性的公共卫生干预措施至关重要，旨在减少高风险地区男性可预防死亡的发生。

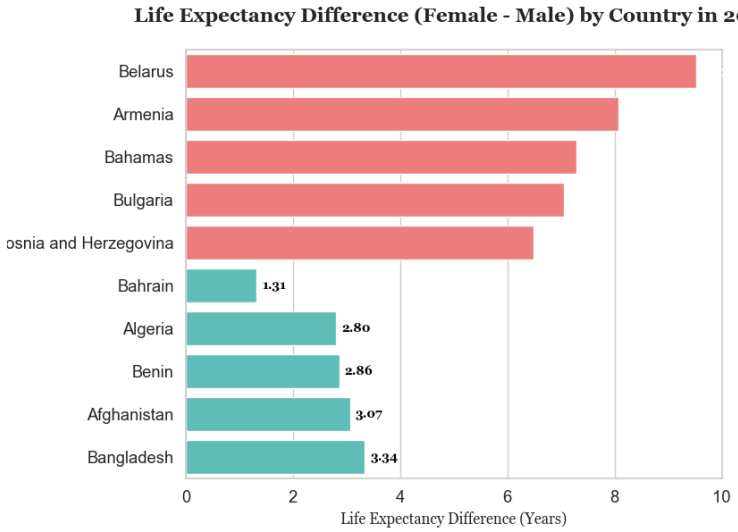


图3：2023年各国预期寿命差（女性－男性）。该水平条形图展示了2023年各国的预期寿命差（女性减去男性），并突出显示差距最大的前五个国家——白俄罗斯（9.5年）、亚美尼亚（8.1年）、巴哈马（7.3年）、保加利亚（7.0年）以及波斯尼亚和黑塞哥维那（6.5年）——和差距最小的后五个国家——巴林（1.31年）、阿尔及利亚（2.80年）、贝宁（2.86年）、阿富汗（3.07年）以及孟加拉国（3.34年）。该可视化凸显了各国之间基于性别的预期寿命存在显著差异。

水平条形图有效地按预期寿命差距对国家进行排序，展现出各种不同的结果。白俄罗斯以女性比男性高出9.5年的显著优势位居首位，其后是亚美尼亚和巴哈马。另一端，巴林的差距最小，仅为1.31年；阿尔及利亚、贝宁、阿富汗和孟加拉国的差异也相对较小。这一可视化强调，从医疗保健质量、生活方式到社会经济状况等国家层面的因素，在决定寿命性别差距的幅度方面发挥着决定性作用。它也发出了行动呼吁，指出了可通过有针对性的干预显著改善男性健康状况的国家。

揭开冒险面纱：探索差距驱动因素

婴儿死亡率：主要驱动因素

全球预期寿命性别差距的主要成因是婴儿死亡率。数据表5显示，在研究的全部28个国家中，男婴的死亡率都高于女婴，全球平均差异约为每100名活产婴儿多0.25例死亡。男婴在生物学上的脆弱性——例如免疫系统较弱、出生体重较低以及更容易感染——使他们从出生起就处于不利地位，并显著加剧了总体预期寿命差距。即使在澳大利亚和比利时等医疗保健发达的国家，男婴的死亡率仍略高，这表明这种差异根植于生物学因素。

婴儿死亡率的影响在高风险地区尤为显著。阿富汗、安哥拉和孟加拉国等国家的婴儿死亡率位居全球前列，男性婴儿死亡率通常超过每100名活产婴儿5例。在这些情况下，婴儿死亡率性别差距可能超过每100名活产婴儿1.5例死亡，从而显著加剧整体预期寿命差距。相比之下，在整体死亡率较低的发达国家，这一差距有所缩小，但由于固有的生物学差异，仍然存在。这一普遍规律表明，社会因素虽可减轻或加剧这一差距，但其根源在于生命最初的日子。

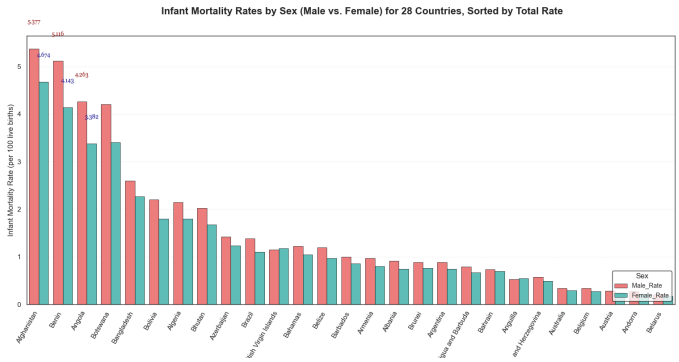


图4：28个国家按性别划分的婴儿死亡率（男性与女性），按总死亡率排序。28个国家按性别划分的婴儿死亡率（男性与女性），按总死亡率排序：这张并列条形图比较了28个国家中男性和女性婴儿每100名活产的死亡率，并按婴儿死亡总率降序排列。图表显示，所有国家的男性死亡率均高于女性，其中阿富汗和安哥拉的总体死亡率最高。

并列条形图生动地展示了全部 28 个国家中男性婴儿死亡率始终较高的模式。阿富汗和安哥拉尤为突出，其总计死亡率超过每 100 例活产 5 例，且男性死亡率显著高于女性

死亡率。即使在澳大利亚和安道尔等总体死亡率较低的国家，男性死亡率仍略高于女性死亡率。这一视觉证据进一步表明，婴儿死亡率是造成预期寿命性别差距的普遍因素，其影响在面临重大健康挑战的地区最为严重。

成年死亡率：隐藏因素

虽然婴儿死亡率是主要因素，但成年死亡模式也会造成寿命差距，尤其是在特定年龄组中。数据表4显示，在65岁时，每个国家男性的中心死亡率都高于女性，全球平均差异约为0.9年。男性在晚年持续处于劣势，反映出他们面临更高的心血管疾病、事故及其他死因风险，而这些因素对男性的影响尤为显著。数据显示，即使在日本和澳大利亚等高收入国家，65岁男性的死亡率仍然更高，这表明生物学和行为因素在成年期仍持续发挥作用。

这些成年死亡率差异的地区分布进一步凸显了医疗服务可及性和生活方式方面的不平等。在65岁时，非洲男女的中心死亡率均为最高，男性约为2.75，女性约为2.05，这表明医疗服务提供和社会经济状况面临系统性挑战。相比之下，欧洲的总体死亡率最低，男性为 ~1.75，女性为 ~0.8，这反映了完善的公共卫生系统和更健康的生活方式。这一模式表明，尽管生物学因素可能使男性更容易面临较高死亡率，但社会决定因素会显著影响这种劣势的幅度。

小提琴图详细展示了不同性别和地区的65岁人群中心死亡率分布差异。在每个地区，蓝色小提琴图（女性）的位置都低于红色小提琴图（男性），这证实了女性在这一年龄段的死亡率较低。图中还揭示了显著的地区差异：非洲的男女死亡率均为最高，且分布范围较广，表明变异性较高；欧洲的死亡率最低，分布范围也较窄。这一可视化凸显了女性在晚年享有的优势在各地区普遍存在，但在医疗保健和生活条件较好的地区更为显著。

历史背景：法国 1816–1843

法国1816年至1843年的历史数据为理解预期寿命性别差距的成因提供了宝贵背景。数据表6显示，在此期间，婴儿死亡率（0岁年龄组）是造成差距的主要因素，平均贡献约为1.17年，是成年死亡率（15–39岁年龄组，~0.12年）贡献的九倍多。这一发现与现代数据一致，并表明男婴的生物学脆弱性在历史上始终是一个影响因素。

数据还显示，15–39岁年龄组的贡献存在显著波动，其中某一年对差距的贡献接近0.9年——这可能是战争或疾病对年轻男性造成了更大影响所致。这一历史发现凸显了外部

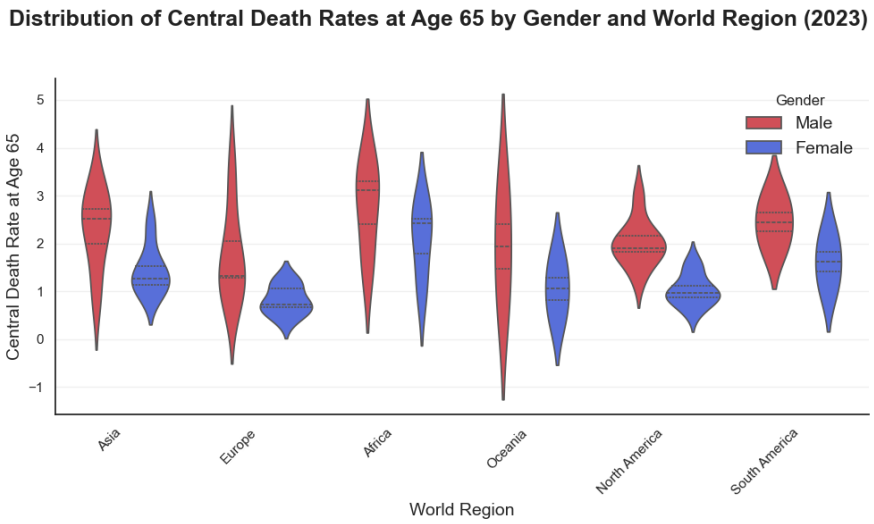


图 5：2023年按性别和世界区域划分的65岁人群中心死亡率分布。该小提琴图展示了2023年六个世界区域中65岁男性（红色）和女性（蓝色）的中心死亡率分布，显示女性在所有地区的死亡率始终低于男性，这也是女性预期寿命更长的原因之一。

这些事件可能会暂时改变死亡模式，并加剧性别差距。即使在19世纪，婴儿期仍是主要驱动因素，这进一步印证了生物学在塑造寿命差异方面起着基础性作用，而社会和环境因素则会随着时间推移调节这些影响。

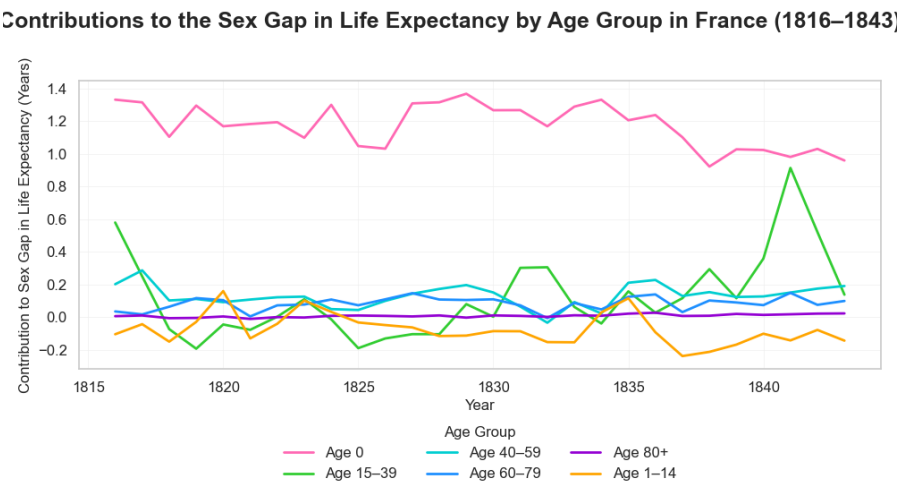


图6：法国不同年龄组对预期寿命性别差距的贡献（1816–1843）。该图展示了1816年至1843年法国不同年龄组对预期寿命性别差距的贡献，显示婴儿期（0岁，粉色线）始终是造成差距的最大因素，女性寿命长于男性主要是因为男性婴儿死亡率较高。其他年龄组（如15–39岁，绿色；1–14岁，橙色）的贡献虽有波动，但与婴儿期相比仍然较小，这凸显了早期生命阶段的生存差异在解释这一时期女性为何比男性长寿方面占据主导地位。

这张折线图清晰地表明，1816 年至 1843 年间，婴儿期（年龄为 0 岁，粉色线）是法国预期寿命性别差距的主要来源，其影响远超其他年龄组。粉色线始终徘徊在 1.2 年左右，而其他年龄组（如 15–39 岁组，绿色线；1–14 岁组，橙色线）的贡献虽有波动，但仍显著较低。这一视觉证据证实，即使在 19世纪，性别差距的主要成因仍是男婴死亡率较高。与 15–39 岁组的波动相比，婴儿期的贡献随时间变化相对稳定，这凸显了这种差异背后持久的生物学基础。

异常值与离群值

在法国的历史数据中，15–39岁年龄组呈现出显著的异常值与离群值。散点图矩阵显示波动极大，其中某一年对性别差距的贡献接近 0.9 年——这可能是战争或疾病对……造成的影响

年轻男性所受影响尤为显著。这个异常值在 15–39岁年龄组的直方图中清晰可见：分布高度偏态，右侧长尾明显，证实存在极端值。

这些异常似乎与影响中年人群的更广泛趋势有关，这一点从 15–39岁年龄组与 40–59岁和 60–79岁年龄组之间的正相关中可见一斑。然而，较年轻年龄组之间并无明显模式，这表明导致青春期性别差距的因素不同于婴儿期或幼儿期的相关因素。这项分析凸显了外部事件如何暂时改变死亡模式并加剧性别差距，也为理解生物学与环境在塑造寿命差异过程中的复杂相互作用提供了宝贵洞见。

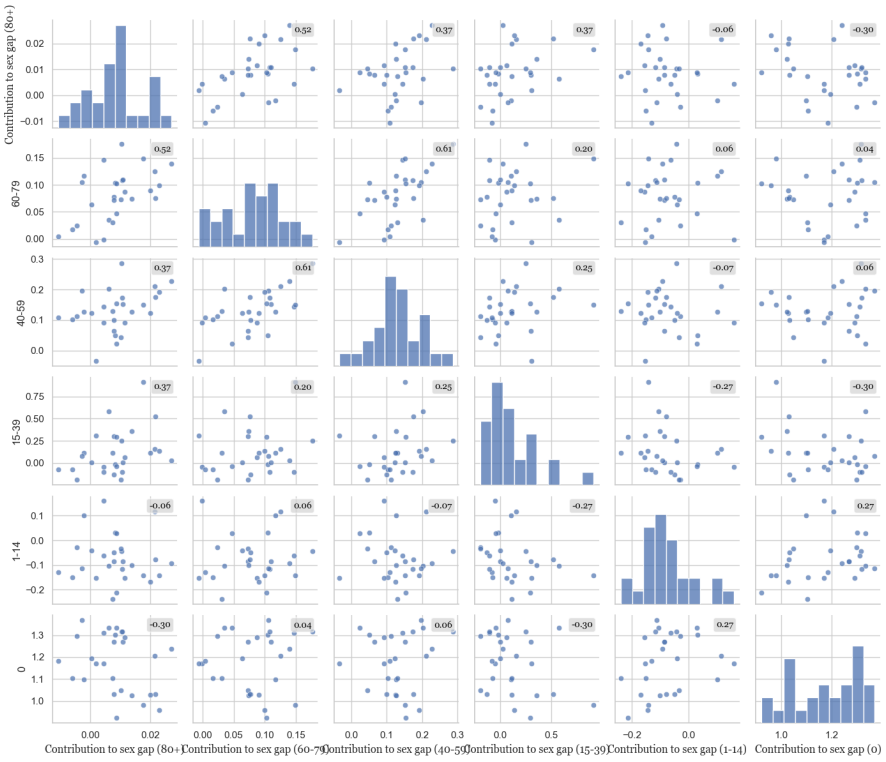


图 7：各年龄组对预期寿命性别差距贡献的相关矩阵（法国，1816–1843）。各年龄组对预期寿命性别差距贡献的相关矩阵（法国，1816–1843）：散点图矩阵展示了不同年龄组对预期寿命性别差距的贡献之间的关系，并突出显示了 15–39岁年龄组与较年长年龄组之间的正相关，说明各年龄段的死亡模式共同解释了女性寿命为何长于男性。

散点图矩阵揭示了不同年龄组对预期寿命性别差距的影响之间存在着复杂的相关性网络

。值得注意的是，15–39岁年龄组与40–59岁年龄组和60–79岁年龄组之间呈正相关，相关系数分别为0.25和0.37。这表明，年轻男性死亡率较高的年份，中年和老年男性的死亡率也往往较高，这可能是由于多代人共同面临风险因素或受到历史事件影响。相比之下，与较年轻年龄组（0岁和1–14岁年龄组）的相关性较弱或为负，表明不同生命阶段的死亡率驱动因素彼此独立。

总结：综合证据，展望未来

生物与社会图景

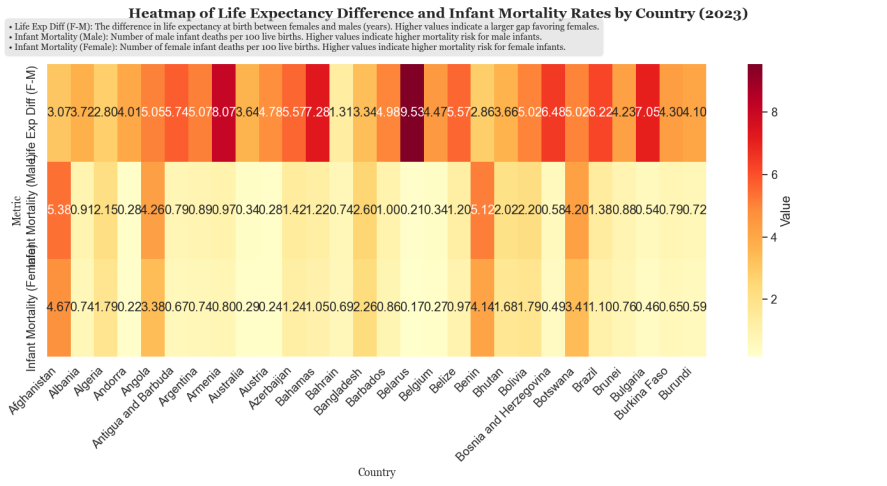
综合现有证据可见，女性寿命长于男性，这是生物学优势与社会决定因素共同作用的结果。从生物学角度看，女性的免疫系统更强、生理韧性更好，因此婴儿死亡率更低，形成了生命早期的优势。这一生物学基础在成年后仍持续发挥作用：女性患心血管疾病、遭遇事故和发生物质滥用的风险更低，而这些因素对男性的影响尤为显著。

从社会层面看，区域差异反映了社会经济状况。医疗保健、营养和安全条件较好的国家（如日本、澳大利亚），预期寿命差距更大，因为女性的生物学优势得以更充分地发挥。相反，在冲突地区或医疗保健条件较差的地区（如贝宁、阿富汗），由于男女双方的死亡率都较高，预期寿命差距较小。生物学与社会之间的相互作用决定了这一差距的幅度，使其成为一种需要多方面解决方案的复杂现象。

这张热力图将各国的预期寿命差与婴儿死亡率有效结合，揭示出一个明显规律：男女婴儿死亡率都较高的国家（如阿富汗、贝宁），往往预期寿命差距中等偏大，这表明尽管男女双方都面临较高风险，女性的生物学优势仍然显现。相比之下，婴儿死亡率较低的国家（如澳大利亚、日本）预期寿命差距较大，这表明男女双方的生存状况都较好，但女性的寿命明显更长。这一视觉综合结果凸显了生物学与环境在塑造寿命差异方面的双重作用。

中国案例研究：百年变迁

中国是一个引人注目的案例，展示了社会变迁如何影响长寿方面的性别差距。数据表3显示，中国的预期寿命差距从1950年的约3.5年增至2023年的5.6年，这反映了医疗保健水平的改善以及男性与生活方式相关疾病的增加。大饥荒期间（20世纪60年代）出现的下降表明，外部冲击可能会暂时逆转长期健康趋势。这一历史轨迹表明，随着社会发展，这种差距可能会扩大，因为



y rates, suggesting better survival for both sexes but disproportionately longer lives for women. Countries with high infant mortality rates (e.g., Afghanistan, Benin) tend to have moderate-to-high life expectancy gaps, indicating thi

图8：各国预期寿命差与婴儿死亡率热力图（2023年）。各国预期寿命差与婴儿死亡率热力图（2023年）：该可视化展示了不同国家女性与男性的出生时预期寿命差（女－男），以及男性和女性的婴儿死亡率。预期寿命差的数值越高，表示女性的预期寿命优势越大；婴儿死亡率越高，则分别表明男婴和女婴面临的风险越高。颜色强度表示各项指标的幅度，颜色越深，数值越高。白俄罗斯和比利时等国的女性预期寿命差距最大，而阿富汗和贝宁等国的男女婴儿死亡率均较高。

这归因于吸烟、压力和心血管疾病等因素在男性中日益普遍。

这一趋势凸显了有针对性的公共卫生干预的重要性。女性的生理优势可能会持续存在，但通过开展关注心血管健康、心理健康和职业安全的举措，减少男性可预防死亡，可以让所有人的健康状况得到改善。中国案例表明，这一差距并非一成不变，而是会随着社会和经济变化而演变。

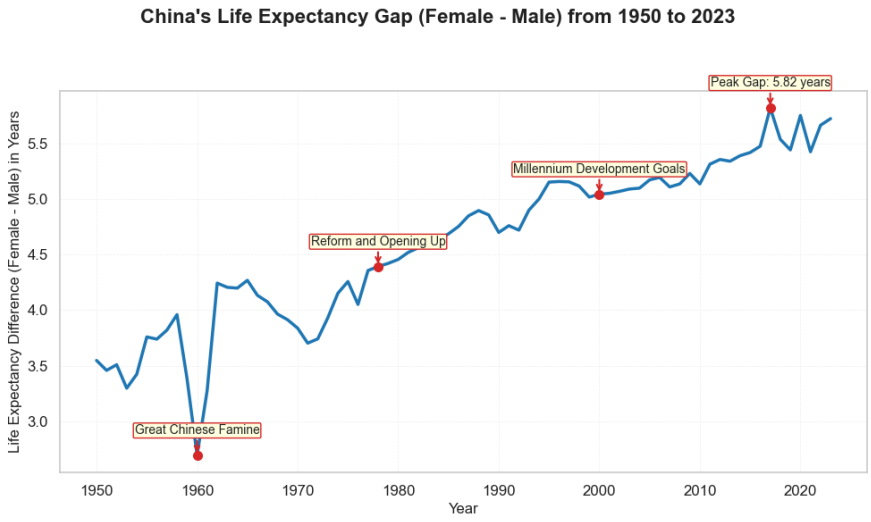


图9：中国。1950年至2023年，中国女性与男性之间的预期寿命差距稳步扩大，达到5.82年的峰值；在中国大饥荒、改革开放和千年发展目标时期等重要历史事件期间，差距出现了显著波动。

1950年至2023年中国预期寿命差距的时间序列折线图呈现出明显的上升趋势，差距从约3.5年扩大到5.6年。中国大饥荒期间（约1960年）出现的下降，表明社会政治动荡可能会暂时逆转健康方面的进展。此后的恢复和持续增长反映了医疗保健和生活水平的改善，以及男性中与生活方式相关的疾病日益增多。这一历史轨迹为了解国家政策和经济发展如何随时间塑造长寿模式提供了宝贵洞见。

最终反思：呼吁公平

总之，尽管生物学在女性寿命为何长于男性这一问题上发挥着基础性作用，尤其是女性婴儿死亡率较低，但社会决定因素

对于这一差距的幅度同样至关重要。医疗保健、教育、经济机会和性别规范都会影响两性的生存结果。在高风险地区减少男性可预防死亡，能够让所有人的人口整体更加健康，促进公平与福祉。

此处呈现的数据凸显了采取有针对性的公共卫生干预措施的必要性，以应对影响男性的特定风险因素，例如心血管疾病、事故和物质滥用。通过投资建设全面的医疗保健体系并倡导健康生活方式，社会可以努力缩小寿命性别差距，这不仅惠及男性，也造福所有人。归根结底，理解生物学与社会交织而成的复杂图景，是为子孙后代创造更健康、更公平未来的关键。

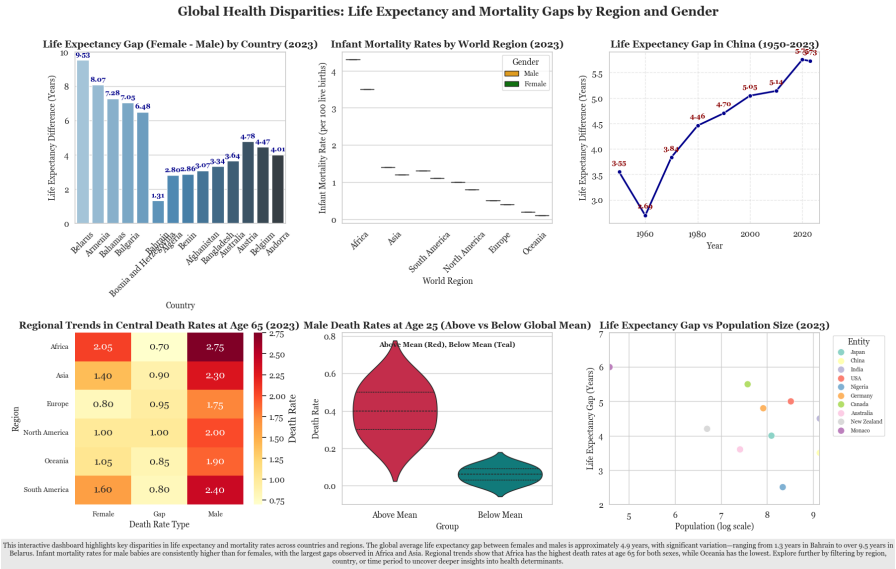


图10：全球健康差距：不同地区和性别的预期寿命与死亡率差距。全球健康差距：不同地区和性别的预期寿命与死亡率差距——这个交互式仪表板显示，全球女性的寿命比男性长，平均预期寿命差距为4.9年，从巴林的1.3年到白俄罗斯的9.5年以上不等。数据凸显了显著的地区和性别差异：所有地区的男性婴儿死亡率都较高，尤其是在非洲和亚洲; 非洲和亚洲65岁男性的死亡率最高; 而日本和摩纳哥等国家尽管人口众多，差距却较小。这些趋势表明，生物、行为和社会经济因素共同导致了寿命差距。进一步按地区、国家或时间段筛选，探索更深层的健康决定因素。

交互式仪表板全面概述了全球寿命性别差距

，总结了预期寿命差、婴儿死亡率和区域趋势等关键指标。它通过视觉呈现证实，全球女性的寿命长于男性，但各国家和地区之间存在显著差异。用户可以按地区、国家或时间段筛选数据，进一步探索仪表板，从而深入了解造成这一现象的复杂因素。最终，它为政策制定者、研究人员和公众提供了有力工具，帮助他们发现需求所在，并共同努力，为所有人创造更健康、更公平的社会。

H.2 失败案例分析

以下是一个失败案例，揭示了我们框架中一种关键的级联失效机制。该报告的可视化密度极低，整份报告中只有一张基础图表，严重限制了生成内容的分析深度和决策支持价值。可视化密度和质量低下，源于一个更深层的技术问题：在实时交互过程中，分析智能体 A'反复生成无法正常执行的可视化代码，原因包括语法错误、数据类型不匹配或库函数调用不兼容。代码执行失败时， A' 返回错误消息或占位输出，而非有效的图表图像，破坏了文本与视觉交错的呈现机制。关键的是，这些早期失败产生了累积性的负面影响，波及写作智能体 W——在多次请求可视化却未能获得预期视觉证据后，W 逐渐学会在报告生成过程的后续阶段减少甚至放弃请求新的可视化内容。这种行为适应意味着，随着报告逐步展开， W 越来越依赖纯文本描述，而没有发挥协作框架的核心优势——生成文本与视觉交错内容，最终导致报告缺乏视觉支持，分析价值有限。

双重能源危机：全球能源格局中的排放与不平等

背景铺陈：能源不公的两面

世界正处于能源演进的关键时刻，面临着定义21世纪的两场相互交织的危机：能源系统持续产生温室气体排放，以及数亿人长期无法获得现代能源服务。这些挑战不仅关乎技术或经济，更是关乎道德的迫切课题，反映出根深蒂固的全球不平等。全球能源需求持续增长，为满足这一需求提供动力的能源却仍主要来自化石燃料，助长了气候变化。与此同时，仍有超过7亿人无法获得电力，近25亿人依赖污染性燃料做饭，健康因此受到威胁，经济机会受到限制，贫困循环也因此延续。

2023年的数据揭示了能源获取不平等的严峻程度。根据数据表3，全球电力普及率为87.7%，但这一平均值掩盖了巨大的差异。各地普及率悬殊，从科摩罗的最低值11.6%，到北美洲和欧洲等地区的100%普及。类似地，数据表1显示，全球人口中仅有71.7%依靠清洁燃料做饭，非洲的比例仅为32%，而大洋洲则达到100%。这些图表数据表明，能源获取问题尚未解决; 它仍是全球不平等的显著表现，最脆弱的人群承受着最沉重的负担。

这张可视化有力地展示了能源获取方面显著的区域差异，证实全球电力获取率平均值为 87.7%、清洁烹饪获取率平均值为 71.7% 只是具有误导性的抽象概括。非洲是受影响最严重的地区，电力获取率仅为 45%，清洁烹饪燃料获取率更是只有 32%，远低于全球平均值。相比之下，欧洲、北美洲和大洋洲均实现了这两项服务的普遍覆盖，表明能源公平是可以实现的。亚洲和南美洲处于中间水平，但一个关键发现是：在每个地区，

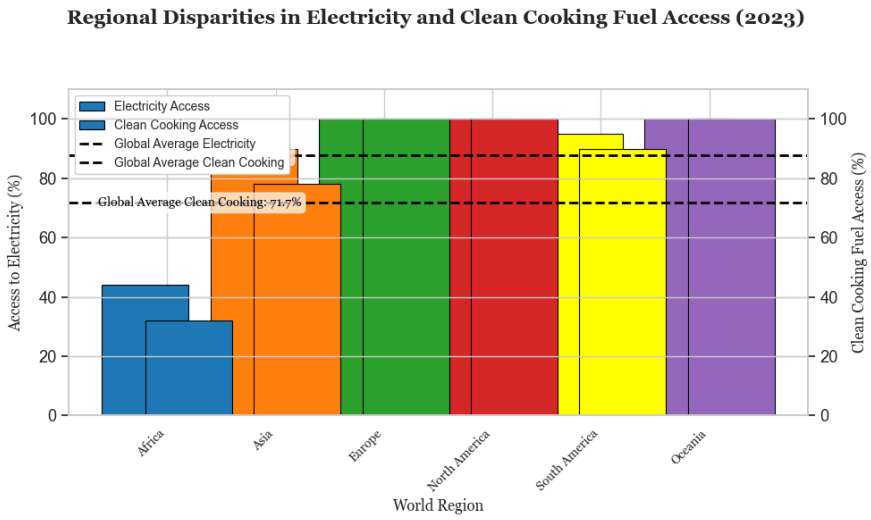


图1：电力和清洁烹饪燃料获取的区域差异（2023年）。该图展示了2023年全球各地区在电力和清洁烹饪燃料获取方面的差异，凸显了地区间的显著差距。非洲的电力获取率（45%）和清洁烹饪燃料获取率（32%）均为最低，而欧洲、北美洲和大洋洲在这两方面均接近普及（100%）。亚洲和南美洲的获取率处于中等水平，且所有地区的电力获取率均高于清洁烹饪燃料获取率。全球平均值——电力为88%，清洁烹饪为71.7%——以虚线标示，凸显了持续存在的能源不平等，以及向低排放能源系统转型所面临的更广泛挑战。

电力获取率超过清洁烹饪获取率，表明向清洁烹饪转型的进程落后于电气化。这一差距凸显出一个具体且紧迫的挑战：即使国家获得了电力，人们往往仍继续依赖有害的固体燃料烹饪，而固体燃料是室内空气污染的主要来源，也是全球温室气体排放的重要贡献因素。

揭示这场冒险：能源不平等的驱动因素与模式

经济发展与能源获取之间的关系是全球能源格局的核心驱动因素。数据表3显示，人均国内生产总值与电力获取之间存在显著的正相关关系。人均国内生产总值超过20,000美元的国家（例如北美洲和欧洲的国家）已实现100%的电力获取率。在这一阈值以下，情况变得更加复杂。在非洲和亚洲部分地区，人均国内生产总值通常低于10,000美元，电力获取率差异很大，从科摩罗的11.6%到一些国家的90%不等，这表明经济产出并不是唯一的决定因素。这种差异表明，政策选择、基础设施投资和治理在决定能源获取结果方面发挥着至关重要的作用。

存在一些显著的异常值，例如加勒比地区的小岛屿国家或其他国家，尽管人均国内生产总值相对较低，却实现了100%的电力获取率。这表明，有针对性的政府投资、区域合作，或采用具有成本效益的可再生能源技术，可以克服经济制约，实现普遍的能源获取。对于清洁烹饪，数据表1显示其与人均国内生产总值也存在类似但不那么显著的正相关。非洲的清洁烹饪获取率再次最低，仅为32%；而大洋洲、北美洲和欧洲则接近100%。然而，人均国内生产总值与清洁烹饪获取之间的差距更为显著，这表明这一转型更加复杂，可能需要采取超越一般经济发展的具体干预措施，例如为清洁炉灶提供补贴，或发展本地沼气系统。

能源结构面临的挑战同样显而易见。数据表 2显示，1985年到2024年期间，化石燃料一直是全球发电的主要来源，平均发电量为12,090 TWh。尽管可再生能源快速增长，从2,058 TWh增至9,847 TWh，核能也从1,488 TWh增加到2,764 TWh，但化石燃料发电量从6,285 TWh增至18,240 TWh，绝对增长幅度十分显著。这表明，全球能源系统正在扩张，但这种增长是由化石燃料而非清洁替代能源推动的，这一趋势与气候目标不相容。

隐性代价：能源不平等与碳排放

能源获取方面的差距与全球碳排放危机密不可分。数据表0揭示了经济产出与排放之间的复杂关系。尽管人均国内生产总值与人均二氧化碳排放之间总体呈正相关——收入较高的国家往往排放量也较高——但这种关系远非线性，并且受到区域能源系统和工业结构的显著影响。全球人均排放量为5.87吨，但各地差异极大，从一些非洲国家的0.25吨到某个高排放亚洲国家的19.34吨不等。

最引人注目的现象是存在显著的异常值。例如，亚洲某个国家的人均国内生产总值为 76,600 美元，人均二氧化碳排放量为 19.34 吨；而亚洲另一个人均国内生产总值相近的国家，排放量却低得多。这表明，能源来源的选择和工业效率是关键因素。同样，人均国内生产总值适中（约 60,000 美元）的欧洲国家，排放量也可能很高（16–17 吨），这说明能源和工业领域的高碳强度，即便在人均收入不高的情况下，也会推高排放量。

相比之下，许多非洲国家的人均国内生产总值虽然很低，排放量也非常低（0.25–6 吨），这体现了它们对化石燃料的依赖程度较低。然而，排放水平低并不代表环境表现优异，而是反映了能源贫困。非洲大多数人口无法获得现代能源，只能依赖生物质能和其他传统燃料；这些燃料不仅有害健康，而且效率低下，从长远来看也难以持续。

清洁能源获取与减排之间的联系十分明确。已实现电力和清洁烹饪燃料近乎普及的国家——北美洲、欧洲和大洋洲——人均排放量也位居最高之列。这是因为这些国家的能源系统虽然现代化，却往往仍然严重依赖化石燃料。真正的解决方案在于将能源获取与碳排放脱钩。目标必须是利用可再生能源和低碳能源，为所有人普及现代能源服务——电力和清洁烹饪。这需要在全球范围内对可再生能源基础设施进行大规模投资，尤其是在低收入地区，从而确保各国在发展过程中能够跨越化石燃料阶段。

前进之路：公正能源转型

这些数据清晰而紧迫地呼吁开展公正且公平的能源转型。当前的模式——高收入国家维持高排放，而低收入国家却在为基本的能源获取而苦苦挣扎——既不可持续，也极不公正。解决方案在于开展全球性行动，扩大可再生能源部署，重点关注最需要帮助的地区。

首先，我们必须优先确保所有人都能获得电力和清洁烹饪服务。这需要

为农村和偏远社区有针对性地投资分布式可再生能源系统，例如太阳能家庭系统和微型电网。此外，还需要开发价格可负担的清洁烹饪技术，例如改良炉灶和沼气系统，并通过补贴和宣传活动予以支持。

其次，高排放国家必须大幅降低其碳强度。这需要迅速淘汰煤炭和其他化石燃料，同时大规模投资风能和太阳能、能源储存以及电网现代化。如数据表2所示，可再生能源的增长必须加快，超过化石燃料的增长速度。

最后，国际社会必须弥补资金缺口。向清洁能源未来转型需要数万亿美元的投资。发达国家有道义义务向发展中国家提供资金和技术支持，使其能够建设现代化的低碳能源系统，而不必重蹈过去高排放的覆辙。

21世纪的冒险不仅关乎技术创新，更在于创造一个让能源获取成为普遍人权、让气候行动成为共同全球责任的世界。数据揭示了挑战的规模，也展现了实现更加公平、可持续未来的可能性。这段征程始于正视现行体系的不公，并致力于为所有人实现大胆、包容且公正的能源转型。

报告大纲生成提示

任务
围绕用户意图生成一份引人入胜的数据报告大纲。

任务详情
- 结合数据摘要，按照线性叙事结构生成报告大纲。

- 线性叙事结构是指包含开端、中段和结尾的叙述结构。可以把它想象成铺陈背景、展开冒险，并以令人满意的结论收尾。

- 大纲中的每个要点都应细分为更小的子要点，以突出数据的具体方面。这些方面可能包括：重要数据或模式、值得注意的例外情况或偏差，以及随时间推移的比较或变化。必要时添加可视化（例如图表）的说明。

- 数据报告的总体主题应聚焦于 {user_意图}。请确保这一倾向
在整个大纲中保持一致。
- 请记住，一份引人入胜的数据报告的精髓不仅在于数字，还在于你如何讲述，
因此，纳入可视化说明至关重要。
- 要具体、清晰，最重要的是要有吸引力。生成的大纲必须连贯地
并且在逻辑上与数据的属性相关联。请尽可能具体。

用户意图
{user_意图}

输出格式
请按照以下结构，在单个 Markdown 代码块中生成大纲：
"""
数据报告标题

<与用户意图一致的章节标题>
- 涵盖数据特定方面的要点
- ...

...
- ...
"""

图9：报告撰写器使用的报告大纲生成提示。

增量报告撰写提示

任务
你的任务是根据提供的大纲生成一份全面的数据分析报告。

任务详情
1. **聚焦用户意图**：报告主题必须与提供的用户意图保持一致：
{user_intent}。
2. **遵循大纲**：报告必须严格遵循其结构和叙述流程
在大纲中定义 .
3. **详细阐述各个部分**：基于大纲，针对每个部分展开详细分析，提炼洞见，
并撰写清晰、专业的正文
4. **可视化请求**：需要可视化时，根据数据摘要生成一条简短的可视化
请求，并将其置于 '<visualization></visualization>' 标签中。
"可视化请求"是指用自然语言纯文本说明需要进行
用于支持分析的可视化。'<visualization></visualization>' 标签中的内容将
提示数据分析师智能体创建图表。
5. **可视化结果**：数据分析师智能体生成图表后，某些结果会
附加在 '<visualization_result></visualization_result>' 标签中，其中包含图表图像
，紧随可视化请求标签 '<visualization></visualization>' 之后。
6. **续写行为**：每个 '<visualization_result></visualization_result>' 标签之后，
请用 2–4 句话继续撰写报告，提供分析、洞见和清晰、专业的正文。

用户意图
{user_意图}

大纲
{outline}

可视化格式
在'<可视化></可视化>'中用自然语言简洁地描述可视化请求。
只需说明可视化目标和要传达的信息。
示例：
"""
<可视化>
比较各地区 2024年 的月均销售额，以找出表现最佳和较差的地区
。重点关注地区名称和月均销售额；排除少于 10 个的地区
2024年的记录。
</可视化>
"""

输出格式
请仅以 Markdown 格式输出报告文本，不要附加任何说明。

图10：报告撰写器使用的增量报告撰写提示。

可视化规范规划提示

你是一名专业的数据分析师和图表设计师。

任务
- 你的任务是分析用户输入，并生成一份包含必要
信息 .
- 可视化描述必须放在 <visualization></visualization> 标签中。

{chart_style}

数据表
数据表提供用于可视化的原始数据。

{summaries}

{tables}

输出格式
输出的可视化描述必须严格遵循以下 YAML 格式：
"""
图表_类型： <图表类型>
标题： <可视化的标题>
数据： <所有待可视化的数据>
标签： <坐标轴标签、图例及其他文本标签的描述>
"""

图 11： 数据增强分析智能体 A使用的可视化规范规划提示 P_{ana} ′.

图注生成提示

任务 您的任务是为提供的图片生成简洁且具有描述性的说明文字，图片标题为{title}，用户意图为{user_intent}。提供的图片是根据“可视化请求”生成的图表：{visualization_request}您应根据“可视化请求”准确、清晰地生成说明文字。## 输出格式 仅以纯文本输出说明文字，不作任何解释。

图 12： 数据增强分析智能体使用的面向分析的标题生成提示 $P_{caption}$ A ′.

可视化规范示例 1

可视化请求：
使用标题为“全球吸烟趋势：2000—2022年”的折线图展示2000年至2022年的全球吸烟率趋势，呈现明显的下降走势。标注关键数据点：2000年为34.3%，平均值为26.7%，2022年为21.7%。注明22年间下降了12.6个百分点。

可视化规范：图表_类型：折线图_图表标题：全球吸烟趋势：2000—2022年 数据：- 年份：2000年 吸烟或使用烟草的_成年人_比例（年龄标准化）：34.322533 - 年份：2005年 吸烟或使用烟草的_成年人_比例（年龄标准化）：30.745779 ... - 年份：2021年 吸烟或使用烟草的_成年人_比例（年龄标准化）：22.133286 - 年份：2022年 吸烟或使用烟草的_成年人_比例（年龄标准化）：21.748644 标签：横轴标签：年份_ _纵_轴_标签：吸烟或使用烟草的成年人比例（年龄标准化，%） 注释：- 文本：“2000年为34.3%” 位置： [2000, 34.322533]颜色：“#000000” 字号：10_箭头：是 - 文本：“平均值：26.7%” 位置： [2012, 26.677333]颜色：“#000000” 字号：10_箭头：是 - 文本：“2022年为21.7%” 位置： [2022, 21.748644]颜色：“#000000” 字号：10_箭头：是 - 文本：“22年间下降12.6个百分点” 位置： [2010, 20]颜色：“#000000” 字号：10_对齐方式：居中 图例：无

图13：折线图可视化规范示例。

可视化规范示例 2

可视化请求：
创建一张水平条形图，标题为“各国香烟税收（2022年）”，展示税收占比最高的前10个国家（如保加利亚、波斯尼亚和黑塞哥维那、巴西）以及最低的后10个国家（如文莱、贝宁、安提瓜和巴布达）。在全球中位数53.6%处添加一条竖直虚线。

可视化规范：
图表_类型：水平_条形_图表
标题：“各国香烟税收（2022年）”
数据：
- 实体：巴西
税款_占_香烟_价格_的_比例_：80.25
- 实体：澳大利亚
税款_占_香烟_价格_的_比例_：76.98
...
- 实体：巴巴多斯
税款_占_香烟_价格_的_比例_：42.99
- 实体：阿尔及利亚
税收_占_香烟_价格_的_比重：44.16
标签：
x_轴：“税收占香烟价格的比重（%）”
y_轴：“国家”
中位数_线：“全球中位数（53.6%）”
图例：null

图14：水平条形图的可视化规范示例。