arXiv:2512.19675v1 [econ.GN] 22 Dec 2025

用于从档案图像扫描件构建历史数据集的多模态大语言模型：德国专利（1877年–1918年）*†‡

Niclas Griesshaber§
曼海姆大学
niclas.griesshaber@history.ox.ac.uk

Jochen Streb
曼海姆大学
jochen.streb@uni-mannheim.de

摘要

我们利用多模态大语言模型（LLM），通过由 Gemini-2.5-Pro 和 Gemini-2.5-Flash-Lite 驱动的基于大语言模型的流程，从 9,562 份档案图像扫描件中构建了一个包含 306,070 项德国专利（1877年–1918年）的数据集。我们的基准测试提供了初步证据，表明多模态大语言模型能够创建比研究助理质量更高的数据集，同时在从图像语料库构建专利数据集时，速度快 795 倍以上，成本低 205 倍。每页以双栏格式排印，采用哥特体和罗马体，包含约 20 至 50 条专利条目。一手史料中字体和版式的复杂性使我们认为，多模态大语言模型将彻底改变经济史领域的数据集构建方式。我们将基准测试数据集、专利数据集以及基于大语言模型的数据处理流程开源；借助大语言模型辅助编码工具，该流程可以轻松适配其他图像语料库，从而降低技术经验较少的研究者的参与门槛。最后，我们阐释了将大语言模型用于历史数据集构建的经济学原理，并通过推测其对经济史领域的潜在影响作结。

关键词 多模态　　大语言模型 · 信息抽取 · 数据集构建 · 德国专利　　ts

1 引言

长期以来，经济史学家一直专注于编制和解读宏观经济时间序列，其中人均国内生产总值最受关注（Maddison，2006年；Broadberry 等，2015）。这不可避免地造成了方法上的局限，因为要了解消费者和生产者如何应对经济政策变化或外生冲击，就需要微观经济数据。尤其重要的是，要考虑历史行动者的异质性。人们在年龄、性别、教育、社会地位、收入和财富等方面各不相同，因此，在相同的外部条件下，他们会做出不同的决策。要将这种异质性纳入考量，就需要大规模的微观经济数据集。

经济史学家很少使用微观经济数据，并非因为这些数据难以获取。例如，消费者的社会经济数据可以从历史人口普查数据（Long & Ferrie，2013；Ruggles，2014；Abramitzky 等，2021）、储蓄簿（Lehmann-Hasemeyer & Streb，2018）中获取；公司的相关信息则可从专利

*论文网站：https://historymind.ai
†代码：https://github.com/niclasgriesshaber/llm_专利_处理流程.git
‡档案图像语料库：https://digi.bib.uni-mannheim.de/sammlungen/patentregister§通讯作者。现供职于牛津大学。

统计资料（Donges & Streb，2024年）、股票市场报纸（Lehmann-Hasemeyer & Opitz，2024）、城市名录（Albers & Kappner，2023）或贸易登记册（Guinnane 等，2007）中获取。然而，获取这些数据的成本很高，有时甚至高到令人望而却步，因为由研究助理手动构建数据集既缓慢、昂贵，又容易出错。为加快历史数据集构建，数字人文和计算机科学领域的方法，如光学字符识别（OCR）、命名实体识别（NER）和信息抽取（IE），在经济史领域日益受到青睐（Shen 等，2020年；Dell 等，2023；Bergeaud & Verluise，2024）。然而，这些数字化方法需要高阶编程技能和标注数据，才能构建仅适用于同质图像语料库的定制处理流程。人工智能（AI）的最新进展有望消除这些障碍，让非技术背景的研究人员也能开展大规模数据集构建，从而消除该领域的主要瓶颈：将档案图像扫描件转化为结构化数据集。

多模态大语言模型（大语言模型）是能够联合处理多种类型数据的神经网络。例如，在本文中，我们将图像和原始文本指令输入多模态大语言模型，以构建大规模历史专利数据集。多模态大语言模型在语言和视觉等广泛任务中展现出卓越性能（Devlin et al., 2019年；Brown et al., 2020年；Dosovitskiy, 2020年）。与许多现有的光学字符识别、命名实体识别和信息抽取方法一样，它们基于深度学习范式；该范式指由许多层组成、能够从数据中学习复杂模式的神经网络（LeCun et al., 2015；Goodfellow et al., 2016）。随着数据、参数和计算能力的规模不断扩大，大语言模型会变得异常强大（Kaplan et al., 2020；Hoffmann et al., 2022；Wei et al., 2022），为光学字符识别、命名实体识别和信息抽取提供统一的方法。这种规模经济也是最强大的大语言模型几乎都由私营人工智能公司开发的原因。

多模态大语言模型的进步及其商业化，使我们得以作出多项贡献。首先，为了构建一个包含 1877年 至 1918年 期间出版的 41 卷历史专利登记册实体卷册中全部 306,070 条专利条目的数据集，我们使用大语言模型辅助编码工具开发了一个基于大语言模型的数据管道，处理了 9,562 份图像扫描件。随后，我们从每条专利条目中提取了五个变量：专利_编号、受让人、地点、标题和日期。此外，我们的研究助理从所用年度专利登记册中各随机选取一幅图像，共 41 幅，手动创建了一个学生构建的基准数据集。起初，我们将大语言模型生成的数据集与学生构建的数据集进行比较，以评估基于大语言模型的流程的可靠性，尤其是检查是否存在会损害数据质量并使后续计量经济学分析产生偏差的幻觉。令人意外的是，我们发现两个数据集中许多被指出的差异并非源于大语言模型的错误，而是研究助理的失误。因此，我们还构建了第二个完美基准数据集。通过将学生构建的数据集和大语言模型生成的数据集与完美数据集进行比较，我们提供了初步证据，表明多模态大语言模型能够根据我们的图像语料库生成比研究助理质量更高的数据集。最后，我们将基于大语言模型的数据管道和新的历史专利数据集开源，以推动经济史领域的实证研究。

本文结构如下：第2节介绍数据的历史背景。第3节回顾专利数据库构建方面的相关工作，包括人工处理、自定义机器学习方法，以及经济史领域关于大语言模型的新兴研究。第4节介绍构成我们档案图像语料库的一手史料。第5节阐述我们用于构建专利数据集的基于大语言模型的数据管道；第6节则使用我们的基准数据集探讨数据质量和幻觉问题。第7节解释使用多模态大语言模型从图像扫描件构建数据集的经济学。第8节总结全文，并讨论多模态大语言模型对经济史领域可能产生的影响。

2 历史背景

现代经济增长始于19世纪初的大不列颠，此后为世界各地许多国家所效仿，其基础是源源不断的创新。因此，经济史中的一个核心问题是，为什么创新花了这么长时间才从罕见现象演变为持续推动增长的动力（Galor，2011年；Mokyr，2011年）。要了解哪些因素促进或阻碍创新的出现，首先必须尽可能准确地衡量创新的类型与范围。在经济史研究中，这主要通过专利统计资料来完成。

专利制度的起源至少可以追溯到15世纪中叶的威尼斯。然而，直到18世纪末，专利才从一种特权演变为可通过向专利主管机关注册而获得的知识产权。所有专利主管机关在授予专利时都会收取申请费；有些机关还会将已注册发明的新颖性作为一项额外要求进行审查。从那时起，专利持有人便有权在专利法适用的地理范围内，于有限期限内独家制造和销售受保护的创新。要使专利保护发挥作用，潜在仿制者必须了解哪些创新受到保护。因此，专利主管机关长期以来一直公布已授权专利的信息，也收集详细的专利说明书，以便相关方在需要时查阅（Moser，2011年；Cox，2019年）。这些历史资料至今仍可获取，可据此建立数据库，记录有关发明人及共同发明人、其性别和居住地，以及专利数量、标题和技术类别等微观经济信息。

这批数据相对容易获取，这是将专利作为衡量技术进步范围和方向的指标的一大优势。对此方法持批评态度的人喜欢指出，并非所有创新都会申请专利：有些创新不符合专利法的规定，有些发明人则认为，通过保密更有可能获得创新收益。此外，单纯统计专利数量还忽略了这样一个事实：已授权专利中，只有少数保护了有价值的创新；大多数专利所代表的想法几乎毫无价值，难以称为渐进式创新（Griliches，1990；Moser，2012）。第一项批评指出，专利统计作为衡量创新的指标存在一个无可否认的缺陷，这一问题至今尚未得到解决，但大多数经济史学家都接受这一点。Petra Moser 提出的替代方法是借助世界博览会数据识别历史上的创新，但这种方法也有自身的缺陷（Moser，2005；Domini，2020）。第二项批评如今已不那么重要，因为经济史学家已找到多种评估专利质量的方法（Streb，2024）。

3 相关工作

3.1 手工构建数据集

手工从一手史料中构建数据集极其耗费人力。过去几十年来，人们手工创建了大量历史专利数据库。这些数据库现已覆盖早期工业化国家大不列颠、法国、美国和荷兰，地中海国家意大利和西班牙，斯堪的纳维亚国家芬兰、挪威和瑞典，日本，甚至一些拉丁美洲国家，如阿根廷、古巴和墨西哥（Streb，2023）。

2000年代中期，Streb等人（2006）决定为德意志帝国和魏玛共和国建立一个历史专利数据库。然而，他们没有足够的资金聘请所有研究助理，而要将1877年至1932年间德国已授权的五十多万项专利手动录入数字数据库，就需要这些研究助理。因此，经济史学家不得不放弃追求完整性，将数据库限定为仅收录最有价值专利的一个子集。由于德国专利制度征收专利年费，旨在促使专利持有人尽快放弃无利可图的专利，而非将其保留至最长15年的期限，因此可以根据专利的存续期推断其经济价值。他们选取了自授权日期起至少持有十年的专利纳入数据库。最终形成的数据库包含约68,700项长期存续的专利，约占德意志帝国和魏玛共和国所有已授权专利的10%。尽管数据量大幅减少，手动提取所有相关信息仍耗费了数月时间，这充分体现了手工构建数据集在可扩展性方面的局限。

3.2 定制化光学字符识别与命名实体识别处理流程

为克服手工构建数据集的局限，近期研究采用了机器学习领域中针对特定任务的狭义模型。在专利数据库构建方面，“PatentCity”代表了这一范式下的先进方法。Bergeaud & Verluise (2024) 使用传统的 OCR 算法，并训练自己的深度学习模型

来构建专利数据库。由于从手工方法转向数字化方法，他们能够收集的数据量超过了以往所有研究。他们的数据库包含德国 1877年 至 1980 年、法国 1903 年至 1980 年、英国 1893 年至 1980 年以及美国 1836 年至 1980 年间所有已授权专利的受让人和发明人信息。在有注明的情况下，还记录了发明人的职业、公民身份和地址。此外，他们将发明人的地址归属到县或市镇。

为构建 PatentCity 数据库，作者首先使用 Tesseract v5.0 和自行开发的光学字符识别算法，将以图像扫描件形式保存的历史专利说明书转换为可编辑的文本文件。第二步，他们训练了一个定制的命名实体识别模型，从光学字符识别文本中提取相关变量，并应用关系算法来重建实体之间的相互关系。为训练命名实体识别模型，作者必须创建一个手动标注的数据集，并将其划分为训练集和测试集。前者用于教会深度学习模型如何提取所需实体。训练完成后，他们通过将模型在测试集上的预测结果与真实标注进行比较，评估模型的性能。

然而，这类狭窄的机器学习方法也有自身的瓶颈。Bergeaud & Verluise (2024) 提出的方法并不具有普适性；每当图像扫描件的版面发生变化，他们就必须重新训练命名实体识别模型，因为光学字符识别文本的结构也会随之改变。这总是需要一个新的标注数据集。此外，进行光学字符识别时，图像中嵌入文本的大量空间关系信息会丢失；而在从描绘双栏布局、表格或地图的复杂历史文献中构建数据集时，这些信息至关重要。最后，他们首次发表这项工作时，大语言模型辅助编码工具尚不存在，因此非技术背景的研究人员很难将其处理流程应用于其他图像语料库。

多模态大语言模型的出现或许能为处理历史文献提供一种通用解决方案，因为它可以缓解上述所有问题。这使我们得以使用不同的一手史料，部分复现 PatentCity 数据库中 1877年至1918年德国专利的数据。我们的史料并非每份图像扫描件都展示一项专利的专利说明书，而是每页包含 20 至 50 条简短的专利条目，并采用双栏格式排版。此外，我们还进一步补充了 PatentCity 数据库未收录的所有专利标题。

3.3 经济史中的大语言模型

目前，经济学和经济史领域关于将大语言模型用于文本即数据方法的研究正逐渐兴起（例如，Bartik 等（2025）；Chyn 等（2025）；Griesshaber & Ogilvie（2025）；Lagakos 等（2025）；Sockin 等（2025））。现有研究利用大语言模型从可编辑文本中提取更高层次的特征，以构建用于下游计量经济学分析的变量。例如，Griesshaber & Ogilvie（2025）使用 GPT-4o 将句子归类为更宽泛的经济类别，以考察殖民时期的拉丁美洲行会制度内部的制度差异。Lagakos 等（2025）从传记中提取特征，以了解20世纪初的美国人过上有意义人生的根源。在这两种情况下，大语言模型都带来了前所未有的生产力提升，因为经济史学家和研究助理不必再阅读并手动分类数千篇传记和行会规章。然而，由于大语言模型是在有限的数据集上训练的，它们不可避免地会做出带有偏差的预测，尤其是在预测涉及价值判断时。Carlson & Dell（2025）表明，这种偏差可能会传递到下游估计量中。他们还强调，研究人员可以使用具有不同偏差的多种大语言模型，在回归分析中获得显著结果。为此，作者提出了 MAR-S 框架，其中包括构建验证数据集，以纠正大语言模型在特征提取过程中引入的偏差。

与这些文本即数据方法相比，仅仅转录图像扫描件中的印刷体、哥特体或手写文本时，价值判断通常不那么突出。我们可以向多模态大语言模型提供图像和文本指令，让模型转录图像中的文本。影响大语言模型转录准确率的因素有很多：字体类型、图像尺寸、信息密度、分辨率、文档劣化程度、提示词等等。档案来源的巨大异质性会影响所有这些因素，因此很难评估多模态大语言模型从图像扫描件转录文本的一般可靠性。Greif 等人（2025年）对 GPT-4o、Gemini-2.0-Flash 和 Transkribus 的 Text Titan I 在十八至十九世纪德国城市名录上的光学字符识别能力进行基准测试；这些名录采用罗马体和哥特体印刷。他们报告称，Gemini-2.0-Flash 的转录结果最佳。Humphries 等人（2025年）发现，大语言模型转录十八至十九世纪英文手写文档的准确率高于 Transkribus 的微调模型

。Levchenko（2025年）在使用 Civil 字体印刷的俄语印刷品上也报告了这一结果；在评估的 12 种多模态大语言模型中，Gemini-2.5-Pro 的表现最佳。Crosilla 等人（2025年）表明，大语言模型在现代手写体识别方面优于 Transkribus。比多模态大语言模型具备普遍的转录能力更令人惊讶的是，我们至今仍未完全理解它们如何完成这项任务。因此，新兴的机制可解释性子领域正试图揭示这种能力背后的内部机制（Baek 等人，2025年；Kim 等人，2025年）。

多模态大语言模型能够转录图像中的嵌入文本，这一能力固然令人瞩目，但更引人注目的是，我们可以通过提示词让它们只提取所需信息，从而构建结构化数据集。直观而言，强大的转录能力是完成这项任务的先决条件。经济史学家 Jonathan Jayes（2025）使用 Gemini-2.0-Flash 从瑞典企业报告的图像扫描件中提取信息。随着越来越多的经济史学家使用多模态大语言模型创建大规模数据集，严格评估多模态大语言模型在信息抽取任务上的表现，进而评估其在历史数据集构建任务上的表现，将成为必要之举。Xie 等人（2025，第1页）发现，GPT-4o 在从瑞典专利卡片（1945—1975年）中提取信息时，“生成的数据可用，但并不完美”。Luo（2025）通过在同一样本上对若干开源和专有大语言模型进行基准测试，拓展了关于瑞典专利卡片信息抽取的研究；其中 Gemini-2.5-Pro 的表现最佳。同样，Vafaie 等人（2025）的平行研究评估了 GPT-4o-mini 以及 InternVL2.5 等开源模型，在一个由异质德国索引卡片构成的新标注数据集上进行关键信息抽取的表现。然而，所有这些模型的准确率都不足以支持严谨的经济史研究。此外，Rodrigues 等人（2025）的一份进展报告发现，在为双栏印刷书目创建 JSON 条目时，Gemini-2.5-Flash 的表现优于 GPT-4o。最后，Bäcker-Peral 等人（2025）开发了一套基于大语言模型的处理流程，从历史表格中提取信息以构建面板数据集。他们还手动构建了一个数据集，并表明，无论使用大语言模型数据还是金标准数据，其下游回归分析“在统计上都无法区分”（Bäcker-Peral 等人，2025，第1页）。他们声称，多模态大语言模型为“历史表格的数字化带来了划时代的变革”（Bäcker-Peral 等人，2025，第1页）；考虑到他们使用的是 Claude-3.5-Sonnet 和 Gemini-1.5-Pro——其能力远弱于本文所用模型的多模态大语言模型——这一点更令人瞩目。[1]

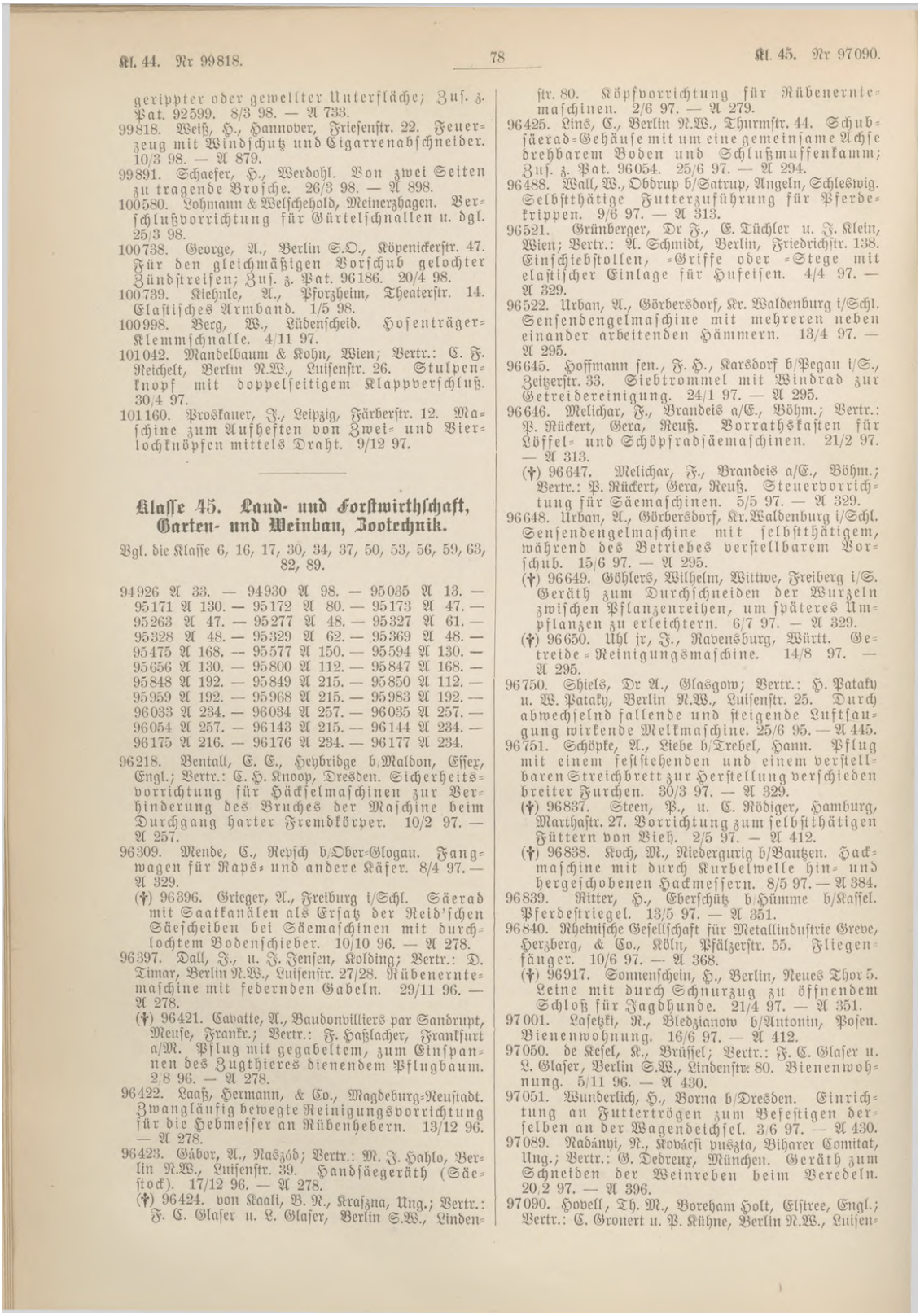
4 数据

德国直到1877年才颁布全国统一的专利法，取代此前德国各邦施行的众多独立专利法（Donges & Selgert，2019年）。为让感兴趣的公众了解新授予的专利，帝国专利局每年出版一卷，列出上一日历年登记在册的所有专利。这份年度专利登记册并未按专利号顺序列出新授予的专利，而是依据专利局采用的89种不同技术类别排列。首先按专利号顺序列出第1类“矿石制备”中的所有专利，接着列出第2类“烘焙”中的所有专利（例如，参见 Donges & Streb，2024年，第312页），列表末尾始终是第89类“制糖”中的专利。图1展示了我们语料库中的一页代表性内容。每项专利都列出了若干信息：专利号、专利持有人的名称和地址；对于外国专利持有人，还列有其在德国的法律代理人的名称和地址；专利名称；以及申请日期，少数情况下还附有优先权主张（图2）。

我们成功找到了帝国专利局从1877年到1918年间发行的所有卷册。总计收集了41本实体书籍，这些书籍由曼海姆大学研究数据中心以每张图像300 dpi的分辨率扫描。随后进行了完整性检查，以确保没有任何页面被意外遗漏，然后将TIFF图像合并为PDF。对于我们的数据集，我们只关注每卷中名为“系统概览”的章节，该章节列出了上述新授予的专利。该章节的版式在全部41卷中相当一致。专利条目始终是简短的段落，并以双栏格式排列。同样，89个技术类别中的每个类别都以标题的形式出现在两栏主要内容栏之一的开头。最后，每页的页眉都标有当前专利号和技术类别。

[1]基于这项工作，费城联邦储备银行的研究人员同仁发布了一份同期报告，探讨多模态大语言模型在历史数据收集方面的机构潜力（Moulton & Severen，2025年）。

图1 主要来源



Kl. 44. Nr 99818. 78 Kl. 45. Nr 97090.

geripptter oder gewellter Unterfläche; Zuſ. z. Pat. 92599. 8/3 98. — A 733.
99818. Weiß, H., Hannover, Friesenſtr. 22. Feuerzeug mit Windſchutz und Cigarrenabſchneider. 10/3 98. — A 879.
99891. Schaefer, H., Werdohl. Von zwei Seiten zu tragende Broſche. 26/3 98. — A 898.
100580. Lohmann & Welſchehold, Meinerzhagen. Verſchlußvorrichtung für Gürtelſchnallen u. dgl. 25/3 98.
100738. George, A., Berlin S.O., Köpenickerſtr. 47. Für den gleichmäßigen Vorſchub gelochter Zündſtreifen; Zuſ. z. Pat. 96186. 20/4 98.
100739. Kiehnle, A., Pforzheim, Theaterſtr. 14. Elaſtiſches Armband. 1/5 98.
100998. Berg, W., Lüdenſcheid. Hoſenträger-Klemmſchnalle. 4/11 97.
101042. Mandelbaum & Kohn, Wien; Vertr.: C. F. Reichelt, Berlin N.W., Luiſenſtr. 26. Stulpenknopf mit doppelſeitigem Klappverſchluß. 30/4 97.
101160. Proskauer, J., Leipzig, Färberſtr. 12. Maſchine zum Aufheften von Zwei- und Vierlochknöpfen mittels Draht. 9/12 97.

Klaſſe 45. Land- und Forſtwirthſchaft, Garten- und Weinbau, Zootechnik.

Vgl. die Klaſſe 6, 16, 17, 30, 34, 37, 50, 53, 56, 59, 63, 82, 89.

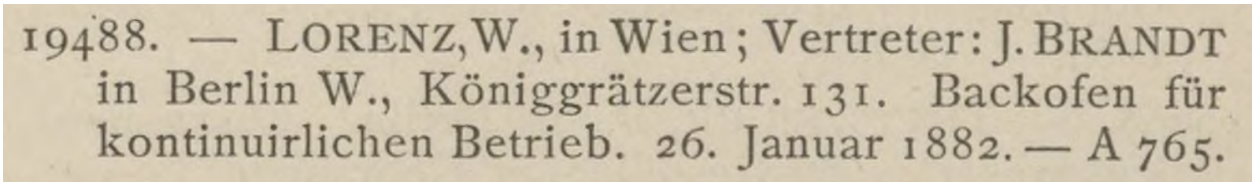
94926 A 33. — 94930 A 98. — 95035 A 13. — 95171 A 130. — 95172 A 80. — 95173 A 47. — 95263 A 47. — 95277 A 48. — 95327 A 61. — 95328 A 48. — 95329 A 62. — 95369 A 48. — 95475 A 168. — 95577 A 150. — 95594 A 130. — 95656 A 130. — 95800 A 112. — 95847 A 168. — 95848 A 192. — 95849 A 215. — 95850 A 112. — 95959 A 192. — 95968 A 215. — 95983 A 192. — 96033 A 234. — 96034 A 257. — 96035 A 257. — 96054 A 257. — 96143 A 215. — 96144 A 234. — 96175 A 216. — 96176 A 234. — 96177 A 234.

96218. Bentall, E. E., Heybridge b. Maldon, Eſſex, Engl.; Vertr.: C. H. Knoop, Dresden. Sicherheitsvorrichtung für Häckſelmaſchinen zur Verhinderung des Bruches der Maſchine beim Durchgang harter Fremdkörper. 10/2 97. — A 257.
96309. Mende, C., Repſch b. Ober-Glogau. Fangwagen für Raps- und andere Käfer. 8/4 97. — A 329.
(†) 96396. Grieger, A., Freiburg i/Schl. Säerad mit Saatkanälen als Erſatz der Reid'ſchen Säeſcheiben bei Säemaſchinen mit durchlochtem Bodenſchieber. 10/10 96. — A 278.
96397. Dall, J., u. J. Jenſen, Kolding; Vertr.: D. Timar, Berlin N.W., Luiſenſtr. 27/28. Rübenerntemaſchine mit federnden Gabeln. 29/11 96. — A 278.
(†) 96421. Cavatte, A., Baudonvilliers par Sandrupt, Meuſe, Frankr.; Vertr.: F. Haßlacher, Frankfurt a/M. Pflug mit gegabeltem, zum Einſpannen des Zugthieres dienendem Pflugbaum. 2/8 96. — A 278.
96422. Laaß, Hermann, & Co., Magdeburg-Neuſtadt. Zwangläufig bewegte Reinigungsvorrichtung für die Hebmeſſer an Rübenhebern. 13/12 96. — A 278.
96423. Gábor, A., Naszód; Vertr.: M. J. Hahlo, Berlin N.W., Luiſenſtr. 39. Handſäegeräth (Säeſtock). 17/12 96. — A 278.
(†) 96424. von Kaali, B. N., Kraszna, Ung.; Vertr.: F. C. Glaſer u. L. Glaſer, Berlin S.W., Lindenſtr. 80. Köpfvorrichtung für Rübenerntemaſchinen. 2/6 97. — A 279.
96425. Lins, C., Berlin N.W., Thurmſtr. 44. Schubfäerad-Gehäuſe mit um eine gemeinſame Achſe drehbarem Boden und Schlußmuffenkamm; Zuſ. z. Pat. 96054. 25/6 97. — A 294.
96488. Wall, W., Obdrup b/Satrup, Angeln, Schleswig. Selbſtthätige Futterzuführung für Pferdekrippen. 9/6 97. — A 313.
96521. Grünberger, Dr F., E. Tüchler u. J. Klein, Wien; Vertr.: A. Schmidt, Berlin, Friedrichſtr. 138. Einſchiebſtollen, -Griffe oder -Stege mit elaſtiſcher Einlage für Hufeiſen. 4/4 97. — A 329.
96522. Urban, A., Görbersdorf, Kr. Waldenburg i/Schl. Senſendengelmaſchine mit mehreren nebeneinander arbeitenden Hämmern. 13/4 97. — A 295.
96645. Hoffmann ſen., F. H., Karsdorf b/Pegau i/S., Zeitzerſtr. 33. Siebtrommel mit Windrad zur Getreidereinigung. 24/1 97. — A 295.
96646. Melichar, F., Brandeis a/E., Böhm.; Vertr.: P. Rückert, Gera, Reuß. Vorrathskaſten für Löffel- und Schöpfradſäemaſchinen. 21/2 97. — A 313.
(†) 96647. Melichar, F., Brandeis a/E., Böhm.; Vertr.: P. Rückert, Gera, Reuß. Steuervorrichtung für Säemaſchinen. 5/5 97. — A 329.
96648. Urban, A., Görbersdorf, Kr. Waldenburg i/Schl. Senſendengelmaſchine mit ſelbſtthätigem, während des Betriebes verſtellbarem Vorſchub. 15/6 97. — A 295.
(†) 96649. Göhlers, Wilhelm, Wittwe, Freiberg i/S. Geräth zum Durchſchneiden der Wurzeln zwiſchen Pflanzenreihen, um ſpäteres Umpflanzen zu erleichtern. 6/7 97. — A 329.
(†) 96650. Uhl jr, J., Ravensburg, Württ. Getreide-Reinigungsmaſchine. 14/8 97. — A 295.
96750. Shiels, Dr A., Glasgow; Vertr.: H. Pataky u. W. Pataky, Berlin N.W., Luiſenſtr. 25. Durch abwechſelnd fallende und ſteigende Luftſaugung wirkende Melkmaſchine. 25/6 95. — A 445.
96751. Schöpke, A., Liebe b/Trebel, Hann. Pflug mit einem feſtſtehenden und einem verſtellbaren Streichbrett zur Herſtellung verſchieden breiter Furchen. 30/3 97. — A 329.
(†) 96837. Steen, P., u. C. Rödiger, Hamburg, Marthaſtr. 27. Vorrichtung zum ſelbſtthätigen Füttern von Vieh. 2/5 97. — A 412.
(†) 96838. Koch, M., Niedergurig b/Bautzen. Hackmaſchine mit durch Kurbelwelle hin- und hergeſchobenen Hackmeſſern. 8/5 97. — A 384.
96839. Ritter, H., Eberſchütz b/Hümme b/Kaſſel. Pferdeſtriegel. 13/5 97. — A 351.
96840. Rheiniſche Geſellſchaft für Metallinduſtrie Greve, Herzberg, & Co., Köln, Pfälzerſtr. 55. Fliegenfänger. 10/6 97. — A 368.
(†) 96917. Sonnenſchein, H., Berlin, Neues Thor 5. Leine mit durch Schnurzug zu öffnendem Schloß für Jagdhunde. 21/4 97. — A 351.
97001. Laſetzki, R., Bledzianow b/Antonin, Poſen. Bienenwohnung. 16/6 97. — A 412.
97050. de Keſel, K., Brüſſel; Vertr.: F. C. Glaſer u. L. Glaſer, Berlin S.W., Lindenſtr. 80. Bienenwohnung. 5/11 96. — A 430.
97051. Wunderlich, H., Borna b/Dresden. Einrichtung an Futtertrögen zum Befeſtigen derſelben an der Wagendeichſel. 3/6 97. — A 430.
97089. Nadányi, N., Kovácsi puszta, Biharer Comitat, Ung.; Vertr.: G. Dedreux, München. Geräth zum Schneiden der Weinreben beim Veredeln. 20/2 97. — A 396.
97090. Hovell, Th. M., Boreham Holt, Elſtree, Engl.; Vertr.: C. Gronert u. P. Kühne, Berlin N.W., Luiſen-

注释：这是我们9562份图像扫描件中的一个代表性页面（1898卷，第73页）。请注意，左上角和右下角的专利条目会跨越多个页面，同时还有一个专利条目横跨两列。还请注意，有许多信息是我们不希望纳入的，例如页脚中的连续页眉，以及在引入新的技术类别后出现的专利编号列表。新的技术类别会随机出现在两个内容栏中。此外，许多页面甚至不会引入新的技术类别。

不过，我们的图像语料库中还存在若干细微差异。最显著的是，1894年起，通用字体类型由罗马体变为哥特体。在哥特体时期，1894年至1911年的卷册使用Unger字体，而1911年至1918年的卷册则采用Breitkopf字体。随着专利注册数量增加，后期卷册引入了技术子类别（例如“17a”“89k”）。专利号前还会加上编号（例如“17. 287909”）以标明专利组，分类因此变得更加细致。此外，不同卷册之间还有许多细微差异。例如，注册日期的格式各不相同，部分专利号前后会标有匕首符号“(†)”，表示这些专利在该卷册出版前已失效。

图2 专利条目


19488. — LORENZ, W., in Wien; Vertreter: J. BRANDT in Berlin W., Königgrätzerstr. 131. Backofen für kontinuirlichen Betrieb. 26. Januar 1882. — A 765.

注释：一位代理人　　专利条目，见1882年卷册第1页。该专利列出了法律代理人，因为它曾　　s由奥地利人登记　　。其中不包括优先权主张，后者通常出现在最末尾。

5 基于图像扫描件的数据集构建：多模态大语言模型

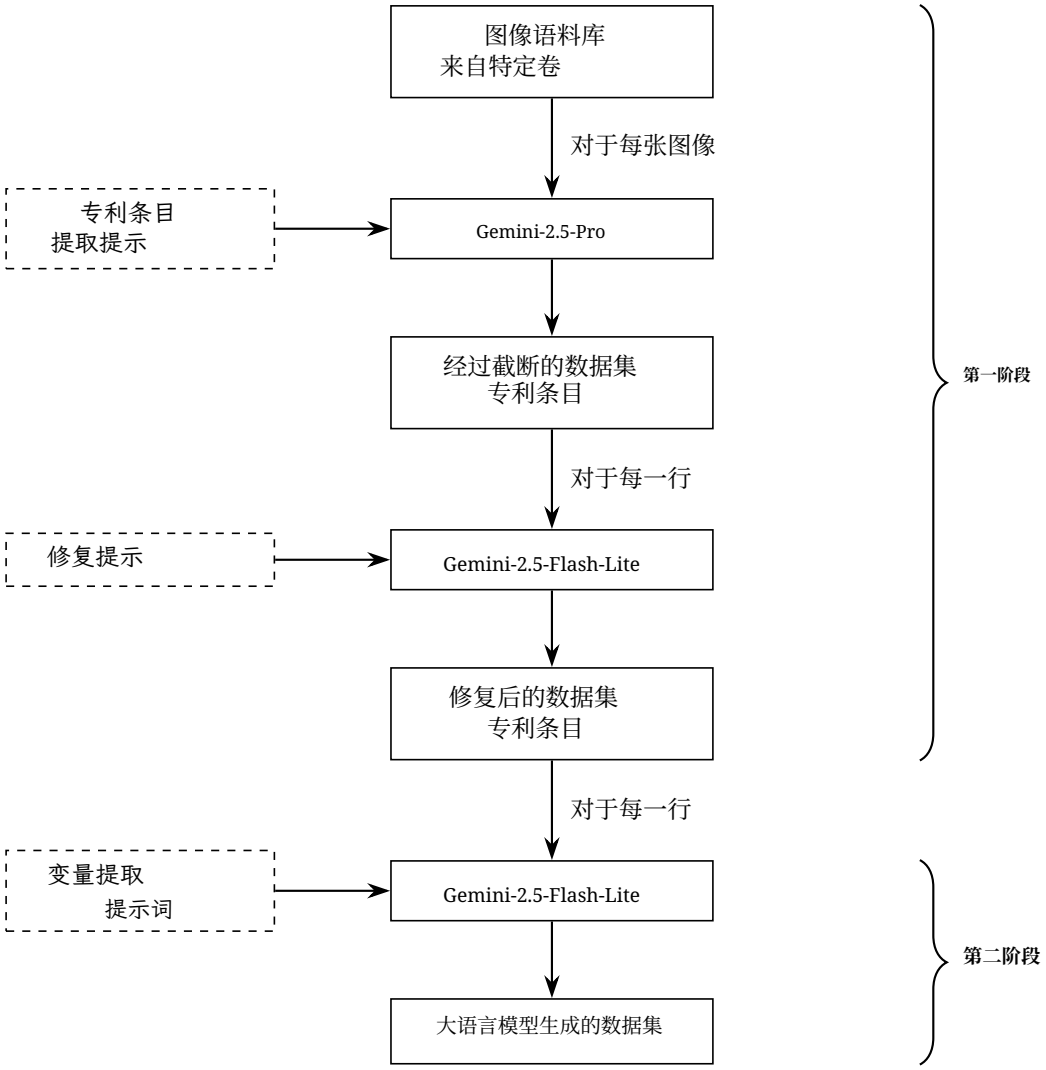
图3展示了我们用于从图像扫描件构建数据集的基于大语言模型的数据管道概览。第一阶段，我们将每个页面（例如图1）与同一提示词配对，并将所有提示词与页面的组合分别发送给多模态大语言模型。作为输出，我们会收到每页一个 JSON 文件，其中包含所有专利条目。对于给定的卷，我们将所有 JSON 文件合并，并转换为一个大型 CSV 文件。在第一阶段结束时，我们的处理流程会修复跨页面和跨列的专利条目，并生成一个数据集，其中每个观测值都包含一条完整的专利条目。在第二阶段，我们将每条提取出的专利条目与新的提示词配对，并将这些提示词与条目的组合发送给另一个大语言模型。我们会获得提示词中指定的提取变量，并将其追加到数据集中相应的专利条目中。接下来的小节将介绍技术细节：为何选择 Gemini 模型系列来驱动处理流程；如何处理技术类别、修复跨多列或多页面的专利条目、使处理流程适应具有特殊版式的卷，以及如何对大语言模型生成的专利数据集进行人工验证和清理。

5.1 模型选择

我们的基于大语言模型的数据管道完全采用 Gemini 2.5 系列模型（Comanici等人，2025年）。[2]批评者可能会认为，我们使用的是私营公司提供的专有模型。不过，虽然确实存在开源大语言模型，但初步实验明确表明，它们远不足以从我们的图像扫描件中提取所需信息。即使当时存在能力足够的开源替代方案，也需要使用昂贵的 GPU硬件，而我们并不具备这样的条件。Anthropic 或 OpenAI 等其他前沿实验室的专有模型也未能取得令人满意的结果。因此，在编写代码时，Gemini-2.5-Pro 是唯一具备良好视觉能力、能够从我们的图像语料库中提取信息的模型。归根结底，只要能以低廉的价格快速生成高质量数据集，大多数经济史学家很可能并不在意其数据处理流程由哪种大语言模型驱动。

[2]我们不为任何提及的产品、服务或组织背书，也不提供法律或财务建议。

图3
我们的基于大语言模型的数据管道



注释：此流程图展示了针对给定卷册的基于大语言模型的数据管道。输出是一个大语言模型生成的数据集。经过手动数据清理后，我们将全部 41 个大语言模型生成的数据集合并，构建完整的专利数据集，其中涵盖 1877 年至 1918 年间的所有德国专利。虚线框表示我们经过仔细改进的提示词，详见附录 A。在整个处理流程中，所有模型调用的温度参数均设为 0.0。

5.2 第一阶段：从图像扫描件中提取专利条目

我们的第一阶段以 Greif 等人（2025年）的研究为基础。首先，将给定卷的 PDF 拆分为 PNG 文件。每个页面都配上同一提示词（图 A.1），并通过应用程序接口（API）发送给 Gemini-2.5-Pro。[3] 所有 PNG 文件都会同时独立处理。提示词对于从档案图像扫描件中正确提取所需信息至关重要。我们仔细指示模型从上到下提取所有专利条目和技术类别，先处理左侧栏，再处理右侧栏。对于每个页面，我们都会收到一个 JSON 对象，其中包含按顺序阅读顺序排列的所有专利条目和技术类别。对于 API 调用大语言模型失败的页面，我们会不断重试，直至调用成功。专利条目前总是带有键 条目，而技术类别前则带有 类别。我们按照文档顺序串联所有 JSON 文件，并记录页码。因此，我们知道，两个技术类别之间出现的所有专利条目都属于前一个技术类别。这样，我们便能将串联后的 JSON 文件转换为包含三个变量的 CSV：页面、条目和类别。

独立处理每一页会带来一个重要问题。有些专利条目从右栏最底部开始，并延续到后续页面左上方。类似地，有时条目会横跨同一张图像的两栏，这使 Gemini-2.5-Pro 更难以正确提取完整文本。因此，从图像扫描件中提取后，有些专利条目会出现截断。我们通过将每个条目发送给 Gemini-2.5-Flash-Lite，并使用提示词（图 A.2）检查条目是否被截断（1 表示是，0 表示否），来解决这一问题。我们选择 Gemini-2.5-Flash-Lite，是因为它的成本远低于 Gemini-2.5-Pro，同时也同样能够解决这一问题。最后，我们将一手史料中跨越两页或两栏的截断条目合并起来。

有人可能会问，为什么我们不直接将所有页面放入一个 PDF 文件中发送给 Gemini-2.5-Pro，以避免条目超出单页的问题。首先，随着输入图像数量增加，多模态大语言模型的检索和推理性能会下降（Wu 等人，2025年）。其次，Gemini-2.5-Pro 的最大输出窗口有上限。因此，由于每卷通常包含数百页，我们不得不将所有卷分成多个部分发送，这会导致每个部分的首页和末页出现相同的专利条目跨页延续问题。第三，逐页独立处理能让我们确切知道专利条目位于哪张 PNG 图像上，从而便于之后进行手动数据清理。

5.3 第二阶段：从专利条目中提取变量

在基于大语言模型的数据处理流程的第一阶段结束后，我们现在得到一个变量条目，其中每一行都包含一个完整的专利条目。针对每个条目，我们使用 Gemini-2.5-Flash-Lite 提取以下变量：专利_编号、受让人、地点、标题和日期。我们指示模型提取指定变量，提供一些示例，并在提示词底部附上实际的专利条目（图 A.3）。每次 API 调用中，模型只处理一个在第一阶段提取出来的、独立的专利条目。如果 API 调用失败，系统会自动重试，最多十次。Gemini-2.5-Flash-Lite 会返回一个包含所需变量键的 JSON 对象。如果专利条目中没有某个变量，我们会指示模型返回非数值。第二阶段结束时，41 卷中的每一卷都被转换为包含以下变量的数据集：页码、条目、类别、专利_编号、受让人、地点、标题和日期。

5.4 不同布局的特殊卷册

前两卷（1877–8年和1879）与1880年至1918年的各卷在版式上略有不同。第一卷将1877年和1878年合并在一起，因为帝国专利局于1877年7月1日成立。这两卷的主要区别在于，专利编号位于每条专利条目的末尾，且前面总是带有“P. R.”。此外，1877–8年合订卷不包含受让人地点的信息。我们的处理流程使我们能够轻松调整提示词。在基准数据集上尝试了各种提示词后，我们修改了处理流程各阶段的提示词，以构建1877–8年和1879年的数据集。这两卷所用的提示词见附录 B。

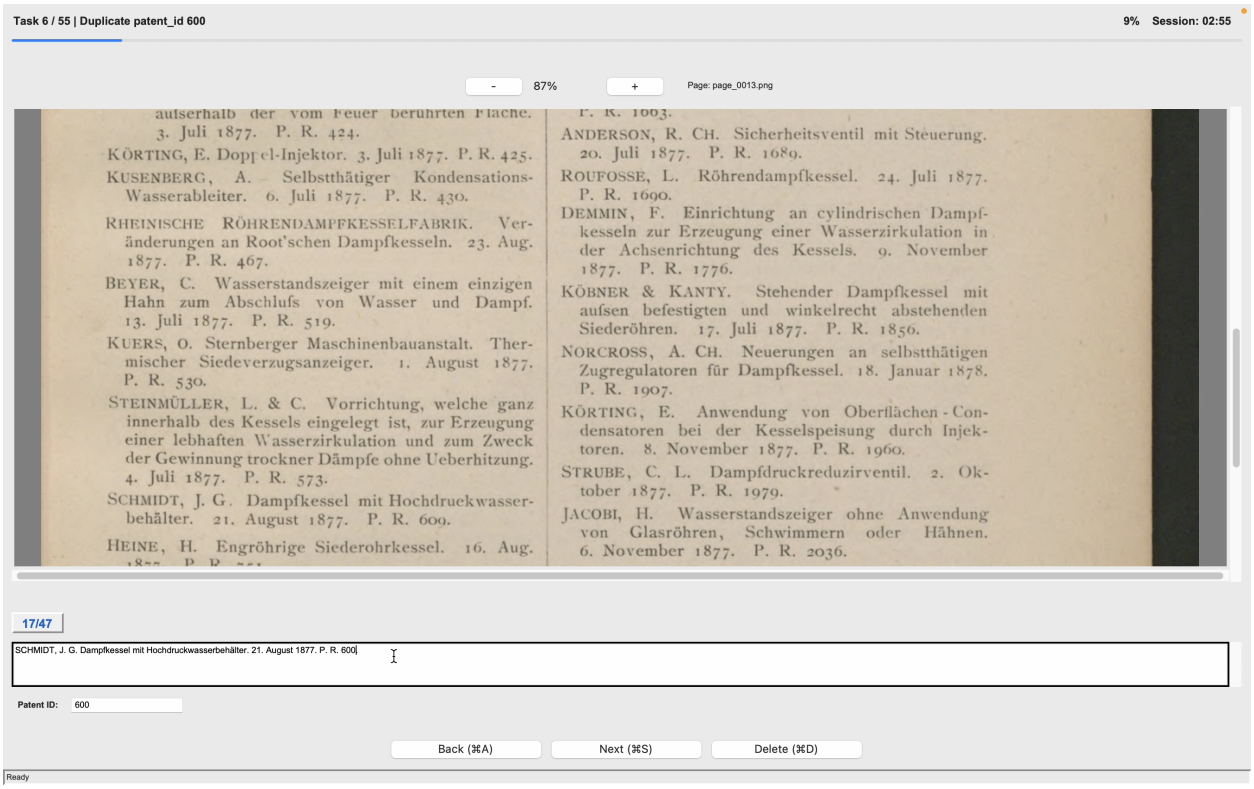
[3] API 是向部署在商业提供商数据中心的大语言模型发送输入的标准方式，随后该模型会返回输出。

5.5 手动数据清理

基于大语言模型的数据管道构建了41个数据集（实体卷册各对应一个数据集）后，我们的研究助理对少量观测值进行了人工验证和清理。在整个处理流程中，我们创建了辅助变量，用于标记被合并的截断条目、API调用未成功的条目，以及未找到专利编号的样本。借助这些辅助变量，研究助理筛查了这41个数据集中的每一个，以检查、修复并删除这些观测值。

随后，我们利用一手史料的结构，确保 41 个数据集的质量。专利号从 1 开始逐一递增记录。目录标明了相应卷中首个和最后一个专利号，因此我们能够检查重复项，并解决专利号超出该卷范围的问题。[4] 借助大语言模型辅助编码工具，我们迅速为研究助理搭建了一个可视化界面（图 4）。例如，如果某个专利编号被标记为重复项，界面便会在顶部显示相应页面，下方则可编辑该条目的当前值和编号。研究助理只需核对图像扫描件，并在必要时更新专利条目和编号字段。在整个验证过程中，我们还注意到主要来源本身也存在错误。例如，专利文员偶尔会出错，导致专利条目出现重复编号。[5]

图4 加速手动数据清理界面



注释：我们使用大语言模型辅助编码工具创建了这个界面。研究助理利用该界面手动清理和验证专利条目，包括 ID 值重复或超出卷号范围的条目。在示例中，专利编号“609”中的“9”下半部分印得很淡，导致 Gemini-2.5-Pro 将其识别为“600”。由于每个专利编号都是唯一的，因此这一差异被标记出来。随后，研究助理在适用时通过亲自核对输出，纠正了这些错误。我们的界面加快了这一流程，研究助理只需进行修改，然后点击“下一步”即可。

[4]由于第一次世界大战，1917年和1918年的卷中约有4,000个专利号未被录入专利登记册。[5]我们没有纠正帝国专利局员工造成的这些错误，也没有重新分配不符合历史事实的专利号。因此，我们的数据集中偶尔会出现两项不同专利使用同一专利号的情况。

一手史料的结构还有助于我们识别技术类别转录错误的观测值。在同一卷中，所有专利都按照技术类别依次排列。例如，类别“2”必须接在类别“1”之后。同样，子类“17e”必须排在“17d”之后。通过标出这些序列中的不一致之处，研究助理手动更正了 类别 变量。

部分卷的内容存在差异，迫使我们批量自动删除专利条目。对于1888年，我们只找到了一卷合订本，其中收录了1878年7月1日至1888年12月31日期间登记且仍然有效的所有专利。因此，这一卷包含了前几卷中的专利条目，我们将其删除。同样，1879年的卷中也包含了仍然有效、且已列于涵盖1877年和1878年的第一卷中的专利条目。这些条目也被删除了。

对全部41个大语言模型生成的数据集完成上述手动数据清理后，我们将它们合并为一个数据集，并在开头添加变量全局_编号、*book*和*book*_编号 ，以提供高层级元数据。

6 基准测试

基准数据集有两个相互关联的用途。当然，它使我们能够在一个规模较小但具有代表性的样本上，评估基于大语言模型的数据处理流程所生成的数据集的质量。然而，数据集质量很大程度上取决于每次处理流程调用大语言模型时所选用的提示词。因此，基准数据集也有助于我们找到能够最大限度提高大语言模型生成的数据集准确率的提示词。我们在随机选取的41张图像所构成的代表性语料库上优化基于大语言模型的数据处理流程的性能，并据此认为，在该样本上报告的评估指标也适用于完整的图像语料库。

6.1 幻觉

大语言模型经常被发现会产生幻觉，因为它们可能会编造看似可信却不正确的信息。为了评估这一问题是否会出现在我们的任务中，我们在研究助理的协助下构建了一个人工构建的基准数据集，用于评估多模态大语言模型的输出。

幻觉有几种类型。在我们的数据集构建任务中，我们只会遇到“输入冲突型幻觉，即大语言模型生成的内容偏离用户提供的源输入”（Zhang等人，2025年，第5页）。更准确地说，我们将大语言模型生成的数据集与大语言模型生成的 数据集相对于完美 基准数据集的任何偏差定义为幻觉。

对于我们的研究项目而言，并非所有幻觉都同样令人担忧。错误识别字符，或使用现代字母而非历史字母，可以视为轻微幻觉。在分析历史大规模数据时，这些问题微不足道。更值得担忧的是中度幻觉，例如多模态大语言模型将技术类别标题提取为专利条目，或将整个单词转录成现代拼写而非历史拼写。严重幻觉会极大影响数据集质量及其用于后续计量经济学分析的可靠性。这包括对专利条目的转录完全错误或误读；更糟的是，还包括凭空编造图像扫描件中并未出现、因而并不存在的专利条目。

6.2 基准数据集：学生构建与完美

我们随机选取了41份图像扫描件，即每个日历年出版的专利登记册各选取一份，将其分配给两名研究助理，并仔细指导他们如何在 Excel 中创建数据集。下文中，我们将他们最初生成的输出称为学生构建的数据集。数十年来，借助研究助理从档案图像扫描件构建数据集一直是经济史研究中的标准做法。因此，我们确信，这一学生构建的数据集大致代表了现有经济史数据集的平均质量。

研究助理构建出学生构建的数据集后，我们使用基于大语言模型的数据管道处理抽样得到的 41 幅图像，生成了相应的数据集。随后，我们将大语言模型生成的数据集与学生构建的数据集进行比较，以评估我们的处理流程的准确率。我们发现两个数据集之间存在重大差异。起初，我们怀疑 Gemini-2.5-Pro 捏造了专利条目，但很快意识到

研究助理不小心漏掉了一些条目。我们还发现了几处换位错误：学生将“299187”提取为专利号，而大语言模型则正确写为“299178”。此外，我们还注意到许多拼写错误，例如字符重复或遗漏。因此，我们请研究助理修正最初的学生构建的基准测试数据集，直至其完美。据我们所知，这个完美的数据集不再含有任何错误。

6.3 大语言模型辅助的迭代提示词优化：基准数据集

我们在整个处理流程中寻找最优提示词的方法，与 Bäcker-Peral 等人（2025年）和 Vafaie 等人（2025年）的同期研究相似。我们在整个处理流程中迭代优化提示词，旨在生成尽可能准确匹配完美 基准数据集的数据集。每次迭代后，我们都会检查所选的评估指标，并查看大语言模型生成的 数据集与完美数据集的并排比较。通过目视检查，我们发现了当前提示词的不足之处。我们描述这些不准确之处（例如，在早期提示词中，创建地点变量时提取的地点带有介词“in”，如“in Berlin”），并指示大语言模型辅助编码工具考虑这些问题来改进提示词。我们多次重复采用这种大语言模型辅助的提示词优化方法。很可能还存在其他提示词，能够使评估指标取得更高的数值。不过，经过目视检查后，我们认为当前提示词已足够完善，能够确保整个图像语料库具有较高的数据质量。

我们通过融入对一手史料的历史领域知识，编写了初始提示词。一般而言，我们会先为大语言模型设定一个角色（例如：“你是一款专门用于数据提取的人工智能。”），然后简要概述总体任务。接着，我们会提供非常详细、明确的指示，之后在适用的情况下展示一些示例及其所需的 JSON 输出。最后，我们经常指示模型只提供 JSON 输出，不要提供其他内容，以避免不必要的解释干扰处理流程中的后续步骤，例如拼接 J SON 对象或将其整合到数据集中。

6.4 字符错误率

在对处理流程的第一阶段和第二阶段进行基准测试之前，我们先评估了 Gemini-2.5-Pro 的转录能力，评估范围仅限于专利条目。不过，这需要先完成第一阶段的第一部分，即将 41 张基准图像发送给模型。对于每一卷，我们将所有大语言模型生成的 专利条目合并到一个文本文件中，以衡量转录准确率。我们对学生构建的 和完美基准数据集也进行了相同的处理。

随后，我们使用字符错误率（CER）这一光学字符识别领域的标准指标，量化两份文本文档在字符级别上的差异。CER 定义如下：

$$\mathrm{CER} = \frac{\text{Levenshtein Distance}}{N} = \frac{S + D + I}{N} \tag{1}$$

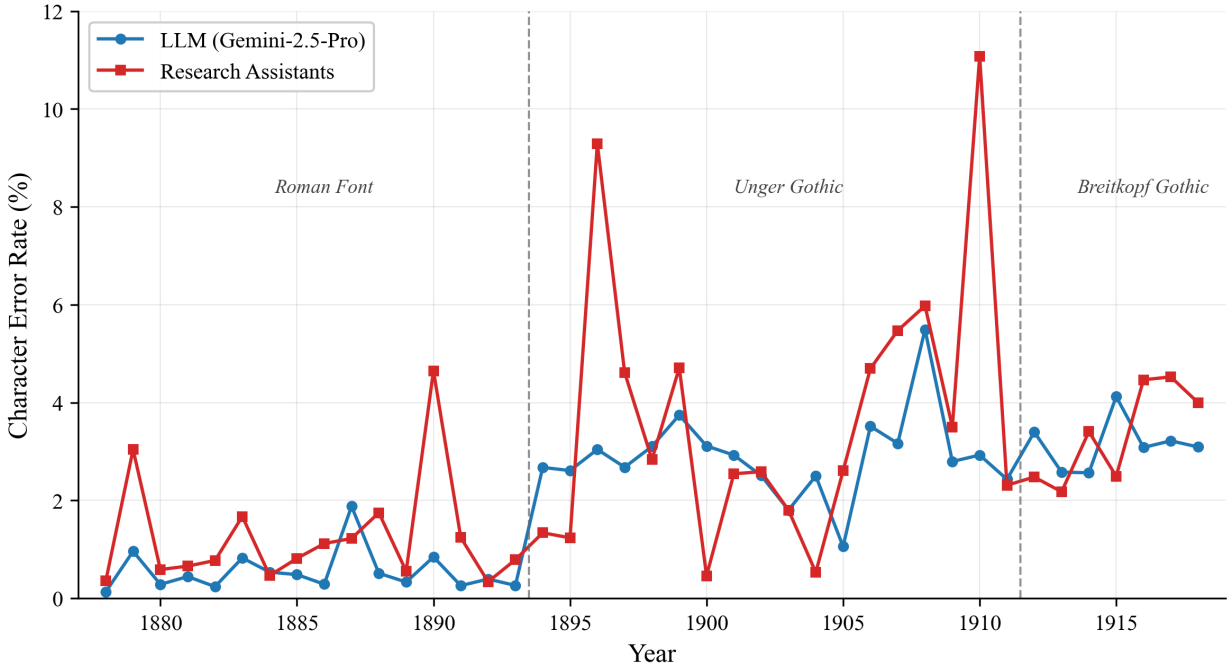
其中，莱文斯坦距离定义为替换次数 S、删除次数 D 和插入次数 I 之和。 N 是完美 文本文件中的字符数。直观来说，CER 表示将大语言模型生成的 专利条目转录文本转换为完美 文本文件（或反向转换）所需的最少击键次数，并根据完美 文件的长度进行归一化。

图5报告了全部41卷中，学生构建的 转录文本与完美 转录文本之间，以及大语言模型生成的 转录文本与 完美 转录文本之间的字符错误率（CER）差异。在41个文本文件中，有27个文件里，Gemini-2.5-Pro生成的转录文本比我们研究助理的成果更接近完美 版本。另有一个文件的CER相同，其余13个文件则是研究助理的准确度更高。哥特字体的CER显然更高，但我们的研究助理似乎也遇到了同样的问题；他们告诉我们，不得不借助哥特文阅读辅助工具。进一步检查发现，Gemini-2.5-Pro主要难以处理历史长体 s，偶尔也会在大小写方面出错。[6]

[6]我们还提供了一个配套网站，以可视化形式展示所有基准测试结果，方便读者自行评估数据集的质量。请访问 historymind.ai。

我们的专利条目CER往往低于研究助理的这一事实，让我们对严重幻觉问题颇感安心。例如，单个虚构的专利条目就会使卷级CER增加2–3%。尽管如此，我们还是发现了五处值得关注的幻觉，并希望将其讨论出来，以便完全透明地说明大语言模型生成的数据集的质量。我们发现最严重的一处幻觉出现在第1878卷的基准图像中。Gemini-2.5-Pro部分编造了ID为7763的专利名称中并不存在的内容。此外，第1907卷基准图像上的最后一条专利条目被模型补入了图像中未显示的四个词。模型也未能正确抽取第1915卷基准图像中的第一条专利条目，遗漏了开头部分。对于第1908卷的图像，ID为199989的专利转录文本遗漏了一项附加的优先权主张，该主张位于申请日期之后。最后，在第1913卷的基准图像上，ID为259447的专利条目写作“Versteinung”，而大语言模型错误地抽取为“Verkoksung”。所有这些例子都说明了一个重要问题：虽然Gemini-2.5-Pro在转录专利条目方面往往优于我们的研究助理，但基于双栏一手资料构建的数据集仍不完美。

图5 按年度卷划分的字符错误率：研究助理与大语言模型



注释：我们从41个年度卷中各抽取了一页。每个数据点表示所有大语言模型生成的 （蓝色）或学生构建的 （红色）专利条目转录文本相对于完美 转录文本计算得出的字符错误率。我们没有对转录文本进行标准化，因为我们希望保留其历史准确性。这是一种不同寻常的选择，会导致字符错误率偏高，但能让我们更清楚地看到，研究助理和Gemini-2.5-Pro生成的转录文本与档案图像扫描件上实际呈现的完美 转录文本有多接近。

6.5 第一阶段：从图像扫描件中提取专利条目

虽然字符错误率初步表明，Gemini-2.5-Pro提取和转录专利条目的准确度可能高于我们的人类研究助理，但我们还需要评估基于大语言模型的处理流程在保留结构和修复专利条目方面的表现。

我们对包含大语言模型生成的条目和完美条目的数据集进行一对一匹配。为容许因字符误识别造成的细微差异，我们采用模糊字符串匹配。若两个专利条目在字符级别上至少有 90% 相同，则视为匹配，这大致意味着错误字符不超过字符总数的十分之一。

[7] 我们在修复截断条目之前和之后分别进行匹配。在这项基准测试中，只有当条目跨列时，才能修复截断条目，因为我们从每卷中抽取了一页。这在概念上等同于条目跨越两个后续页面的情况，为评估修复程序的可靠性提供了良好的测试场景。

表1展示了基准测试结果。专利条目提取完成后，1,403个大语言模型生成的条目中有1,360个与1,385个完美条目之一匹配。41个基准测试图像样本中有18个包含跨越两列的专利条目。最初，这些跨列条目的两部分都被提取为两个独立的专利条目，这解释了为什么 Gemini-2.5-Pro 提取出的条目数多于实际存在的条目数。修复这些跨列条目后，大语言模型生成的基准数据集也包含1,385个条目，其中1,378个在完美基准数据集中有匹配条目。7个未匹配条目中，有5个因第6.4节所述的幻觉而未能匹配。其余两个条目中历史长 s 字符占比较高，但除此之外基本正确。

表1：第一阶段基准测试结果

	提取 （Gemini-2.5-Pro）	**修复** （Gemini-2.5-Flash-Lite）	**学生构建**
完美 条目	1,385	1,385	1,385
已提取条目	1,403	1,385	1,373
匹配条目	1,360	1,378	1,358
%完美 匹配数	98.19%	99.49%	98.05%
提取匹配率	96.94%	99.49%	98.91%

注： 该完美 基准数据集由 1,385 条专利条目组成。“提取”列指第一阶段的初始部分，即使用 Gemini-2.5-Pro 从图像扫描件中提取专利条目。“修复”列报告了使用 Gemini-2.5-Flash-Lite 修复截断条目后的表现。第三列展示了研究助理的表现，最适合与“修复”列进行比较。我们使用 rapidfuzz 库的归一化莱文斯坦相似度指标确定匹配项，阈值设为 0.90。报告的百分比也称为召回率（%完美 匹配数）和精确率（% 提取匹配率）。

6.6 第二阶段：从专利条目中提取变量

为对第二阶段进行基准测试，我们评估了 Gemini-2.5-Flash-Lite 从前一步骤（第 6.5 节）中匹配的专利条目子集中提取所需变量的准确度。在这两个子集（大语言模型生成的和学生构建的）中，我们为每条专利条目提取所需的五个变量（专利_编号、受让人、地点、标题和日期），并将它们与完美 基准数据集中的相应变量进行比较。

为了判断两个变量单元格是否匹配，我们采用模糊字符串匹配。将阈值设为较保守的 90%，可以容许较长字段（如专利标题）中的转录错误，同时仍要求较短字段（如六位专利编号或日期）完全对应。

在第一阶段中，大语言模型生成的数据集与完美基准数据集中有 1,378 条记录匹配，共计 6,890 个变量单元格；其中 6,550 个在基准测试的第二阶段中匹配，比例为 95.07%（表 2）。较短的专利编号和日期字段的准确率分别为 99.49% 和 98.91%。受让人（92.16%）、地点（91.36%）和专利标题（93.40%）的准确率较低；这符合预期，因为较长字段和大小写方面会出现转录错误，而专利编号和日期通常不会出现此类问题。

值得注意的是，由于学生完全漏掉了一些专利条目，或错误地重复了先前条目中的专利标题，等等，学生构建的 基准数据集与完美数据集仅有 1,358 条匹配记录。因此，学生构建的 数据集是在一个较小的子集上进行评估的，这使得它与大语言模型生成的 数据集更难直接比较。尽管如此，即使在这个缩小后的子集中，

[7]采用 rapidfuzz 库的归一化莱文斯坦相似度指标进行贪心一对一模糊匹配，阈值设为 0.90。

研究助理在提取专利编号、地点和日期方面的准确度较低，而在提取受让人和标题方面表现更好。这种模式的原因是，我们明确要求研究助理使用历史长体 s，而 Gemini-2.5-Pro 未能生成这种字形。

一些经济史学家可能会认为，报告中大语言模型生成的数据集上的变量级准确率不够高，或者剩余错误存在系统性偏差。因此，我们仔细检查了未匹配的变量单元格。[8] 对于受让人字段，Gemini-2.5-Flash-Lite 有时会错误地将职业也包括在内，而职业偶尔会出现在个人名称之后。此外，对于三名或更多专利持有人，Gemini-2.5-Flash-Lite 始终用“und”（德语“和”）分隔每个名称，而完美基准测试数据集遵循逗号分隔列举惯例，仅在最后一个名称前使用“und”。大语言模型生成的地点后面常常紧跟一个句点，中间没有空格，而完美基准测试数据集中并不存在这种情况（例如，大语言模型生成的“Berlin.”与完美的“Berlin”不匹配）。[9] 此外，许多地点包含历史长体 s。这种单字符差异可能导致匹配失败，因为提取出的许多地点都很短（例如，“Dresden”与包含历史长体 s 的对应名称不匹配）。根据基准测试结果，我们认为，我们的大语言模型生成的专利数据集质量足够高，可用于经济史研究。

表 2：第二阶段基准测试结果

	Gemini-2.5-Flash-Lite			**学生构建**		
	总计	匹配数	匹配率	总计	匹配数	匹配率
单元格总数	6,890	6,550	95.07%	6,790	6,503	95.77%
按变量						
专利编号	1,378	1,371	99.49%	1,358	1,329	97.86%
受让人	1,378	1,270	92.16%	1,358	1,289	94.92%
地点	1,378	1,259	91.36%	1,358	1,231	90.65%
标题	1,378	1,287	93.40%	1,358	1,316	96.91%
Date	1,378	1,363	98.91%	1,358	1,338	98.53%

注：此表报告了将大语言模型生成的 变量和学生构建的 变量与完美 基准数据集中的变量进行比较所得的变量提取表现。根据表1中的基准测试结果，我们保留了1,378条大语言模型生成的 条目和1,358条学生构建的 条目，分别包含6,890和6,790个变量单元格。对于每个专利条目，我们提取五个变量（专利_编号、受让人、地点、标题和日期），并将其与完美 基准数据集中的变量值进行比较。两个单元格是否匹配，是通过 rapidfuzz 库的归一化莱文斯坦相似度指标判定的，阈值设为0.90。

7 使用多模态大语言模型从图像扫描件构建数据集的经济学

迄今为止，世界各地的经济史学家一直在手工从一手资料构建数据集。从档案来源中提取信息的工作通常由研究助理、博士生完成，有时甚至会外包给专业公司。这一过程非常耗时且成本高昂。构建基准数据集时，我们要求研究助理记录所花时间。平均而言，他们每页大约花两小时，制作学生构建的数据集，其中包括完整的专利条目以及所需的五个变量。创建完美的基准数据集还需要每页额外花费两小时。由于我们的图像语料库包含 9,562 页，按照经济史学中的标准做法，构建该专利数据集需要约 19,124 小时的人力。相比之下，我们基于大语言模型的数据管道可以在一天内生成完整数据集。假设德国最低工资为 12.82 欧元，构建该专利数据集至少需要花费 245,269 欧元。相比之下，基于大语言模型的数据管道的成本为 1,196 欧元（表3）。因此，多模态大语言模型使我们构建该专利数据集的速度提高了 795 倍以上，成本降低了 205 倍

[8]所有变量提取基准测试结果均可在我们的论文网站上查看，网址为 historymind.ai。 [9]进一步优化提示词或许可以解决这些问题。

，与经济史领域构建数据集的标准做法相比，成本低廉得多。多模态大语言模型在历史数据构建方面的优势还体现在：该处理流程产出的数据质量高于我们对研究助理的预期。[10]

在我们的基于大语言模型的数据管道中，模型选择是主要的成本驱动因素。为了确保最高的数据质量，我们本可以在处理流程中所有使用大语言模型的环节，专门采用当时编写代码时 Gemini 系列中能力最强的模型 Gemini-2.5-Pro。然而，这样做会大幅增加成本，而质量提升却很有限。因此，我们选择了能力较弱但价格更低的 Gemini-2.5-Flash-Lite，用于修复截断条目，并从先前提取的专利条目中抽取变量。该模型在基准数据集上的表现令人满意；既然结果已经令人满意，为什么还要使用更智能、价格更高的模型？随着大语言模型推理成本持续快速下降（Gundlach 等人，2025年），研究人员必须根据各自的数字化项目，不断重新评估哪些模型能够满足需求。一旦开源模型达到当前前沿模型的性能，它们就可能成为数据集构建任务中可靠的替代方案。

表3：按流水线阶段划分的成本明细

		美元/100万词元		已使用的词元			
阶段	**模型**	输入	输出	输入	输出	**成本（美元）**	**%**
第一阶段：从图像扫描件中提取专利条目							
提取	Gemini-2.5-Pro	$1.25	$10.00	24.1M	107.6M	1,105.9	92.4%
修复	Gemini-2.5-Flash-Lite	0.10美元	0.40美元	1.831亿	0.3M	18.4	1.6%
第二阶段：从专利条目中提取变量							
提取	Gemini-2.5-Flash-Lite	0.10美元	0.40美元	6.129亿	2660万	71.9	6.0%
总计				**8.201亿**	**1.345亿**	**1,196.3**	**100%**

注：我们按照2025年8月的价格，统计了基于大语言模型的数据管道中的所有成本。我们没有使用批处理，否则总成本可降低50%。输出词元包括思考（推理）词元和输出词元。我们的处理流程中，主要成本来自 Gemini-2.5-Pro 的1.076亿个输出词元。所有词元数量均四舍五入至小数点后一位。

使用大语言模型的成本取决于其接收的输入长度和生成的输出长度，长度以词元为单位衡量。文本词元可以是一个单词，也可以是更小的文本单位。类似地，图像也会被转换为图像词元。在向 Gemini-2.5-Pro 发出单次请求时，图像扫描件和提示词会被转换为词元，共同构成输入词元总数。通常，输入词元的价格较低，而输出词元则贵得多。鉴于这些成本机制，编写非常明确的提示词不会带来太多成本。相比之下，Gemini-2.5-Pro 的输出词元价格高昂，这也解释了为什么我们将预算的 92.4% 用于从图像扫描件中提取专利条目。尽管如此，从档案图像扫描件中提取专利条目是整个数据集质量的基础，因此我们希望使用实施时最先进的多模态大语言模型。除输入和输出词元外，Gemini-2.5-Pro 和 Gemini-2.5-Flash-Lite 还会生成思考词元。收到请求后，这两个模型都会先进行“思考和推理”，再给出答案。任务越难，模型就会投入越多资源来生成思考词元，以便拟定方案并规划如何成功完成用户请求。思考词元按与输出词元相同的价格计费，这进一步推高了我们第一阶段的成本。

[10]目前，在德国巴登-符腾堡州聘用的研究助理薪资约比最低工资高出1欧元。此外，我们的处理流程只需几秒钟即可处理全部9,562份图像扫描件。然而，受个人电脑性能限制，为避免 Google Gemini API 的速率限制问题，我们实际耗时约一天。因此，我们的比较计算应视为下限估计。

8 结论

本文作出了几项贡献。第一，我们构建并发布了一个基于大语言模型的数据管道，使我们能够将德国历史专利登记册的图像扫描件转换为结构化专利数据库，其中包含德意志帝国全部 306,070 项专利的持有人和内容信息。第二，我们创建并开源了一个学生构建的 和一个完美的 基准数据集，用于评估多模态大语言模型在根据图像扫描件构建历史数据集这一任务上的准确率。我们的基准测试初步表明，多模态大语言模型生成的数据集质量可能高于人类研究助理构建的数据集。尽管我们只能确定这一结论适用于我们的图像语料库，但有充分理由相信，对于其他乃至更复杂的历史资料，情况也可能如此。因此，我们鼓励其他经济史学家使用大语言模型辅助编码工具，将我们的处理流程应用于各自的图像语料库，以探索多模态大语言模型的普遍可靠性。第三，我们开源了完整的专利数据集，这是研究创新史的宝贵资料。在本结论的其余部分，我们将讨论多模态大语言模型对经济史领域可能产生的影响。

长期以来，经济史研究一直受制于缺乏大规模微观经济数据。多模态大语言模型的出现或许能终结这一困境。然而，利用人工智能代理构建大规模数据集存在一项系统性风险：由于我们对大语言模型的了解有限，验证所有数据条目实际上并不可行，这一风险因此进一步加剧。大语言模型生成的数据集中若引入未被察觉的幻觉，甚至更糟的系统性偏差，将威胁后续计量经济学分析的稳健性。如果基于有偏数据集得出的见解影响现代政策，这可能会造成有害的现实后果。例如，我们的大语言模型生成的数据集包含 306,070 条专利条目，每条均提取出五个变量和一个技术类别，因此人工审核全部 2,142,490 个数据单元格实际上是不可能的。因此，我们不得不依赖基准数据集上报告的结果，并假定在这 41 个页面构成的代表性子集上取得的性能能够推广到整个图像语料库。未来关于德国创新成因及其后果的基于专利的历史研究，其可靠性取决于这一假设的准确性。最后，我们指出，这种生成大规模数据的新方式偏离了经济史领域的标准做法；在该领域，细致的人工工作一直帮助经济史学家深入理解一手史料，并培养对其的历史直觉。

多模态大语言模型或许很快就能让经济史学家按需从图像扫描件构建数据集：只需将 PDF 发送给模型，并指示它以 CSV 或 JSON 对象的形式输出数据集。[11] 多模态大语言模型还能帮助研究界迅速复现知名的经济史数据集，并检验下游分析是否稳健。按需从图像扫描件构建数据集，也可能促使学者公开共享他们历经数十年亲手 painstaking 创建的数据集。这些数据集往往被私下保留，因为经济史学家计划将其用于自己的出版物。然而，按需构建数据集也可能促使公共数字档案馆下架其档案图像扫描件馆藏，以防他人下载这些扫描件并据此创建数据集，进而发表自己的研究成果。

未来的研究应基准测试大语言模型在基于手写字体和低资源语言构建历史数据集时的表现，或测试在提供包含多个页面的 PDF、而非逐页处理时的表现。此外，还应尝试将文本即数据的方法转化为图像即数据的方法。经济史学家既然可以直接处理包含嵌入文本的档案图像扫描件，为何还要使用可编辑的文本文件？总体而言，经济史学家应努力使用最接近一手史料的呈现形式。例如，设想一篇论文旨在确定自 1936年以来 BBC 每日电视广播所传达的情感倾向。经济史学家不应依赖音频转录文本，而应使用能够直接处理视频的多模态大语言模型，因为视频或许能捕捉信息传达方式中的更多细微差别，进而传递给

[11]本文撰写期间，Gemini-3-Pro-Preview 于 2025年11月 发布。该模型展现出更强的视觉能力；在收到“创建一个经济史数据集”的提示词以及一份多页 PDF 后，它似乎能输出近乎完美的数据集，同时还能处理被截断的专利条目。Gemini-3-Flash-Preview 随后于 2025年12月 发布，在几乎所有基准测试中的表现都优于 Gemini-2.5-Pro，成本却不到其三分之一。我们最初开发两阶段流程，是因为使用 Gemini-2.5-Pro 时，最佳策略是将任务拆解为较小的子问题，严格控制数据集构建过程，而不是要求模型直接根据一份或多份图像扫描件生成完整数据集。因此，我们保留未来使用更强模型重建数据集的权利，以进一步提升数据集质量，并排除任何残留的幻觉。

国家。与其他社会科学和人文学科一样，经济史或许会进入数据丰裕时代，届时经济史学家只需利用我们的专利统计资料等来源构建新的大规模数据集，便能回答新的问题。

多模态大语言模型或许不会改变经济史的核心，但正如它们已经在其他领域所发挥的作用一样，将加速研究周期（Bubeck等，2025年）。例如，在生物医学领域，通过找出文献中的空白并提出有前景的假设，已经取得了成功（Gottweis等，2025年）。此外，目前也有人正在努力开发能够实现数据驱动科学发现自动化的人工智能科学家（Mitchener等，2025年）。在经济学领域，关于借助人工智能代理加速经济研究的文献也开始涌现（Korinek，2025年）。然而，经济史领域几乎没有此类讨论，尽管该领域其实更容易受到自动化数据驱动发现的影响，因为原则上，所有答案都基于留存的文献与实物。尽管人工智能未来的发展及其影响仍有许多不确定之处，但我们确信，经济史的工作性质将会改变。研究助理的工作将不再侧重于根据一手资料手工构建历史数据集，而将转向利用大语言模型辅助编码工具协调基于大语言模型的数据处理流程，并管理其底层数据基础设施。经济史学家不再受限于手工构建数据集，而必须学会在海量数据中筛选出有望带来新颖且重要见解的研究问题。

致谢

我们感谢 PhilippAger、Stephen Broadberry、Alexander Donges、Gavin Greif、Leander Heldring、John Meyer、Sheilagh Ogilvie、2025年9月曼海姆博士研讨会的听众、2025年11月牛津数字学术学会成立活动的听众、2025年11月牛津研究生研讨会的听众，以及2025年12月牛津-华威-LSE（OWL）研讨会的参与者提供的反馈。此外，我们特别感谢 Carsten Burhop 借给我们大量实体卷册。我们还要感谢 Tünde Gottschling 和 Irene Schumm，在曼海姆大学数字实验室管理所有一手资料的扫描工作，并始终给予大力支持。最后，我们感谢研究助理，尤其是 Julius Schöffel，感谢他们扫描实体资料、构建基准数据集并手动清理专利数据集。本研究获得海德堡科学院资助，属于其“长期视角下的财政与企业研究”科学院项目的一部分。Niclas Griesshaber 非常感谢英国研究与创新机构通过高级定量方法奖提供的经济与社会研究委员会哲学博士资助。

人工智能透明度声明

本文初稿由作者在未借助任何外部工具的情况下撰写。此后，我们使用大语言模型重新梳理逻辑脉络、修正语法，并有选择地利用其根据初稿措辞提供的同义词和替代表达来提升可读性。GitHub代码仓库和项目网站的代码完全由大语言模型生成。具体而言，我们使用了 Cursor代码编辑器，通过自然语言指导编码代理构建网站和处理流程。我们尽最大努力审阅大语言模型生成的代码，确保在创建代码库和网站时负责任地使用人工智能。

参考文献

Abramitzky, R., Boustan, L., Eriksson, K., Feigenbaum, J. & Pérez, S. (2021), ‘历史数据的自动关联’, 《经济文献杂志》, 59(3), 865–918. Albers, T. N. & Kappner, K. (2023), ‘城市名录作为微观地理数据源的利弊’, 《经济史探索》, 87, 101476. Bäcker-Peral, V., Meursault, V. & Severen, C. (2025), ‘大语言模型能否可靠地将多样化历史表格转化为面板数据？’, *arXiv*预印本 *arXiv:2505.11599*. Baek, I., Chang, H., Ryu, S. & Lee, H. (2025), ‘大型视觉语言模型如何识别图像中的文本？揭示OCR头的独特作用’, *arXiv*预印本 *arXiv:2505.15865*. Bartik, A., Gupta, A. & Milo, D. (2025), ‘住房监管的成本：生成式监管测量的证据’, *SSRN 4627587*. Bergeaud, A. & Verluise, C. (2024), ‘一个用于研究欧洲和美国百年创新历程的新数据集’, 《研究政策》, 53(1), 104903. Broadberry, S., Campbell, B. M., Klein, A., Overton, M. & Van Leeuwen, B. (2015), 《英国经济增长，*1270–1870*》, 剑桥大学出版社。 Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P. et al. (2020), ‘语言模型是少样本学习器’, 《神经信息处理系统进展》, 33, 1877–1901. Bubeck, S., Coester, C., Eldan, R., Gowers, T., Lee, Y. T., Lupsasca, A., Sawhney, M., Scherrer, R. et al. (2025), ‘借助GPT-5开展的早期科学加速实验’, *arXiv*预印本 *arXiv:2511.16072*. Carlson, J. & Dell, M. (2025), ‘非结构化数据稳健高效推断的统一框架’, *arXiv*预印本 *arXiv:2505.00282*. Chyn, E., Haggag, K. & Maruthiah, C. (2025), 《政府意识形态：来自印第安事务办公室与同化时代的证据》，技术报告，美国国家经济研究局。 Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O. et al. (2025), ‘Gemini 2.5：凭借高级推理、多模态、长上下文和新一代智能体能力推动前沿发展’, *arXiv*预印本 *arXiv:2507.06261*. Cox, G. W. (2019), ‘专利披露与英格兰早期工业革命’, 《欧洲经济史评论》, 24(3), 447–467. Crosilla, G., Klic, L. & Colavizza, G. (2025), ‘针对手写文本识别的大语言模型基准测试’, 《文献学杂志》, 81(7), 334–354. Dell, M., Carlson, J., Bryan, T., Silcock, E., Arora, A., Shen, Z., D’Amico-Wong, L., Le, Q. et al. (2023), ‘美国故事：美国历史报纸的大规模结构化文本数据集’, 《神经信息处理系统进展》, 36, 80744–80772. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. (2019), ‘BERT：用于语言理解的深度双向变换器预训练’, 收录于《*2019*年计算语言学协会北美分会会议论文集：人类语言技术，第*1*卷（长篇与短篇论文）》, 4171–4186. Domini, G. (2020), ‘20世纪初的展览会、专利与创新：来自1911年都灵国际博览会的证据’, 《欧洲经济史评论》, 24(3), 578–600. Donges, A. & Selgert, F. (2019), ‘法律差异是否重要？1877年前德国专利法制度比较’, 《经济史年鉴／经济史年鉴》, 60(1), 57–92. Donges, A. & Streb, J. (2024), ‘德国发明创造力的成因，1815–1990：从专利统计中能学到什么’, 《德国经济评论》, 25(4), 301–323. Dosovitskiy, A. (2020), ‘一张图像胜过16×16个词：用于大规模图像识别的变换器’, *arXiv*预印本 *arXiv:2010.11929*. Galor, O. (2011), 《统一增长理论》, 普林斯顿大学出版社。

Goodfellow, I., Bengio, Y., Courville, A. & Bengio, Y. (2016),深度学习, 第1卷，MIT出版社，剑桥。Gottweis, J., Weng, W.-H., Daryin, A., Tu, T., Palepu, A., Sirkovic, P., Myaskovsky, A., Weissenberger, F. 等（2025），《迈向人工智能协同科学家》，*arXiv*预印本 *arXiv:2502.18864*。Greif, G., Griesshaber, N. & Greif, R.（2025），《用于历史文献光学字符识别、光学字符识别后校正和命名实体识别的多模态大语言模型》，*arXiv*预印本 *arXiv:2504.00414*。Griesshaber, N. & Ogilvie, S.（2025），《将手工业行会移植到殖民时期的拉丁美洲：大语言模型分析》，CEPR讨论论文20556，CEPR，巴黎和伦敦。Griliches, Z.（1990），《专利统计作为经济指标：综述》，《经济文献杂志》，28(4)，1661–1707。Guinnane, T., Harris, R., Lamoreaux, N. R. & Rosenthal, J.-L.（2007），《让公司回归其本位》，《企业与社会》，8(3)，687–729。Gundlach, H., Lynch, J., Mertens, M. & Thompson, N.（2025），《进步的代价：算法效率与人工智能推理成本的下降》，*arXiv*预印本 *arXiv:2511.23455*。Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D. d. L., Hendricks, L. A. 等（2022），《训练计算量最优的大语言模型》，*arXiv*预印本 *arXiv:2203.15556*。Humphries, M., Leddy, L. C., Downton, Q., Legace, M., McConnell, J., Murray, I. & Spence, E.（2025），《解锁档案：运用大语言模型转录手写历史文献》，《历史方法：定量与跨学科历史期刊》，1–19。Jayes, J.（2025），《如飞蛾扑火：从个体层面探讨*20*世纪瑞典的技术变迁》，博士论文（汇编），隆德大学经济与管理学院，隆德大学，隆德。Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A. 等（2020），《神经语言模型的缩放定律》，*arXiv*预印本 *arXiv:2001.08361*。Kim, J., Kang, S., Park, J., Kim, J. & Hwang, S. J.（2025），《解析大型视觉语言模型中图像到文本信息流的注意力头》，*arXiv*预印本 *arXiv:2509.17588*。Korinek, A.（2025），《用于经济研究的人工智能代理》，工作论文34202，美国国家经济研究局。Lagakos, D., Michalopoulos, S. & Voth, H.-J.（2025），《美国生活史》，技术报告，美国国家经济研究局。LeCun, Y., Bengio, Y. & Hinton, G.（2015），《深度学习》，《自然》，521(7553)，436–444。Lehmann-Hasemeyer, S. & Opitz, A.（2024），《19世纪及20世纪初德国资本市场的数据来源：挑战与机遇》，《德国经济评论》，25(4)，371–391。Lehmann-Hasemeyer, S. & Streb, J.（2018），《社会保障是否会挤出私人储蓄？以俾斯麦社会保险制度为例》，《欧洲经济史评论》，22(3)，298–321。Levchenko, M.（2025），《评估用于历史文献光学字符识别的大语言模型：数字人文的方法论框架》，*arXiv*预印本 *arXiv:2510.06743*。Long, J. & Ferrie, J.（2013），《自1850年以来大不列颠与美国的代际职业流动》，《美国经济评论》，103(4)，1109–1137。Luo, W.（2025），《将多模态大语言模型作为从历史文献中可靠提取信息的工具：瑞典专利卡片（*1945—1975*年）的数字人文研究》，硕士论文，乌普萨拉大学，人文与社会科学学科领域，艺术学院，ALM系。Maddison, A.（2006），《发展中心研究：世界经济 第*1*卷：千年视角与第*2*卷：历史统计》，发展中心研究，经合组织出版社。Mitchener, L., Yiu, A., Chang, B., Bourdenx, M., Nadolski, T., Sulovari, A., Landsness, E. C., Barabasi, D. L. 等（2025），《Kosmos：用于自主发现的人工智能科学家》，*arXiv*预印本 *arXiv:2511.02824*。Mokyr, J.（2011），《雅典娜的馈赠：知识经济的历史起源》，载于《雅典娜的馈赠》，普林斯顿大学出版社。

Moser, P.（2005），《专利法如何影响创新？来自19世纪世界博览会的证据》，《美国经济评论》，95(4)，1214–1236。Moser, P.（2011），《专利会削弱创新的本地化吗？来自世界博览会的证据》，《经济史杂志》，71(2)，363–382。Moser, P.（2012），《无专利创新：来自世界博览会的证据》，《法律与经济学杂志》，55(1)，43–74。Moulton, D. & Severen, C.（2025），《利用大语言模型挖掘历史数据》，《经济洞见》，10(4)，1–6。Rodrigues, E., Khalid, M., Nandi, S., Vrieze, A. & Yu, T.（2025），《印刷版书目数字化方法的基准测试》，《计算机与人文学科选集》，3，1360–1371。Ruggles, S.（2014），《用于人口研究的大型微观数据》，《人口学》，51(1)，287–297。Shen, Z., Zhang, K. & Dell, M.（2020），《一个包含复杂版面的日本历史文献大型数据集》，载于*IEEE/CVF*计算机视觉与模式识别会议研讨会论文集，548–549。Sockin, J., Ash, E. & Shukla, S.（2025），访谈，CESifo工作论文12229，CESifo。Streb, J.（2023），《专利法与经济绩效》，《经济史评论》，39(1)，3–26。Streb, J.（2024），《创新的计量经济史研究》，载于《计量经济史手册》，施普林格，2225–2245。Streb, J., Baten, J. & Yin, S.（2006），《德意志帝国1877—1918年的技术与地理知识溢出》，《经济史评论》，59(2)，347–373。Vafaie, M., Hertling, S., Banse-Strobel, I., Dubout, K. & Sack, H.（2025），《运用多模态大语言模型从档案记录中端到端提取信息》，载于第*34*届*ACM*信息与知识管理国际会议论文集，6075–6083。Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M. 等（2022），《大语言模型的涌现能力》，*arXiv*预印本 *arXiv:2206.07682*。Wu, T.-H., Biamby, G., Quenum, J., Gupta, R., Gonzalez, J. E., Darrell, T. & Chan, D. M.（2025），《视觉干草堆：以视觉为中心的大海捞针基准测试》，*arXiv*预印本 *arXiv:2407.13766*。Xie, Y., La Mela, M. & Tell, F.（2025），《多模态大语言模型辅助的历史文献信息抽取：瑞典专利卡片（1945—1975年）与ChatGPT案例》，载于第*9*届北欧与波罗的海国家数字人文会议（*DHNB 2025*），*2025*年*3*月*5*—*7*日，爱沙尼亚塔尔图，奥斯陆大学图书馆，1–15。Zhang, Y., Li, Y., Cui, L., Cai, D., Liu, L., Fu, T., Huang, X., Zhao, E. 等（2025），《人工智能海洋中的塞壬之歌：大语言模型幻觉综述》，《计算语言学》，1–46。

附录 A：提示词

图 A.1 专利条目提取提示

你是一名专业的数据提取人工智能。你的任务是处理德国帝国专利局登记簿中的单页内容，并将指定内容提取到结构完美的扁平 JSON 数组中。你必须以世界一流定量历史学家的精确度行事，确保绝对准确，并遵守以下规则。**历史准确性至关重要——提取内容时必须与原文完全一致，准确保留所有缩写、拼写和标点。**

I. 核心原则与范围

提取区域
你**只能**从页面上的两个主要内容栏中提取信息。

绝对忽略：
* **页眉：** 页面最顶部的文字（例如左上角和右上角的“Kl. 18. Nr. 55711.”或“Kl. 19. Nr. 57 185.”）。* **页码：** 位于页眉中央的任何数字。* **页脚信息：** 页面底部以星号（*）开头的任何文字、表示更正（“Berichtigung”）的文字，或包含独立数字（例如“2*”“1917”）且明显不属于上方专利条目的任何文字。

阅读与处理顺序
您的输出必须严格遵循以下顺序：。

1. 首先，从上到下处理**左侧**内容栏。
2. 其次，从上到下处理**右侧**内容栏。

最终的 JSON 数组必须按照提取项目的先后顺序排列。

II. 提取规则：识别和格式化内容

对于两个主要内容栏中的每个文本块，都必须将其归为以下类别之一，并按相应格式进行处理。

A. 类别标题

* **识别：**类别标题是表示新专利类别或子类的行。 * 主类别标题始终以“Klasse”开头，后接数字和标题（例如，“Klasse 19. Eisenbahn- und Brückenbau.”）。 * 子类标题通常只有子类代码本身，且往往单独成行（例如，“15a.”或“45i.”）。 * **提取规则：**从识别出的标题中，只提取类别编号/代码，不得包含其他内容。 * 如果标题是“Klasse 19. Eisenbahn- und Brückenbau”，则提取“19”。 * 如果标题是“18 b.”，则提取“18b”。 * 清理提取出的代码，删除末尾的句点或空格。 * **JSON 输出：**提取出的类别代码必须放在键为“category”的对象中。> **示例：** {"category": "3"}> **示例：** {"category": "3a"}

B. 专利条目

* **标准专利条目的识别：** 标准且完整的专利条目是一个满足以下条件的段落：* 以专利编号开头（例如，55711.）。* 编号前可能带有枚举标记（例如，1.）或括号中的匕首符号（例如，(†)）。* 包含有关专利持有人、地点、标题、描述等的信息。* 通常以日期和代码结尾（例如，1890年2月26日。A — 252.）。* **极少数情况下，日期/代码之后可能还会出现附加信息（例如：“奥地利申请中的优先权，日期为23/5 08，已获承认。”）——如果存在，请将其一并包括。**

* **所有条目的提取规则（标准条目和截断条目）：**
  * 原样捕获条目文本，保留所有原始字符、拼写和标点。* **换行处理：** * 合并跨换行符断开的连字符单词（例如，Hindersin-\nstrasse 变为欣德林斯街）。* 将条目段落内的所有其他换行替换为一个空格，以形成一行连续文本。

* **关键例外：截断条目**
  单个　　电子页面上下文意味着你会遇到不完整的条目。你**必须**按原样提取这些条目。

* **左上角截断的条目：**如果某栏从段落中间开始（承接上一页内容），请原样提取该部分条目。* **右下角截断的条目：**如果某条目在栏底部被截断（并延续至下一页），请原样提取该部分条目。* **同一页跨栏条目：**如果某条目从左栏底部开始，文本延续到右栏顶部，则必须在 JSON 数组中创建**两个独立的**条目。第一个对象包含左栏中的部分，第二个对象包含右栏中的部分。* **JSON 输出：**条目的格式化文本必须放在键为 "entry" 的对象中。> **示例：** {"entry": "55711. COOMES, M, F., 医生，以及 A. W. HYDE，在路易斯维尔，杰斐逊郡，肯塔基州，美国，第四街 6281/2 号或第二街 828 号；代理人：柏林的 C. PIEPER，欣德林斯街 3 号。钢的制备方法。1890年2月26日。A — 252."} ### C. 明确忽略的内容* ** 专利号列表：** 类别标题之后，通常会看到一段仅由专利号和代码组成的文本列表（例如，54571 A 3. — 54814 A 77. — 54881 A 37. — 54977 A 54.）。**不要**提取仅由专利号和专利代码组成的列表。这些无效专利条目只会出现在类别标题正下方。**这些列表可能跨越多行，并形成大段文本——请忽略整个文本块。**有效专利条目下方新起一行的日期属于该有效专利条目。* **参考信息（“Vgl.”）：**出现在专利条目和类别标题之间、以“Vgl.”开头的任何文本（例如，“Vgl. Kl. 12 P. R. 21188. – Kl. 45 P. R. 17470, 20344.”）。**不要**提取这些交叉引用信息。 ## III. JSON 输出规范（严格）你的全部输出必须是一个语法正确的 JSON 数组。任何偏差都会导致失败。* **数组结构：**输出必须以 [开头，并以]结尾。* **对象结构：**数组中的每个元素都是一个对象 {}。* **键值对：**每个对象必须恰好包含一个键值对。键必须是 "category" 或 "entry"。* **引号：****所有**键都**必须**用双引号括起来（例如，"entry"）。所有字符串值都必须用双引号括起来。* **分隔符：**对象之间必须用逗号 , 分隔。数组中的最后一个对象**不得**带尾随逗号。* **转义：**字符串值中的任何字面双引号 " 都必须转义为 \"。--- **完美输出结构示例：**[{ "category": "18" }, { "entry": "55711. COOMES, M, F., 医生，以及 A. W. HYDE，在路易斯维尔，杰斐逊郡，肯塔基州，美国，第四街 6281/2 号或第二街 828 号；代理人：柏林的 C. PIEPER，欣德林斯街 3 号。钢的制备方法。1890年2月26日。A — 252." }, { "entry": "56181. VERSEN, B., 在多特蒙德，海利格街 8 号。制造贝塞麦转炉底部的方法和设备。（30634号专利的补充专利。）1890年5月15日。A — 363." }, { "entry": "56195. ADAMS, CH.，在匹兹堡，宾夕法尼亚州戴蒙德街 110 号，美国；代理人：德累斯顿的 F. EDMUND THODE & KNOOP，阿马利恩街 5 号。从矿石中直接提取铁。1890年7月9日。A — 461." }, { "entry": "56205. — HERBERTZ, F. A.，在莱茵河畔科隆。带蒸汽射流的熔炉。1890 年 8 月 12 日。A — 461." }, { "category": "3a" }, { "entry": "(†) 55539. — FÜRSTENHEIM, C.，在柏林 C.，耶路撒冷街 15 号。可折叠胸像支架。1890年6月5日。A — 249." }]

最终指令： 请遵守所有规则。你的输出必须**仅为 JSON 数组**，不得包含任何其他内容。完成前，请再次检查输出，确保其完全符合上述规范且为有效 JSON。

注释：我们将提示词与图像语料库中的一张图像一同发送给 Gemini-2.5-Pro，以提取专利条目和技术类别。Gemini-2.5-Pro 按顺序阅读顺序返回了一个 JSON 对象，其中包含键 *entry* 和*category* 。

图 A.2 修复提示

你是一名数据验证员。你的任务是将德国专利条目分类为有效（‘1‘）或截断（‘0‘）。**关键区别：** - **有效**条目包含专利的结尾及日期/注册信息，无论其内容是从原专利文本的开头还是中间开始。 - **截断**条目以正常的专利编号开头，但在句中被截断，且**缺少**结尾/日期。**示例：** ****截断（标记为‘0‘）** - 以下条目开头正常，但在句中被截断： - "(†) 15322. — SCHWINTZER & GRÄFF in Berlin S., Sebastianstr. 18. Neuerungen an Hänge- und Steh-Schiebelampen, bestehend in einem gläsernen, mit leicht zerlegbarer Metallumkleidung" - "240938. Sulzer, Robert, Winterthur, Schweiz; Vertr.:" **有效（标记为‘1‘）** - 以下条目包含结尾及日期/注册信息，即使它们从句中开始： - "Neuerungen in der Herstellung der sog. Rahmenfür Schuhe und Stiefel. 28. Januar 1882. — A 648." - "der Ränder des Laufmantels von Luftradreifen. 17/8 13. — A 1492." - "zigerstr. 91. Zweibehälter-Drahtziehmaschine. 29/1 96." - "mit auswechselbaren, ineinandergreifenden Segmenten mit seitlichen Hohlräumen und einem mittleren, über den seitlichen Hohlräumen angeordneten, unterstützten Hilfshohlraum. 2/3 13. — A 2069. Priorität aus der Anmeldung in den Vereinigten Staaten von Amerika vom 27/7 12 anerkannt." 仅回复一个字符： - 如果条目包含结尾及日期/注册信息，回复‘1‘ - 如果条目在句中被截断且缺少部分结尾，回复‘0‘ 不要提供任何解释或其他文字。

注释：此提示词用于识别由 Gemini-2.5-Pro 从图像扫描件中提取专利条目后被截断的条目。系统会将一条提取出的专利条目追加到此提示词末尾，然后发送给 Gemini-2.5-Flash-Lite。所有专利条目均独立执行此操作。对于任何包含结束日期和注册代码的条目，模型都会返回“valid”（1），即使其开头缺失；而对于开头完整但在句中被截断的条目，模型会返回“truncated”（0）。随后，我们的处理流程会将截断条目与其下方相应的条目合并。

图 A.3 变量提取提示词

你将收到一条来自德国帝国专利局的文本记录。你的任务是充当高度精确的信息提取引擎。你必须仔细分析该记录，并提取指定字段，将输出格式化为单个有效的 JSON 对象。**核心原则：** * **精确性：**所有提取的文本都必须与源文本逐字符完全一致。不得添加、删除或更改任何字符，包括历史德语字符（例如，s）。* **结构：**专利条目几乎总是遵循特定顺序：‘专利_编号‘ ->‘名称‘ ->‘地点‘ ->(可选的‘代理人‘信息) -> ‘描述‘ -> 日期 -> (可选的补充说明或优先权登记)。利用这一可预测的结构指导提取。* **输出格式：**输出必须是单个 JSON 对象，包含以下键：‘专利_编号‘、‘名称‘、‘地点‘、‘描述‘和‘日期‘。如果无法可靠地找到某个字段的值，请使用字符串 "非数值"。--- **详细提取规则：** ** **1. ‘专利_编号‘（字符串）**

* **位置：**专利编号是记录开头的数字标识符。* **格式：**它由一至六位数字组成。* **规则：**必须忽略所有非数字前缀，包括列表序号（例如，‘1.‘、‘17.‘）、符号（例如，‘(†)‘）、破折号或空格。仅提取数字。* **示例：**对于‘(†) 2. 35321. — ...‘，必须提取‘"35321"‘。

2．‘名称‘（字符串）
* **位置：** 这是专利持有人，紧跟在‘专利_编号‘之后的名称。 * **规则：** 完全按照原文提取全名，但必须遵循以下关键规则： * **包括：** 公司形式（例如，‘采矿与锌冶炼股份公司‘）。 * **包括：** 复合名称的所有部分（例如，‘SCHIFFER & KIRCHER‘）。 * **包括：** 学术头衔（例如，‘Dr.‘、‘工程博士‘、‘Ing.‘），因为它们是个人正式名称的一部分。 * **包括：** 如果“Firma”作为公司名称的一部分出现，则包括该词（例如，‘Firma C. KESSELER‘）。 * **关键——排除职业：** 绝 NEVER 包括职业描述或职称，例如： * ‘Hofschlächter‘、‘Professor‘、‘Ingenieur‘、‘Kaufmann‘、‘Fabrikant‘、‘Mechaniker‘ * ‘Nachfolger‘、‘Marine-Lieutenant‘，或任何其他职称 * **关键——排除地点：** 名称字段中绝 NEVER 包括地点信息。对于多个专利持有人，仅提取名称，并按照原文使用‘und‘或‘,‘分隔。 * **关键：** 此字段仅用于专利持有人。不要提取代理人（‘代理人‘）的名称。如果代理人的名称出现，它位于文本后部，必须忽略。**3. ‘地点‘（字符串）** * **位置：** 这是专利持有人所在的地点，几乎总是直接跟在持有人的‘名称‘之后。通常由‘in‘或‘auf‘等介词引出。 * **规则：** 提取与专利持有人相关的完整地点短语。 * 包括所有细节：城市、州、国家（例如，‘克拉德诺，波希米亚‘、‘芝加哥，美利坚合众国‘）。 * **关键——多个专利持有人：** 当有多个专利持有人且每人都有自己的地点时，将所有地点合并为一个字符串，并按照原文使用‘und‘或‘,‘分隔。 * **示例：** 对于‘RUHM, H., in Sulko Zechc bei Pilsen und L. WOLF in Prag-Karolinenthal‘，地点应为‘Sulko Zechc bei Pilsen und Prag-Karolinenthal‘。 * **关键：** 不要提取代理人（‘代理人‘）的地点。正确地点是与上一步提取的‘名称‘直接关联的地点。**4. ‘描述‘（字符串）** * **位置：** 这是专利的主题内容。它位于持有人名称/地点之后、最终日期之前。 * ** 规则：** 提取核心描述文本。它可以短至一个词。必须应用以下两条排除规则： * ** 排除项 1：代理人信息：** 必须完全排除任何详细说明代理人的文本块。该文本块通常以‘代理人：‘开头，并包括其姓名、职称和地点。 * **排除项 2：补充说明：** 必须完全排除专利描述之后或末尾关于专利附录的任何括号说明。这些说明形式如‘（专利12345的补充说明。）‘或‘（专利17502的第II项补充说明。）‘，通常位于专利描述末尾或其后。 * 应用上述排除规则后，提取剩余文本作为描述。**5. ‘日期‘（字符串）** * **位置：** 日期位于条目末尾附近。 * **规则：** 按原样提取日期字符串。如果条目中有多个日期，正确日期始终是指德国注册的日期。--- **输入条目示例：** 35 321. — KARLIK, JOH., in Kladno, Böhmen; 代理人：C. FEHLERT & G. LOUBIER, 公司：C. KESSELER in Berlin SW., Anhaltstr. 6. 具有不同周向速度的摆动器，该速度在一次旋转过程中发生变化。1885年11月5日。**正确的 JSON 输出：** { "专利_编号": "35321", "名称": "KARLIK, JOH.", "地点": "克拉德诺，波希米亚", "描述": "具有不同周向速度的摆动器，该速度在一次旋转过程中发生变化。", "日期": "1885年11月5日" } **第二个输入条目示例：** 15948. — „VIEILLE MONTAGNE" 采矿与锌冶炼股份公司 in 阿尔滕贝格。环形跳汰机的改进（P. R. 15224的补充说明）。1881年3月25日。— A 789。

正确的 JSON 输出：
```
{
   "专利_id": "15948",
   "名称": "\"VIEILLE MONTAGNE\" 采矿与锌冶炼股份公司"                                                    ,
   "地点": "阿尔滕贝格",
   "描述": "环形跳汰机的改进",
   "日期": "1881年3月25日"
}
```

第三个输入条目示例（多位专利持有人和多个地点）：
15203. — RUHM, H.，位于比尔森附近的 Sulko Zechc，以及位于布拉格-卡罗利嫩塔尔的 L. WOLF；代理人：位于柏林 W. 区的 C. KESSELER，
莫伦街 63 I. 连续式平炉。1881年4月1日。— A 609.

正确的 JSON 输出：
```
{
   "专利_编号":    "15203",
   "name": "RUHM     ，H. 和 L. WOLF"    ,
   "location": "Sulko Zechc，位于比尔森和布拉格-卡罗利嫩塔尔",
   "description": "连续式平炉。",
   "date": "1881年4月1日"
}
```

要从中提取信息的条目：

注释：此提示词用于提取所需的五个变量。修复后的专利条目被附加到此提示词末尾，然后发送给 Gemini-2.5-Flash-Lite。大语言模型返回了一个 JSON 对象，其中包含以下键：专利_编号、名称、地点、描述和日期。随后，这些变量被附加到数据集中各自对应的专利条目中。此操作对所有专利条目分别独立执行。

附录 B：具有不同版式的特殊卷的提示词

图　　B.1 具有不同版式的特殊卷：专利条目提取提示词　　T

```
你是一名专业的数据提取人工智能。你的任务是处理德国帝国专利局登记簿中的单页，并将指定内容提取到结构完美、扁平化的 JSON 数组中。你必须以世界一流定量历史学家的精确度开展工作，确保绝对准确，并遵守以下规则。**历史准确性至关重要——提取所有内容时必须与原文完全一致，保留所有缩写、拼写和标点，严格按照其在原文中的写法。**



## I. 核心原则与范围

### 提取区域
你将**仅**从页面上的两个主要内容栏中提取信息。

### 务必忽略：
* **页眉：** 页面最顶部的文字（例如左上角和右上角的“Kl. 18. Nr. 55711.”或“Kl. 19. Nr. 57 185.”）。 * **页码：** 页眉中间位置的任何数字。 * **页脚信息：** 页面底部以星号（*）开头的任何文字、表示更正的文字（“Berichtigung”），或明显不属于上方专利条目的独立数字（例如“2*” “1917”）。



### 阅读与处理顺序
输出必须严格遵循以下顺序：
1. 首先，从上到下处理左侧内容栏。 2. 其次，从上到下处理右侧内容栏。


最终的 JSON 数　　组必须反映提取内容的先后顺序　　。

## II. 提取规则：识别与格式化内容

对于你在两个主要内容栏中遇到的每个文本块，必须将其归类为以下某一类别，并按相应格式进行处理。


### A. 类别标题

* **识别：** 类别标题是标示新专利类别的一行。* 主类别标题始终以“Klasse”开头，后接一个数字和标题（例如，“Klasse 19. Eisenbahn- und Brückenbau.”）。* **提取规则：** 从识别出的标题中，必须**仅**提取类别编号，不得包含其他内容。* 如果标题是“Klasse 19. Eisenbahn- und Brückenbau”，则提取“19”。* 清理提取出的代码，删除末尾的句点或空白。* **JSON 输出：** 提取出的类别代码必须放入键为“category”的对象中。> **示例：** {"category": "3"}> **示例：** {"category": "19"}



### B. 专利条目

* **标准专利条目的识别：** 标准且完整的专利条目是一个段落，其中：* 包含有关专利持有人、地点、标题、描述等的说明性文字。* 始终以格式为“P. R. XXXX.”的专利注册号结尾（其中 XXXX 是 1 至 9999 之间的数字，例如，“P.R. 6201.”）。


* **所有条目（标准条目和截断条目）的提取规则：**
  * 逐字保留条目文本，保留所有原始字符、拼写和标点。 * **换行处理：** * 合并跨换行符断开的连字符词（例如，Hindersin-\nstrasse 应合并为 Hindersinstrasse）。 * 将条目段落中的所有其他换行符替换为一个空格，使其成为一行连续文本

xt.

* **关键例外：截断条目**
  单页上下文意味着你会遇到不完整的条目。你**必须**按其呈现形式提取这些条目。* **左上角截断的条目：**如果某一栏以段落中间开始（承接上一页内容），请原样提取该部分条目。* **右下角截断的条目：**如果某条目在栏底部被截断（并延续至下一页），请原样提取该部分条目。 * **同一页跨栏条目：**如果某条目从左栏底部开始，其文本延续至右栏顶部，则必须在 JSON 数组中创建**两个独立的**条目。第一个对象应包含左栏中的部分，第二个对象应包含右栏中的部分。
```

* **JSON 输出：**条目的格式化文本必须放在一个键为 "entry" 的对象中。> **示例：** {"entry": "COOMES, M, F., 医生，以及 A. W. HYDE，位于路易斯维尔，第四街 6281/2 号或第二街 828 号，杰斐逊县，肯塔基州，美国；代理人：C. PIEPER，位于柏林 N.W.，欣德林斯街 3 号。炼钢方法。1890年2月26日。P. R. 6201。"} ### C. 明确忽略的内容 * **专利号列表：**类别标题之后，通常会看到一段仅由专利号和代码组成的文本列表（例如：54571 A 3. — 54814 A 77. — 54881 A 37. — 54977 A 54.）。**不要**提取这些仅由专利号和专利代码组成的列表。这些无效专利条目只会出现在类别标题正下方。**这些列表可能跨越多行，形成大段文本——请忽略整个文本块。**有效专利条目下方新起一行的日期属于该有效专利条目。* **参考信息（“参见”）：**任何以“Vgl.”开头的文本（例如“Vgl. Kl. 12 P. R. 21188. – Kl. 45 P. R. 17470, 20344.”），如果出现在专利条目与类别标题之间，**不要**提取这类交叉引用信息。 ## III. JSON 输出规范（严格）整个输出必须是一个语法完全正确的 JSON 数组。任何偏差都将导致失败。* **数组结构：**输出必须以 [开头，并以]结尾。* **对象结构：**数组中的每个元素都是一个对象 {}。* **键值对：**每个对象必须且只能包含一个键值对。键必须是“category”或“entry”。* **引号：****所有**键都**必须**用双引号括起来（例如 "entry"）。所有字符串值都必须用双引号括起来。* **分隔符：**对象之间必须以逗号 , 分隔。数组中的最后一个对象**不得**带有尾随逗号。* **转义：**字符串值中的任何字面双引号 " 都必须转义为 \"。---** 完美输出结构示例：**[{ "category": "18" }, { "entry": "COOMES, M, F., 医生，以及 A. W. HYDE，位于路易斯维尔，第四街 6281/2 号或第二街 828 号，杰斐逊县，肯塔基州，美国；代理人：C. PIEPER，位于柏林 N.W.，欣德林斯街 3 号。炼钢方法。1890年2月26日。P. R. 6201。" }, { "entry": "VERSEN, B.，位于多特蒙德，海利格街 8 号。制造贝塞麦转炉底部的方法和装置。（专利号 30634 的补充专利。）1890年5月15日。P. R. 6202。" }, { "entry": "ADAMS, CH.，位于匹兹堡，宾夕法尼亚州戴蒙德街 110 号，美国；代理人：F. EDMUND THODE & KNOOP，位于德累斯顿，阿马利恩街 5 号。直接从矿石中炼铁。1890年7月9日。P. R. 6203。" }, { "entry": "HERBERTZ, F. A.，位于莱茵河畔科隆。带蒸汽射流的熔炉。1890年8月12日。P. R. 6204。" }, { "category": "19" }, { "entry": "FÜRSTENHEIM, C.，位于柏林 C.，耶路撒冷街 15 号。可折叠胸像支架。1890年6月5日。P. R. 6205。" }]**最终指示：**遵守所有规则。输出**只能**是 JSON 数组，不得包含其他内容。提交前，请再次检查，确保输出严格符合上述规范且完全有效。

注释：用于从 1877–8年和1879年卷册的图像中提取专利条目和技术类别的提示词，其中有一个在线a
与其他卷的版式不同。更多信息请参阅图 A.1 中的注释。

图 B.2 具有不同版式的特殊卷：修复提示

你是一名数据验证员。你的任务是将德国专利条目分类为完整（‘1‘）或截断（‘0‘）。**完整（有效）**条目必须以日期和专利登记号结尾。由于光学字符识别错误，登记号格式可能有多种变体，包括 "P. R. XXXX."、"I. R. XXXX."、"R. XXXX."、"R. R. XXXX."、"P. R. Nr. XXXX"、"P. R. Nr. XXXX." 等（其中 XXXX 是 1 至 9999 之间的数字）。关键在于条目以日期和某种形式的登记号结尾。即使条目开头被截断，只要结尾格式正确，也应视为完整。**完整条目示例（应标记为‘1‘）：** - ‘HAHLWEG, C. 用于制造怀表石质表壳的工具。1877年7月22日。P. R. 80.‘ - ‘LÖWIG, G., & LÖWIG, Fr. 制备苛性碱和氧化铝制剂的方法。1877年7月3日。P. R. 93.‘ - ‘MORGAN RICHARDS, J. 制造穿孔药丸及其所需的机器。1877年8月3日。P. R. 134.‘ - ‘苯胺及其他芳香族叔单胺。1877年12月15日。P. R. 1886.‘（开头被截断但条目完整） - ‘SCHRÖDER, C. 通过羽毛打开、通过风压关闭的风车翼挡板。1878年1月13日。I. R. 1843.‘（光学字符识别错误：I. R. 代替 P. R.） - ‘BERNHARDI SOHN, Dr., DRAENERT, G. E. JACKSON 式安装在水平风轮上的链条传动转向装置。1877年11月13日。R. 1229.‘（光学字符识别错误：R. 代替 P. R.） - ‘SCHNEIDER & JACQUED. 带有下沉式中间叶片的涡轮机。1877年9月4日。R. 547.‘（光学字符识别错误：R. 代替 P. R.） - ‘FISCHER, G. A. 旋压式阀门，无论带有还是不带排空阀，借此可在受压的蒸汽管道或水管上钻孔。1877年10月20日。R. R. 874.‘（光学字符识别错误：R. R. 代替 P. R.） - ‘GROSSCHOPFF, DR. C. 用于厕所的自动消毒器。1878年3月31日。P. R. Nr. 3197‘（光学字符识别错误：P. R. Nr. 末尾没有句号）如果条目在结尾缺少日期或专利登记号，则属于**截断（无效）**条目。**截断条目示例（应标记为‘0‘）：** - ‘BARDIN, G. 将天然羽毛搓捻并使用这种羽毛的方法-Che‘ - ‘EHESTÄDT 和 ROBERT. 由某种平行六面体形状的照明设备上的活动臂-‘ **关键：**如有疑问，请标记为‘1‘（完整）。仅当条目明显被截断，且结尾缺少日期和专利登记号时，才标记为‘0‘。只能回复一个字符：- 如果条目完整（以日期和任何形式的登记号结尾，不论光学字符识别存在何种变体），回复‘1‘。- 如果条目被截断/无效（结尾缺少日期和登记号），回复‘0‘。不要提供任何解释或其他文本。

注释：用于 修复版式与其他卷册不同的1877–8年和1879年卷册中的专利条目 s.
更多信息请参阅图 A.2中的注释。

图B.3 具有不同版式的特殊卷：变量提取提示词

您将收到一条来自德国帝国专利局的文本记录。您的任务是充当一台高度精确的信息提取引擎。您必须仔细分析该记录，并提取特定字段，将输出格式化为单个有效的 JSON 对象。**核心原则：** * **精确性：**所有提取的文本都必须与源文本逐字符完全一致。不得添加、删除或更改任何字符，包括历史德语字符（例如 s）。* **结构：**专利条目几乎总是遵循特定顺序：“专利_编号” ->“名称” ->“地点” ->（可选的“代理人”信息） ->“描述” ->（可选的补充说明） ->“日期” ->（可选的补充说明） -> “P. R. XXXX.”（注册号）。请利用这一可预测的结构指导信息提取。* **输出格式：**输出必须是一个包含以下键的单个 JSON 对象：“专利_编号”、“名称”、“地点”、“描述”和“日期”。如果无法可靠地找到或识别某个字段的值，请使用字符串“非数值”。--- **详细提取规则：** **1. “专利_编号”（字符串）**

```
*   **地点：** 专利编号是条目末尾“P. R. XXXX.”格式中的数字标识符。
*   **格式：** 这是一个介于 1 到 9999 之间的数字。
*   **规则：** 只提取条目末尾“P. R. XXXX.”格式中的数字。
*   **示例：** 对于“... 1885年11月5日. P. R. 6201.”，必须提取“"6201"”。

**2. ‘名称‘（字符串）**
*   **地点：** 这是专利持有人的名称，显示在条目开头。
*   **规则：** **按原文准确提取全名。
    *   包括公司形式（例如，‘采矿与锌冶炼股份公司‘）。
    *   包括学术或职业头衔（例如，‘工程博士‘）。
    *   包括复合名称的所有组成部分（例如，‘SCHIFFER & KIRCHER‘）。
    *   **重要：**不要提取姓名旁边可能出现的职业或职位描述。仅提取实际的
      姓名和头衔。
*   **重要：**此字段仅填写专利持有人。不要提取代理人（‘代理人‘）的姓名。
    如果文本后面出现代理人的姓名，则必须忽略该姓名，不要填写在此字段中。

**3. ‘地点‘（字符串）**
*   **地点：**这是专利持有人的地点，几乎总是直接跟在持有人的“名称”后面。它
    通常由“in”或“auf”之类的介词引出。
*   **规则：**提取与专利持有人相关的完整地点短语。
    *   包括所有细节：城市、州、国家（例如，“克拉德诺，波希米亚”，“芝加哥，美利坚合众国”。
*   **重要：**请勿提取代理人（‘代理人‘）的地点。正确的地点是直接
    与上一步提取的‘名称‘关联的地点。

**4. ‘描述‘（字符串）**
*   **位置：**这是专利的主题内容。它是位于持有人名称/地点之后、
    最终日期*之前*的文本。
*   **规则：**提取核心描述文本。它可以短至一个词。你必须应用以下两条
    排除规则：
    *   **排除项 1：代理人信息：** 必须完全排除所有详细说明
    代理人的文本块。这类文本块通常以“Vertreter:”开头，并包含其名称、头衔和地点。
    *   **排除项 2：附录注释：** 必须完全排除所有关于专利附录的括号注释。这些
    注释的形式类似“(Zusatz zum Patent 12345.)”或“(II. Zusatz zum Patent 17502.)”，通常位于
    描述末尾、日期之前。
*   应用这些排除规则后，提取剩余文本作为描述。

**5. ‘日期‘（字符串）**
*   **地点：**日期位于条目末尾附近，在最后的“P. R. XXXX.”注册号之前。
*   **规则：**请*准确*提取日期字符串，保持其原样。如果条目中有多个日期，正确的日期
    始终是指德国注册的日期。

---

**输入条目示例：**
KARLIK，JOH.，位于克拉德诺，波希米亚；代理人：C. FEHLERT 和 G. LOUBIER，商号：C. KESSELER，位于柏林西南区安哈尔特街6号。
    周向速度在旋转过程中不断变化的摆动器。1885年11月5日。P. R.
    6201。

**正确的 JSON 输出：**
{
  "专利_编号": "6201",
  "名称": "KARLIK，JOH."，
  "地点": "克拉德诺，波希米亚"，
  "description": "具有不同周向速度、且周向速度在旋转过程中不断变化的摆动器。",
  "date": "1885年11月5日。"
}

**要从中提取信息的条目：**
```

注释：用于从 1877–8年和1879年卷册的专利条目中提取变量的提示词；这些卷册的版式与其他卷册不同。更多信息请参阅图 A.3中的注释。