面向用户决策的文档结构设计实验框架：食谱实证研究

Rina Kagaw a[1], Masaki Matsubara[1], Rei Miyata[2, 4], Takuya Matsuzaki[3], Yukino Baba[1], 和 Yoko Yamakat a[4]

[1] 筑波大学，筑波，日本 kagawa-r@md.tsukuba.ac.jp[2]名古屋大学，
名古屋，日本 [3] 东京理科大学，东京，日本[4]东京大学，东京，日本

文本类文档必须具备良好的质量，才能确保偏远地区的异步通信有效开展，尤其是在新冠肺炎疫情期间。然而，要界定一种能够改善普通读者决策的理想文档结构（内容与安排）颇具挑战。首先，仅仅汇集专家知识，无法确定适合不同读者的有用内容类型。其次，目前尚未建立从用户视角评估文档实用性的方法。本研究提出了一个实验框架，通过汇总普通读者的见解来识别文档中的有用内容。本研究以200份在线食谱为研究对象，并招募了1,340名业余厨师作为普通读者。所提出的框架识别出食谱中的六种有用内容。随后，多层建模表明，在已识别的六种内容中，在每个烹饪步骤末尾以小标题形式安排合适的食材或备注，能够显著提升食谱的实用性。本框架有助于通过文档进行传播设计。

关键词： 文档结构、决策、烹饪食谱、异步通信

1 引言：为普通人的决策制定文档结构规范

文档使人们能够以异步方式向偏远地区分享多样且复杂的信息，尤其是在全球抗击新冠肺炎疫情之际。文档应经过组织，以实现其交际目标 [1]（例如，小学教科书为年轻学生提供符合其年龄的知识）。文档的任何使用者都必须能够从文档中检索所需信息；同时，文档还需要具备良好的质量，使使用者能够根据文档的交际目标改变自己的想法和行为（图 1(A)）。我们关注的是文档结构的质量，其中包括内容（文档应包含哪些内容？）和安排（内容应如何组织？）[2]。

文档读者有时会根据对文档的理解做出决策。本研究着重于帮助普通读者进行决策，这是文档的交际目标。主要交际目标与读者决策无关的文档，例如小说 [3], ，不在本研究的范围内。

当文档旨在帮助普通读者做出决策时，专家知识可能并不足够，因为专家可能不了解普通读者认为哪些内容有用[1]。此外，文档结构往往由少数领域专家设计[4–6]。因此，需要一种为读者决策定义这类文档结构的框架。食谱就是旨在支持读者决策的一种文档。读者需要根据食谱中提供的细节（例如：“你可以用金枪鱼代替鲑鱼。”）、自身偏好和所处环境，灵活地决定如何烹饪。然而，专业厨师可能不了解业余厨师认为哪些信息有用。

要开发一种方法来确定有助于提升普通读者决策的优选文档结构，面临两项挑战。首先，不同读者认为对决策有用的内容类别往往需要通过实证确定，并且取决于他们各自的理解和动机；因此，仅凭收集专家知识可能无法确定这些类别。其次，读者并不总是在阅读后立即作出决定[2]，，这使得难以直接评估文档对决策是否有用。此外，评估指标也颇具挑战。当文档旨在帮助读者作出决定时，其质量不能仅依据读者理解的准确性来评估。此外，文档评估结果可能因读者的理解和动机而异，这也使得评估文档的实用性颇为困难。

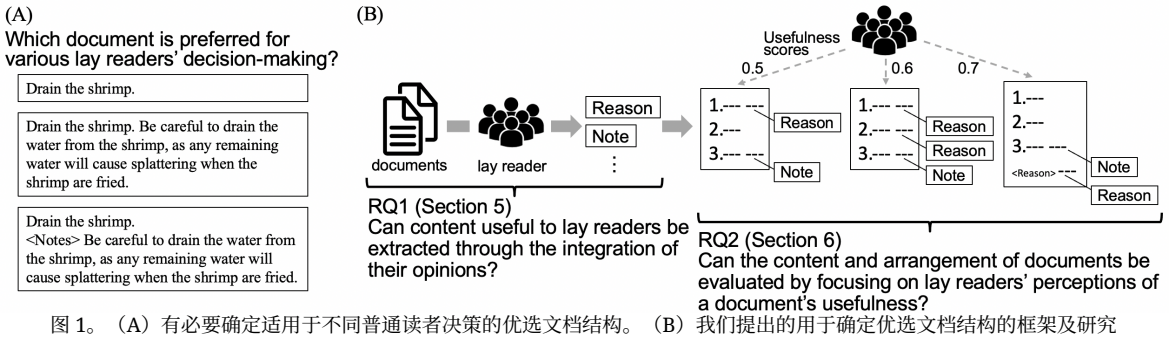

图 1。（A）有必要确定适用于不同普通读者决策的优选文档结构。（B）我们提出的用于确定优选文档结构的框架及研究问题。

2 研究目标与研究问题

本研究旨在确定有助于提升普通读者决策的优选文档结构。为实现这一目标，我们认为，利用普通读者的知识至关重要。作为起点，我们采用读者对文档实用性的感知作为评估指标，以判断该文档是否有助于他们作出决策。本研究的研究问题（图 1（B））如下：

- RQ1：能否通过整合普通读者的意见，提取对他们有用的内容?
- 研究问题2：能否通过关注普通读者对文档实用性的看法，来评估文档的内容与安排?

[1] 我们认为，可理解性、易读性及其他指标与实用性彼此相关，但属于不同的概念。即使一份文档易于理解和阅读，且不存在矛盾，如果读者最终认为它没有实用性，那么这份文档对决策的实用性可能就尚未得到恰当评估。因此，本文采用对读者的实用性作为一项指标（请参见第 2 节）。本文未直接讨论可理解性、易读性、一致性等方面。[2] 例如，食谱使用者有时会在读过食谱一周多之后，才应用其中的内容（例如，将提及不同食材替换方案的烹饪过程）。

本研究开展了一项实证研究。本文的实证研究以在线食谱合集为例，探讨这一文档体裁的内容；其交际目标是帮助读者灵活决策。尽管食谱通常按步骤编排得井然有序，但关于哪些内容（例如“为什么某个步骤是必要的”）和编排方式对不同读者有用，现有研究尚未进行充分探讨。有人可能认为，食谱应以图像或视频的形式记录。然而，在初步调查问卷中，我们请150名使用过食谱的人使用10分制李克特量表（10分：重要，1分：不重要），评价文字、图像和视频在食谱中的重要性；其平均值（标准差）分别为9.05（1.24）、8.63（1.34）和6.60（2.77）。因此，我们认为，食谱使用者比起图像或视频，更重视文字作为信息来源。于是，本文的实证研究以食谱文本为例，聚焦于文本类文档。

本研究首先提出一个实验框架，通过整合普通读者的意见，识别文本类文档中对普通读者有用的内容。随后，作为一项实证研究，我们使用所提出的框架开展实验，实际识别对普通读者有用的内容（第5节）。接着，考虑到读者理解方式和动机的多样性，我们根据普通读者对文档实用性的感知，评估所识别内容的安排（第6节）。

3 相关工作

3.1 文档形式化

研究方法。近期研究通过聚合专家意见，构建了文档形式化框架。例如，引言、方法、结果与讨论（IMRaD）[7] 已广泛应用于科学写作。语步分析已用于对研究论文的描述进行分类[8, 9]。主观资料、客观资料、评估与计划（SOAP）[10] 在全球范围内被广泛认可为日常临床实践中临床记录的理想框架。另一方面，基于普通读者见解聚合的文档形式化尚未得到探讨。随着众包平台日益普及，已有研究揭示了普通人（众包工作者）的处理能力所体现的实证知识。例如，普通人展示了不同画家画作的特征[11], 、暗示说谎的面部表情特征[12], 以及星星的形状[13]。此外，还有许多研究探讨普通人如何修改专业文档[14] ，以及如何根据文档搜索事实[15]。因此，普通人很有可能能够从文本类文档中识别出有用内容及其安排。我们认为，由众包工作者进行评估尤其适用于食谱，因为普通用户在日常生活中会撰写和阅读食谱。

研究对象。以往研究主要关注研究论文 [7, 8, 9] 或临床记录 [10]；这些文档旨在供专家阅读。据我们所知，关于影响普通读者灵活决策的文档（如食谱）之文档形式化的研究很少。由于研究结果与文档的安排有关，标题有助于恰当的记忆 [16–18], ，但标题是否有助于读者决策尚未得到检验。

研究目的。随着在线文档的数量和种类不断增加，不仅食谱，网页文档的信息检索[19, 20], 搜索[21, 22], 主题与体裁估计[23, 24], 以及网页文档的分类[25] 也在积极发展。另一方面，很少有研究考察在线文档[5]基于内容和安排的首选文档结构。

3.2 文本与文档评估指标

人们已制定了各种各样的评估指标，用于衡量特定文档的质量。例如，软件需求规格说明书必须满足 12 项要求[26]。临床系统评估工具用于评估患者教育文档的可理解性和可操作性，并在日常临床实践中使用[27]。以往研究讨论过小学教科书 [28], 、手册 [29–30], 技术传播文档 [31], 和讨论文件[6] 的质量。然而，这些文档的交际目标主要是确保读者正确理解内容或采取相应行动。

基于词语出现频率的可读性评分 [32–36], 被用于评估文本的可读性，也应用于软件开发中的代码可读性 [37, 38]。然而，这些评分并非旨在捕捉文档结构的变化，也就是信息呈现顺序的变化。一些研究 [39–41] 根据通过眼动追踪测得的阅读时间，对用户的句子处理成本进行心理建模。然而，由于这些研究只能评估相邻词语之间的相似性，因此尚不清楚这些方法能否用于更宏观地评估文档中句子语义内容的存在或安排对用户决策的影响。句子处理成本的心理语言学模型提出了“惊奇度”这一概念 [42], ，但这些成本无法直接评估文档的实用性。美国国家航空航天局任务负荷指数（NASA-TLX）[43], 是一种心理负荷测量工具，可用作文档评估的客观指标；但心理负荷这一概念关注的是任务难度或信息处理需求，而这两者都是决策过程中的预处理步骤。因此，NASA-TLX 不适用于本研究。

3.3 食谱文档研究

烹饪是人类的一项基本活动，与人们的身心健康以及国家和地区文化与身份认同密切相关 [44–46]。许多国家和语言都已建立包含用户提交文本的在线食谱合集 [47–53]。用户提交的在线食谱不仅能够促进人们日常生活中的身心健康。大量的用户提交在线食谱可以成为分享其国家和地区文化与身份认同、惠及后代的宝贵资源。另一方面，用户提交的在线食谱缺乏让作者获得读者关于食谱质量反馈的机会。因此，对食谱质量进行实验性界定是值得的。

鉴于机器学习等技术的快速发展，一些研究尝试从食谱中提取信息（例如，烹饪工具 [54] 或重要的烹饪过程 [55]），其他研究则根据图像 [56], 、食谱的标题和食材 [57], 生成文本烹饪说明，或开发食谱编辑系统[58]。作为这些研究的基础，关于烹饪和食材的本体 [59–63] 以及以流程图描述烹饪过程的语料库 [64] 也已建立。然而，这些研究主要关注烹饪步骤或食物本身。据我们所知，尚无研究深入关注超越食材或步骤的信息，例如烹饪操作的原因或替代项。一些研究指出，食谱使用者可能会从食谱中获取深入信息，例如依口味调整 [65], 、替代食材 [66], 以及烹饪步骤的含义 [67]。然而，这些描述是否确实能够提高食谱对读者的实用性，仍不明确。

4 参与者与材料

为了通过汇集普通读者的见解，以实验方式确定一种有助于改善普通读者决策的理想文档结构（内容与安排），我们选取食谱作为文档体裁的范例，其

交际目标是帮助读者进行决策。实验在计算机上进行，所有参与者均从 Lancers [68], 日本商业在线众包平台[3]招募。所有参与者均在线连接至实验网页界面并参加实验。参与者在预调查中表示，他们有烹饪经验并读过食谱；他们都没有在餐馆担任厨师，但每月至少读过几次食谱。因此，本研究将所有参与者视为普通读者中的业余厨师范例。本研究的每项实验均招募了不同的参与者。招募参与者时未设定任何排除标准。

每项实验均使用了从以下食谱中随机选取的若干食谱：由专业厨师撰写并在线发布的 9,820 份食谱 [71]（“专业食谱”），以及由食谱收集网站的用户提交并在线发布的 113,688 份食谱 [72] （“业余食谱”）。有关食谱随机选取的详细信息，请参见第 5 节和第 6 节中各项实验的说明。所有食谱均以日语撰写，所有实验也均以日语进行，尽管为了便于说明，本手稿中以英语呈现。

使用了 Perl 5.30、Python 3.7.2、NumPy 1.17.2、pandas 0.24.2、NetworkX 2.3、R 4.1.0、brms 2.15.0、lme4 1.1–27 和 lattice 0.20–38。本研究经筑波大学医学研究所伦理委员会批准（许可编号：1522）。

5 识别对普通读者有用的内容

5.1 框架概述

为建立一种旨在帮助普通人进行决策的理想食谱结构定义方法，本小节概述了用于识别食谱中有用内容的实验框架。我们提出了一项三步实验，通过收集、汇总并概念化普通读者的不同意见，整合这些意见，从而提炼出普通读者认为有用的内容类别。

首先，为了有效收集普通读者的意见，我们参考了人们在进行比较时能够更详细地识别物体特征的认知特征 [73]。参与者比较了随机选取的两份食谱。随后，他们指出第一份食谱比第二份更有用的原因，并解释第二份食谱比第一份更有用的原因。他们按照以下形式进行自由描述：“就 __ 而言，它对您更有用”。我们仅从这些回答中提取了与“__”对应描述中的唯一元素。随后，我们将这些元素称为“内容候选项”。

接下来，我们根据语义相似性对内容候选项进行聚合。此过程分为两个阶段：确定并捕捉内容候选项之间的相似性，以及根据这些相似性对内容候选项进行聚合。

最后，我们为聚合后识别出的每个类别命名，也就是对这些类别进行概念化。

5.2 普通读者有用内容类别的提取

我们使用食谱进行了这项三步实验，并确定了对普通读者有用的类别。在初步研究中，40名参与者自由描述了专业食谱或业余食谱为何更好。总共有592条描述认为专业食谱比业余食谱更有用，而558条描述则认为业余食谱更有用。基于这些结果，我们无法判断专业食谱还是

[3] 众包平台是供普通人在线完成任务的在线平台，广泛用于招募学术在线实验的参与者，因为这 20 世纪 90 年代末 [69, 70]。

业余食谱更有用。因此，我们使用了一个将专业食谱和业余食谱混合在一起的数据集。每个步骤的详细信息见补充材料A。

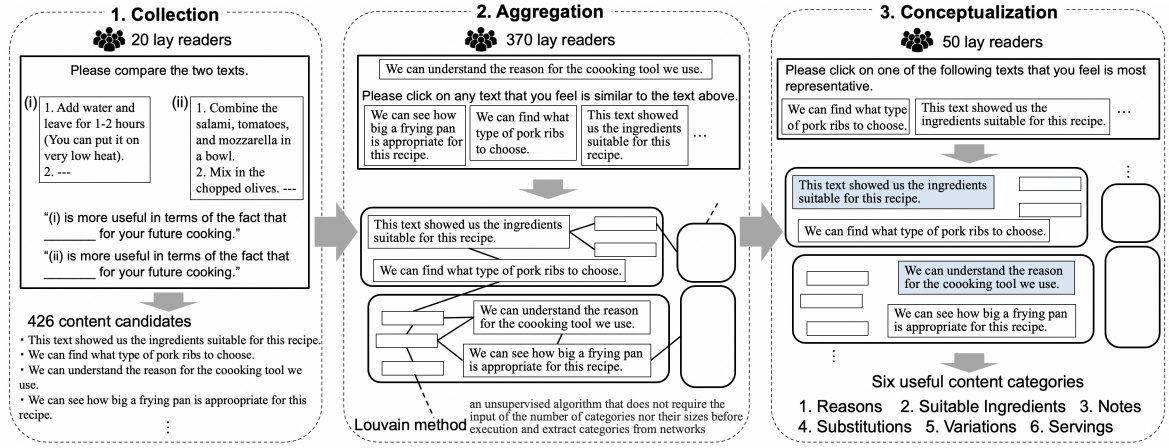

图2. 提取有用内容类别的三步实验程序。

5.2.1 收集

20名参与者比较了两份随机选取的食谱[4]。实验共使用134份随机选取的食谱。参与者指出第一份食谱在哪些方面比第二份更有用，并说明第二份食谱在哪些方面比第一份更有用。他们采用如下形式进行自由描述："就__而言，这对你未来的烹饪更有用"（图2，左上）。每位参与者在每项任务中比较两份食谱，并将此任务重复20次。每项任务和每位参与者看到的食谱配对都各不相同。

总共收集到616条回答。我们仅从这些回答中提取了与"__"对应的描述中的唯一元素（图2，左下）。共提取出426个此类元素。这些元素随后被称为"内容候选项"。

5.2.2 聚合

接下来，我们根据相似性对内容候选项进行了聚合。

我们评估了各内容候选项之间的相似性。然而，要开发能够自动判断具体程度不同的描述之间相似性的技术并不容易。例如，尽管以下两种描述都指出了合适的食材，但两者所用词语的具体程度却有很大差异。

- "这段文字向我们展示了适合这道食谱的食材"

[4] 专业食谱数据集和业余食谱数据集各自采用三层树结构对其子类别进行组织（即甜点——蛋糕——巧克力蛋糕）。两个数据集中，每道食谱最初都对应一个第三层子类别（例如巧克力蛋糕）。在此收集步骤中，每项任务呈现给每位参与者的两道食谱，均从专业食谱数据集和业余食谱数据集中对应同一第三层子类别的食谱中随机选取。这样做是因为我们假设，如果参与者比较两道差异很大的食谱（例如巧克力蛋糕食谱和咖喱食谱），就无法深入思考每道食谱为何有用。需要注意的是，我们将专业食谱和业余食谱混合后，从这个混合数据集中随机选取两道食谱；因此，参与者比较的两道食谱不一定分别来自专业食谱和业余食谱。

- “我们可以找到该选哪种猪肋排”

因此，我们通过多数表决的方式，请参与者判断各对内容候选项是否相似，并对内容候选项之间的相似性进行人工评估。共有370名参与者，每人完成25项任务。每项任务中，我们向参与者展示一个随机选取的内容候选项作为基准，并提供另外100个随机选取的内容候选项作为目标（图2，顶部中间）。随后，参与者选出所有他们认为与基准候选项相似的目标候选项。在本研究中，至少有四名参与者判定为相似的内容候选项对，即视为相似。

为了根据相似性聚合内容候选项，我们创建了一个网络，将每个内容候选项作为一个节点，并将一对被认为相似的候选项之间的连线作为边（图 2，底部中间）。我们对该网络应用了 Louvain 方法，这是一种从网络中提取类别的方法。Louvain 方法适用于聚合实证研究结果，因为通过聚合得到的类别数量无需预先确定。我们提取了 13 个类别，并将每个类别解释为实用食谱所需的类别。

5.2.3 概念化

最后，我们为聚合识别出的每个类别命名。

50 名参与者通过多数投票，从每个类别中选出最能代表该类别内其他内容候选项的内容候选项（图 2，右上方）。如果两个或多个内容候选项所得票数相同，则随机选出其中一个。每个类别中选出的内容候选项被用来为整个类别命名（图 2，右侧中间）。

作者将这 13 个类别分为两大类：食谱的内容（例如，“对烹饪过程相关备注的描述”）和食谱的表达方式（例如，“烹饪说明详细且易于遵循”）。对于名称表述为“对 __ 的描述”、“我们可以理解 __”或“__ 已写明”的类别，“__”被定义为与内容相关的类别。其他类别则涉及表达方式。13 个类别中，六个与内容有关（例如，“与烹饪过程相关的备注”），七个与表达方式有关。本研究旨在确定理想食谱结构（内容和安排）；因此，下文不讨论表达方式类别。

确定了以下六个内容类别：烹饪过程的原因、关于特别合适食材的信息、与烹饪过程相关的备注、使菜肴更美味的替代方案、如何安排（更改）食谱，以及如何上菜和摆盘。我们分别将它们称为原因、合适的食材、备注、替换、变体和份量（图 2，右下；表 1）。

5.2.4 对所得内容类别的评估

我们进行了定性评估，以确认普通读者是否认为每个内容类别都是有用的信息项。

作为第一项定性评估，我们使用 10 点李克特量表（10：有用 – 1：无用）向 400 名参与者提出以下问题：“您认为提及各个内容类别的食谱有用吗？”各内容类别的均值（标准差）分别为：原因 8.16（1.80）、合适的食材 8.48（1.51）、备注 9.02（1.22）、替换 8.39（1.65）、变体 8.33（1.68）以及份量 6.98（2.06）。

对于第二次定性评估，我们发放了一份开放式问卷，共有200名参与者参与，旨在询问：“实用烹饪食谱的文本中包含哪些类型的内容？”一名标注员确认了回答中是否包含各项内容类别（允许选择多个类别）。对于每个内容类别，其百分比的

提及实用食谱中内容描述的人数比例分别为：原因占17%，合适的食材占43%，备注占60%，替换占26%，变体占35%，份量占5%。换句话说，所有内容类别都被认为是有用的。

综上，整合普通读者意见后得到的内容类别被认为是恰当的。

表 1。 Exa各内容类别的句子示例。标注出内容的句子以粗体和斜体显示。

内容类别	句子示例	每份食谱的平均句子数（标准差）
原因	关火，并将铝箔放在碗上。 通过 放置铝箔，您可以安全地移开热碗 远离火源。	0.16 (0.40)
合适的食材	鱼，1片（80克） *马鲛鱼、鲷鱼、鲳鱼等。	1.16 (1.16)
备注	沥干虾。 注意沥干……中的水。 虾，因为任何残留的水都会在……时引起飞溅。 它们会被煎炸。	3.53 (2.61)
替换	一种茎和叶脉呈红色的香草。你可以 也可以使用芝麻菜、沼泽植物等。	0.36 (0.75)
变体	加入水并静置1–2小时。可以将其置于极低 火。	1.08 (1.43)
份量	沥干面条，将其与生菜一起放入碗中，并 充分涂上鸡肉味嚼酱。	0.36 (0.48)

5.3 讨论

本节表明，通过整合读者的意见，可以识别对读者有用的内容。第 5 节中的实验表明，根据实验程序的设计，可以提取文档中的有用内容，这些内容体现了普通读者的实证见解。

6 内容类别的数量与安排对食谱实用性的影响分析

本节旨在建立一种方法，以定义适用于普通人决策的理想食谱结构。对于第 5 节中识别出的每种有用内容类别，当句子数量会显著影响食谱实用性时，我们将考虑普通读者的不同理解和动机，明确各内容类别所对应句子的安排。

第 6.1 节详细介绍了预处理。第 6.2 节详细介绍了收集食谱实用性数据，以及读者的理解和动机的过程。第 6.3 节利用第 6.2 节获得的回答，详细分析了食谱结构对食谱实用性的影响。

目前尚未建立从用户视角评估文档实用性的方法。本研究旨在以读者为中心的方式评估食谱的实用性；随后通过问卷评估这种实用性。我们根据安排，将不含用于表明内容类别的句子的食谱，与含有表明内容类别的句子的食谱进行比较，以此计算实用性。

6.1 预处理

6.1.1 内容类别标注

在以下分析中，我们为专业食谱标注了内容类别；这些食谱未用于第 5 节的实验，因为我们假设业余食谱中包含少量用于表明内容类别的句子。

从随机选取的 100 份专业食谱中，我们剔除了所有提及其他食谱的食谱[5]；剩下 86 份，我们对其进行了标注，标明是否包含某个内容类别。两名经过培训的标注员依据提出的编码规则[6]，标记出表达各内容类别的句子。卡帕系数是衡量标注员间一致性的指标（1 表示完全一致，0 表示完全不一致），其值为 0.74。若两名标注员的判断不一致，他们会进行讨论并共同确定最终结果（表 1）。

6.1.2 实验食谱的制作方法

根据每份已标注的食谱（“原始食谱”），通过删除一条用于说明各内容类别的句子，制作出实验食谱（“已删除食谱”）（图3（A））。已删除的食谱中不含说明内容类别的句子。另从每份原始食谱中，通过将表示内容类别的句子重新编排为五种版式（安排）类型，制作出其他实验食谱（“重排食谱”，图3（B））[7]（A01– A05；图3（B））。这五种类型根据三个方面的组合确定：相关内容是在食谱末尾还是在每个烹饪步骤末尾提及，是否插入换行，以及是否使用标题。对于每个内容类别，最多根据每份原始食谱制作五种重排食谱（详情见补充材料B）。因此，本研究共讨论了36种食谱结构（6种内容类别 × 6 安排（原始食谱和五种重排食谱））。

[5] 在某个在线食谱合集上发表的一些食谱会提及同一合集中的其他食谱（例如，某份沙拉食谱会提及另一份用于制作该食谱所用调味汁的食谱）。本实验聚焦于完整食谱，因此排除了此类食谱。[6] 6558280 [7] 烹饪食谱标注手册：https://zenodo.org/record/ 这些安排来自一项初步实验：研究人员请 120 名有撰写食谱经验的业余厨师撰写食谱，要求每份食谱都包含各个内容类别，但不指定其安排方式。

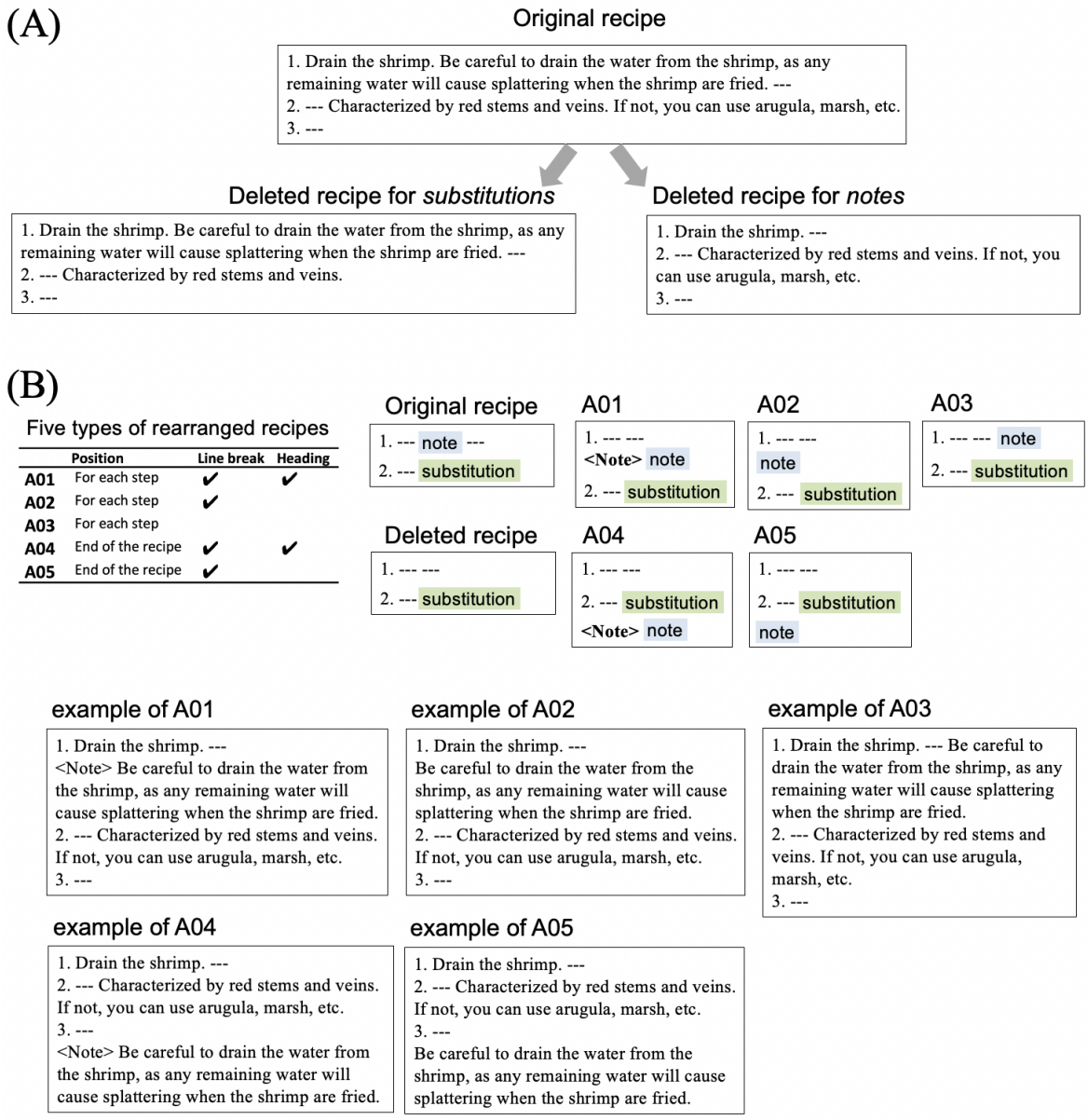

图3. 实验食谱概览。(A) 如果一份原始食谱包含多个内容类别，则会为每个内容类别创建已删除食谱和重排食谱。(B) 以备注为内容类别的原始食谱、已删除食谱和重排食谱概览。原始食谱和重排食谱是本研究讨论的安排，而已删除食谱则是讨论安排时的比较对象。

6.2 评估食谱实用性和读者背景知识的以读者为中心的实验流程

共有300名参与者完成了由三个步骤组成的调查问卷（图4）。

- **步骤 1，实用性**：首先，参与者同时浏览已删除食谱和相应的重排食谱或原始食谱。参与者使用 11 点李

- **步骤 2，理解程度**：我们试图以读者的理解作为判断食谱实用性的依据；然而，当时没有用于评估读者

    1. **背景知识**：句子中有多少个未知词？2. **可理解性**：你能理解这些句子吗？（10：能——0：

- **步骤 3，信息需求**：我们考察了读者阅读食谱时通常希望了解哪些信息，无论具体是哪道食谱。对于每

综合考虑实用性评分、读者的理解程度和信息需求，我们获得了 7,500 份答案。其中，5,574 份答案在所有评分中均无缺陷，并被纳入第 6.3 节所示的分析（详情见补充材料 C）。实用性评分的结果见表2。

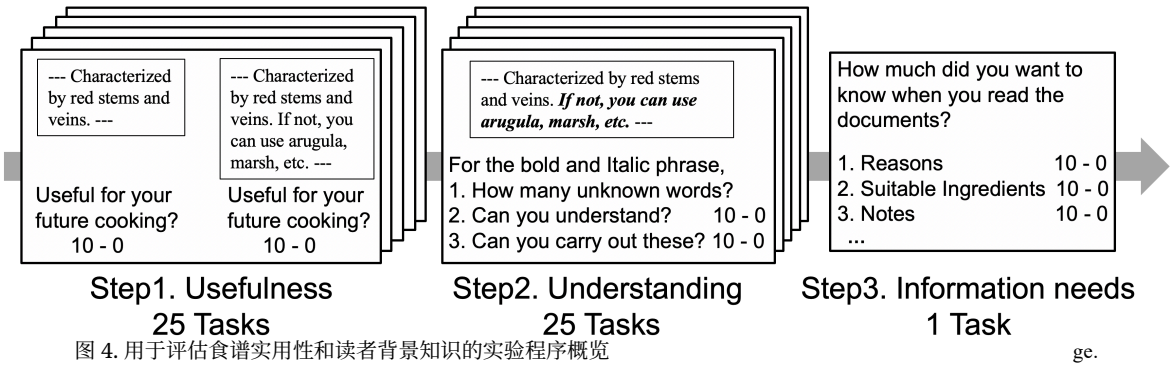

图 4. 用于评估食谱实用性和读者背景知识的实验程序概览 ge.

[8] 李克特量表的等级数仍存在争议。我们参考了[74] 相关研究，该研究表明，增加量表等级数可以降低偏度并使分布趋于正态，因此我们采用了 11 点量表。李克特量表原始数据的平均值（标准差）为 7.76（1.89），这一结果表明答案并未集中在“中间”选项。此外，如文稿所述，我们的分析并未使用量表的绝对值，而是比较已删除食谱与相应的重新排列文档或原始文档之间的数值差异。

表2. 各种食谱结构类型的实用性评分平均值和标准差。

内容类别	原始食谱	安排 A01	A02	A03	A04	A05
原因	0.41 (0.82)	0.28 (0.68)	0.46 (0.78)	0.37 (1.21)	0.26 (0.97)	0.39 (0.94)
合适的食材	0.26 (0.69)	0.46 (0.96)	0.33 (0.93)	0.37 (0.80)	0.34 (0.79)	0.31 (0.87)
备注	0.66 (1.08)	0.65 (1.28)	0.05 (1.85)	0.68 (1.27)	0.58 (1.19)	0.40 (1.31)
替换	0.35 (0.90)	0.31 (0.72)	0.27 (0.60)	0.35 (0.90)	0.37 (0.76)	0.32 (0.64)
变体	0.32 (0.89)	0.44 (0.89)	0.27 (0.81)	0.45 (1.02)	0.43 (0.73)	0.35 (0.85)
份量	0.35 (0.88)	0.38 (0.70)	0.40 (0.83)	0.38 (0.77)	0.51 (1.06)	0.44 (1.06)

6.3 内容类别的数量和安排对食谱实用性的影响分析 **ss**

6.3.1 方法

我们使用按照第 6.2 节所述步骤收集的所有答案，研究了共 36 种食谱结构对食谱有用性评分的影响。

首先，我们检查了参与者理解和信息需求中的哪些数据相对于实用性评分呈层级结构。各项数据的组内相关系数（ICC）和设计效应（DE）[9] [75] 见表3。总的来说，实用性评分被判定为主要在参与者理解和信息需求中，针对参与者未知词的存在或缺失呈现层级结构，因为所有数据的ICC均低于0.10，且参与者未知词的存在或缺失对应的DE最大。

表3. 各项数据的组内相关系数（ICC）和设计效应（DE）。

	ICC	DE		ICC DE
未知词的存在或缺失（1：存在 – 0：不存在）	0.03	17.3	信息需求的原因	0.01 2.17
未知词数量	0.08	10.1	适当替换的信息需求	0.00 1.00
理解	0.00	1.00	信息需求（备注）	0.00 1.50
技能水平	0.00	1.10	替换的信息需求 for 替换	0.00 1.00
			变体的信息需求	0.08 10.0
			份量的信息需求	0.01 2.20

因此，我们采用多层建模，并假设有用性评分会因参与者是否存在未知词而呈现层级结构。对于每个内容类别，我们尝试在假设参与者是否存在未知词具有层级结构的情况下，明确句子的排列方式，即在句子数量会显著影响食谱实用性的情况下，如何安排表示各内容类别的句子。我们针对每个内容类别构建了以下六种模型：原始模型和模型 A01–A05，并在各内容类别中选取广泛适用信息准则（WAIC）值最低的模型进行分析（详情见补充材料 D.1）。模型参数通过四条马尔可夫链蒙特卡罗（MCMC）链进行拟合，每条链包含 10,000 次迭代、9,000 个预热样本，抽稀参数为 1。我们针对每个内容类别比较了以下六种模型。

[9] 设计效应是一个组内数据的平均数量。A 同时考虑组内数据平均数量和组内相关系数的标准。$DE = 1 + (k^* - 1) \times ICC$。$k^*$ 表示 n 组内相关系数大于 0.1 或设计效应大于 2，说明数据具有层次结构。

- 原始模型：食谱实用性受到表示该内容类别的句子数量及其排列方式的影响。原始食谱排列方式中，表

- A01–A05模型：食谱实用性受到表示内容类别的句子数量及其排列方式的影响。根据 A01–A05 的安排

对于每个内容类别的六种模型，我们根据以下方程估计参数。

$$usefullness_{ij} = \alpha + \sum_{k=1}^{6} \beta_k n_{ik} + \eta_l^{(z_{ij})} + \gamma_l^{(z_{ij})} n_{il} + e_{ij} \ (k = 1,..,6)$$

$$e_{ij} \sim N(0, \sigma_e^2)$$

对于第 j 位参与者对第 i 个食谱给出的实用性评分 $usefullness_{ij}$，α 为截距，k 表示各内容类别对应句子的安排（原始值、A01、A02、A03、A04、A05），β_k （$k = 1, . . ,6$）为 n_{ik} 的系数，n_{ik} 表示六种内容类别各自的描述数量。具体而言，我们对 β_k 使用非信息先验，对 $\alpha \sim$ 使用 StudentT(3, 0, 2.5)，并将 StudentT(3, 0, 2.5) 作为 $\sigma_e \sim$ 的先验分布。$z_{ij} \in \{0, 1\}$ 表示各内容类别描述中是否存在未知词。$\eta_l^{(z_{ij})}$表示第 i 个食谱中第 l 种安排（例如，原始模型对应原始值，模型 A01 对应 A01）的安排对截距的随机效应。$\gamma_l^{(z_{ij})}$表示安排对 n_{il} 系数的随机效应。各模型使用与基础模型相同的先验。随机效应标准差的先验分布采用非信息先验；$LKJCholesky(1)$ 是 $l \in \{1, ..., 6\}$ 和 $g \in \{0, 1\}$ 对应的 $\gamma_l^{(g)}$ 与 $\eta_l^{(g)}$ 之间相关矩阵的先验。

6.3.2 结果

结果详情见补充材料 D.2。结果表明，在每个烹饪步骤末尾用小标题（例如“<Suitable Ingredients>”或“<Notes>”）呈现内容句子时，增加与合适的食材（$mean = 0.01$，95% 可信区间（置信区间）[0.00，0.03]）或备注（$mean = 0.03$，95% 置信区间[0.01，0.04]）相关的句子数量，会显著提高食谱的实用性。

本研究表明，所提出的框架能够通过实验找出可显著提升普通读者感知到的实用性的食谱内容及其排列方式。

6.4 讨论

本节通过关注普通读者对食谱实用性的感知，实验性地确定了这样的食谱：其中与合适的食材或备注相关的句子数量有所增加，并在每个烹饪步骤末尾以小标题呈现，从而显著提高了食谱对普通读者的实用性。采用标题进行排列的食谱被发现具有显著实用性，这一结果与先前研究 [16] 一致。然而，本研究首次以食谱为在线文档示例，通过实验验证了这一发现。

7 结论与未来工作

为了建立一种界定适用于普通人决策的理想文档结构的方法，本研究表明，通过整合普通读者的意见，可以识别对他们有用的内容。此外，本研究还通过关注普通读者对食谱实用性的感知，以实验方式确定了能显著提升食谱对普通读者实用性的内容及其安排。

这项以食谱为对象的研究结果表明，只要设计出适当的实验程序（即我们提出的三步框架），就有可能提取有用内容，也就是普通读者的实证见解，并根据这些已识别的有用内容设计文档。我们的结果有助于推进文档传播设计的实践与研究；所得的理想文档结构不仅可作为教育应用的依据，例如要求作者按照该理想文档结构撰写文档，也可用于文档撰写工具的研发，以生成理想的文档结构。

我们提出的三步法能够识别有用内容，并允许将普通人和专家都设定为用户（即，以专家厨师为用户应用我们的框架，就能获得专家的理想食谱结构）。我们提出的方法也不受语言限制。因此，我们的实验框架具有普适性和适用性。另一方面，我们提出的方法能否应用于其他文档，或其他文化背景下撰写的食谱，仍有待验证。

本研究首先采用读者感受到的实用性作为衡量指标，以评估文档是否对读者决策有用；因此，本研究并未评估所提出的文档结构在现实情境（即现实烹饪情境）中是否有助于决策。未来研究应纳入现实情境下基于任务的评估 [76]。本研究所假定的读者背景知识及其他多样的个体差异，仅通过所开发的问卷获取；因此，未来还需开发更准确的方法来估算读者的背景知识及其他多样的个体差异。本研究的另一项局限是，我们没有考察各个要素的具体写法。措辞等表达方式上的差异可能会影响文档的实用性。我们计划开展一项大规模用户实验，制作内容固定、表达方式各异的文档。

致谢

在本文中，我们使用了由乐天集团股份有限公司通过国立信息学研究所的 IDR 数据集服务提供的“乐天数据集”（https://rit.rakuten.com/data_release/）。本研究获得了 JST-Mirai Program Grant Number JPMJMI19G8、JSPS KAKENHI Grant Number JP19K19347 和 JP21H03993 的支持。

参考文献

[1] Miyata R. 机器翻译时代的受控文档编写。伦敦：Routledge，2021年，第12页。[2] Reiter E 和 Dale R. 2000。自然语言生成系统构建。剑桥：剑桥大学出版社，2000年，第59页。[3] Latham KF. 文档体验。《文献学报》2014年；70(4)：544–561。[4] Klischewski R. 公共行政中的电子文档管理本体。《商业》《流程管理期刊》2006年；12(1)：34–47。[5] Han S 和 Kageura K. 网页文档设计如何影响第二语言阅读过程：以日本行政文档为例。载于：第*38*届美国计算机协会国际传播设计会议论文集（编者 J Walwerna、D Hocutt 和 S Pigg S），德克萨斯州丹顿，2020年10月3日至4日，第1–7页。纽约：美国计算机协会。[6] O’Keefe DJ. 2004年。劝说理论与研究的趋势和前景。载于：劝说、社会影响与促从研究选读。波士顿：Pearson/Allyn and Bacon，2004年，第31–43页。[7] Day RA. 科学论文的起源：IMRaD格式。《美国医学写作协会期刊》1989年；4(2)：16–18。[8] John M. Swales. 类型分析：学术与研究场景中的英语。剑桥：剑桥大学出版社，1990年，第140–148页。[9] Vijay K. Bhatia. 类型分析：专业场景中的语言运用。伦敦：Longman，1993年，第29–34页。[10] Weed LL. 医疗记录、医学教育与患者护理：作为基本工具的问题导向记录。俄亥俄州克利夫兰：凯斯西储大学出版社，，1970年，第97–108页。[11] Takahama R、Baba Y、Shimizu N 等。Adaflock：面向人在环预测建模的自适应特征发现。载于：第三十二届*AAAI*人工智能会议论文集，路易斯安那州新奥尔良，2018年2月2日至7日，第1619–1626页。华盛顿特区：AAAI出版社。[12] Cheng J 和 Bernstein MS. Flock：混合式群体—机器学习分类器。载于：第*18*届美国计算机协会计算机支持协作工作会议论文集

协作工作与社会计算，温哥华，不列颠哥伦比亚省，2015年3月14日至18日，第600–611页。纽约：美国计算机协会。[13] Raddick MJ、Bracey G、Gay PL 等。星系动物园：探索公民科学志愿者的动机。《天文教育评论》2010年；9(1)：1539–1515。[14] Nie L、Jiang H、Ren Z 等。基于群体知识的代码搜索查询扩展。IEEE服务计算汇刊 2016年；9(5)：771–783。[15] Bian J、Liu Y、Agichtein E 等。在人群中寻找正确事实：社交媒体上的事实型问答。载于：第17届万维网国际会议论文集，中国北京，2008年4月21日至25日，第467–476页。纽约：美国计算机协会。[16] Sanchez RP、Lorch EP 和 Lorch RF Jr. 标题对文本处理策略的影响。《当代教育心理学》2001年；26(3)：418–428。[17] Sweller J、Van Merrienboer JJ 和 Paas FG. 认知架构与教学设计。《教育心理学评论》1998年；10(3)：251–296。[18] Lorch RF Jr 和 Lorch EP. 组织信号对文本处理策略的影响。《教育心理学杂志》1995年；87(4)：537–544。[19] Luo M、Mitra A、Gokhale T 等。利用神经检索器改进生物医学信息检索。载于：AAAI人工智能会议论文集，线上，2022年2月22日至3月1日，63卷(10)：11,038–11,046。[20] Gadge J 和 Bhirud S. 在分层向量空间模型中计算词项权重的上下文加权方法。《信息科学杂志》2021年；47(1)：29–40。[21] Chandwani G、Ahlawat A 和 Dubey G. 一种采用基于聚类的倒排索引进行文档检索的方法。《信息科学杂志》。提前在线发表，2021年6月14日。DOI：10.1177/0165551521101840。[22] Lai C-H 和 Chang Y-C. 基于群体信任和用户权重分析的文档推荐。《信息科学杂志》2019年；45(6)：845–862。[23] Matthews P 和 Glitre K. 利用情节摘要主题模型进行电影类型分析。《美国计算机协会信息科学与技术学会会刊》2021年；72(12)：1511–1527。[24] Feng J、Mu X、Wang W 等。一种基于三维战略图的主题分析方法。《信息科学杂志》2021年；47(6)：770–782。[25] Han K、Rezapour R、Nakamura K 等。一种基于小规模训练数据的领域专用文档分类专家参与方法。《美国计算机协会信息科学与技术学会会刊》。提前在线发表，2022年10月10日。DOI：10.1002/asi.24714。[26] ISO/IEC/IEEE 29148:2018标准。系统与软件工程——生命周期过程——需求工程。（访问日期：2023年4月3日）。[27] 医疗保健研究与质量局。患者教育材料评估工具（PEMAT）及用户指南（2013年，访问日期：2023年4月3日）。[28] Arai T、Matsuzaki T 和 Iwane H. 通过定量文本分析评估科学教科书，并分析学生阅读技能的心理测量结果。载于：第42届认知科学学会会议论文集，线上，2020年7月29日至8月1日，第3239–3245页。西雅图：Curran Associates。[29] Klischewski R. 公共行政中的电子文档管理本体。《商业流程管理期刊》2006年；12(1)：34–47。[30] Sugino H、Miyata R 和 Sato S. 文档结构形式化及文档元素自动识别：以汽车维修手册为例。载于：亚洲数字图书馆国际会议。瑞士尚贝里：Springer，Cham，2019年，第249–262页。[31] Li Q、Karreman J 和 de Jong MDT. 归纳式与演绎式结构的产品描述：对中国和西方读者的影响。《商业与技术传播杂志》2020年；34(4)：335–363。[32] Flesch R. 一种新的可读性测量标准。《应用心理学杂志》1948年；32(3)：221–233。[33] Smith EA 和 Senter RJ. 自动可读性指数，AMRL-TR，66–22。俄亥俄州赖特-帕特森空军基地：航空航天医学部，美国陆军国防技术信息中心（DTIC），1967年。[34] Coleman M 和 Liau TL. 为机器评分设计的计算机可读性公式。《应用心理学杂志》1975年；60(2)：283–284。[35] Sato S、Matsuyoshi S 和 Kondoh Y. 基于教科书语料库的日语文本可读性自动评估。载于：LREC论文集，摩洛哥马拉喀什，2008年5月26日至6月1日，第654–660页。巴黎：LREC，[36] Obi 3.0.1。（2018年，访问日期：2021年7月27日）。[37] Scalabrino S、Linares-Vasquez M、Poshyvanyk D 等。利用文本特征改进代码可读性模型。载于：2016年IEEE第24届国际程序理解会议（ICPC），得克萨斯州奥斯汀，2016年5月16日至17日，第1–10页。华盛顿特区：IEEE。[38] [illegible] 料库 载于：第29届认知科学学会会议论文集，田纳西州纳什维尔，2007年8月1日至4日，第947–952页。西雅图：Curran Associates。[40] Demberg V 和 Keller F. 来自眼动追踪语料库的数据作为句法处理复杂度理论的证据。《认知》2008年；109(2)：193–210。[41] Asahara M、Ono H 和 Miyamoto ET. “当代书面日语平衡语料库”的阅读时间标注。载于：第26届计算语言学国际会议技术论文集（编者 Y Matsumoto 和 R Prasad），日本大阪，2016年12月11日至16日，第684–694页。：大阪：COLING 2016组委会。[42] Hale J. 作为心理语言学模型的概率Earley解析器。载于：第二届计算语言学协会北美分会语言技术会议论文集，美国宾夕法尼亚州斯特劳兹堡，2001年6月1日至7日，第1–8页。宾夕法尼亚州斯特劳兹堡：计算语言学协会。[43] Hart SG. NASA任务负荷指数（NASA-TLX）：20年后。人因与工效学学会年会论文集 2006年；50(9)：904–908。[44] Fischler C. 食物、自我与身份认同。《社会科学信息》1988年；27(2)：275–292。[45] Min W、Bao B-K、Mei S 等。人如其食：探索丰富的食谱信息以进行跨区域食物分析。IEEE多媒体汇刊 2018年；20(4)：950–964。

[46] Mejova Y、Abbar S 和 Haddadi H。《数字时代对食物的迷恋：全球各地的 #foodporn》。载于：第十届 AAAI 国际网络与社交媒体会议，德国科隆，2016 年 5 月 17–20 日，第 250–258 页。加利福尼亚州帕洛阿尔托：AAAI Press。[47] Allrecipes.com。https://www.allrecipes.com/（1997 年，访问于 2023 年 4 月 3 日）。[48] Food.com。https://www.food.com/（2017 年，访问于 2023 年 4 月 3 日）。[49] Marmiton。https://www.marmiton.org/（1996 年，访问于 2023 年 4 月 3 日）。[50] Chefkoch GmbH。https://www.chefkoch.de/（2012 年，访问于 2023 年 4 月 3 日）。[51] Tudo Gostoso。https://www.tudogostoso.com.br/（2018 年，访问于 2023 年 4 月 3 日）。[52] Cookpad.com。https://cookpad.com/（1999 年，访问于 2023 年 4 月 3 日）。[53] Mediafort。Povarenok.ru，https://www.povarenok.ru/（2010 年，访问于 2023 年 4 月 3 日）。[54] Diwan N、Batra D 和 Bagler G。基于命名实体的食谱建模方法。载于：2020 年 IEEE 第 36 届国际数据工程会议研讨会（ICDEW），美国得克萨斯州达拉斯，2020 年 4 月 20–24 日，第 88–93 页。华盛顿特区：IEEE。[55] Yamakata Y、Imahori S、Maeta H 等。用于从烹饪过程文本中提取由食材、工具和动作组成的主要工作流程的方法。载于：ICME 研讨会论文集，美国华盛顿州西雅图，2016 年 7 月 24–31 日，第 1–6 页。华盛顿特区：IEEE。[56] Salvador A、Drozdzal M、Giro-i-Nieto X 等。逆向烹饪：根据食物图像生成食谱。载于：IEEE/CVF 计算机视觉与模式识别会议论文集，美国加利福尼亚州长滩，2019 年 6 月 15–20 日，第 10453–10462 页。华盛顿特区：IEEE。[57] Lee H、Shu K、Achananuparp P 等。RecipeGPT：基于生成式预训练的烹饪食谱生成与评估系统。载于：2020 年网络会议附属论文集，中国台湾台北，2020 年 4 月 20–24 日，第 181–184 页。纽约：美国计算机协会。[58] Li S、Li Y、Ni J 等。SHARE：分层辅助食谱编辑系统。arXiv。预印本，2021。DOI：arXiv:2105.08185。[59] Cui J、Zhang X 和 Zheng D。基于用户知识需求构建食谱知识图谱。《信息科学杂志》2023。01655515221151139。[60] Dooley DM、Griffths EJ、Gosal GS 等。FoodOn：促进全球食品可追溯性、质量控制和数据集成的协调食品本体。《NPJ 食品科学》2018；2(1)：23。[61] Celik Ertuğrul D。FoodWiki：通过语义网检查食品添加剂副作用的移动应用。《医学系统杂志》2015；40(2)：1–15。[62] Caracciolo C、Stellato A、Morshed A 等。AGROVOC 链接数据集。《语义网》2013；4(3)：341–348。[63] Snae C 和 Bruckner M。FOODS：一种以食物为中心、由本体驱动的系统。载于：第 2 届 IEEE 数字生态系统与技术国际会议，泰国彭世洛，2008 年 2 月 26–29 日，第 168–176 页。新泽西州皮斯卡塔韦：IEEE。[64] Yamakata Y、Mori S 和 Carroll JA。英语食谱流程图语料库。载于：第十二届语言资源与评估会议论文集，法国马赛，2020 年 5 月 11–16 日，第 5187–5194 页。巴黎：欧洲语言资源协会。[65] Sugiyama Y、Yamakata Y 和 Tanaka K。将相似食谱作为程序性信息进行概括，并找出食谱数据中细微但重要的差异。《DEIM 论坛》2013；1(1)：1–8（日文）。[66] Shidochi Y 等。结合烹饪动作特征，在烹饪食谱文本中寻找可替换食材。载于：2009 年 ACM 多媒体烹饪与饮食活动研讨会论文集，中国北京，2009 年 10 月 23 日，第 1–6 页。纽约：美国计算机协会。[67] Takahashi H、Saji N 和 Yanagisawa Y。新手与专家在烹饪操作和流程特征上的差异。《日本烹饪科学杂志》2018；51(6)：315–325。[68] Lancers, Inc. http://lancers.jp/（2008 年，访问于 2023 年 4 月 3 日）。[69] Peer E、Brandimarte L、Samat S 等。超越 Mechanical Turk：用于众包行为研究的替代平台。《实验社会心理学杂志》2017；70：153–163。[70] Liu D、Bias RG、Lease M 等。用于可用性测试的众包。《美国信息科学与技术学会会议录》2012；49(1)：1–10。[71] NHK。《今日烹饪》，https://www.kyounoryouri.jp/（1999 年，访问于 2023 年 4 月 3 日）。[72] 乐天集团股份有限公司。乐天数据集，https://doi.org/10.32130/idr.2.0，（2010 年，访问于 2023 年 4 月 3 日）。[73] Gentner D 和 Markman AB。类比与相似性中的结构映射。《美国心理学家》1997；52(1)：45–56。https://doi.org/10.1037/0003-066X.52.1.45[74] Leung S-O。4 点、5 点、6 点和 11 点李克特量表的心理测量属性与正态性比较。《社会服务研究杂志》2011；37(4)：412–421。DOI；10.1080/01488376.2011.580697。[75] Hox JJ、Moerbeek M、Van de Schoot R。《多层次分析：技术与应用》。第 3 版。纽约：Routledge，2017，第 4–6 页。[76] Colineau, N., Paris, C., and Linden, K. V. (2002). 对程序性指导文本的评估。载于：国际自然语言生成会议论文集，美国纽约，2002 年 7 月 14–16 日，第 128–135 页，纽约：计算语言学协会。

补充材料A

每项实验的设计都将每位参与者的工作时间限制在约一小时，同时考虑到参与者所承受的身体和心理负担。

补充材料A.1 收集

参与者人数定义的详细信息如下：（20名参与者）×（20项任务（食谱对））/（67对食谱）= 5.97，因此，平均而言，约有五名参与者检查了每对食谱。

补充材料A.2 聚合

在此聚合步骤中，参与者人数定义的详细信息如下：

426个内容候选项的组合数为 ${}_{6}C$ = 90242,525。370（参与者）× 25 （任务）× 100 （内容候选项对）/ 90,525 = 10.22，因此，平均而言，约有10名参与者检查了每对内容候选项。

这是因为未来的研究可能会涉及支持人们描述所提取的内容类别，因此将相似性的阈值（即“被四名或更多研究参与者判定为相似的内容候选项对，被视为相似”）设定为使内容类别的数量（见第3.1.3节）介于五类和九类之间[ref1], ，这正是人类利用短期记忆能够保持的元素数量。

[ref1] Miller，George A. 1956年。《神奇的数字七，加减二：关于我们处理信息能力的一些限制 》。Psychol. Rev. 63卷，第2期，81–97。https://doi.org/10.1037/h0043158

附录A.3 概念化

在此概念化　　概念化步骤中，关于内容类别，只有“使用工具的原因”被更名为“关于 c 的原因 烹饪过程”，基于本研究的意图。

补充材料B

表 A-1. 每种内容类别所创建的各类实验食谱数量

内容类别	已删除食谱	安排 A01	A02	A03	A04	A05
原因	13	13	9	5	13	13
合适的食材	40	40	40	40	40	40
备注	40	39	40	32	40	40
替换	21	21	21	21	21	21
变体	44	44	44	17	44	44
份量	30	30	30	1	30	30

创建已删除食谱时，只有在删除指示相应内容类别的句子后导致食谱语法不正确时，才会对表述进行最小限度的调整。创建重排食谱时，如果重排后的食谱与原始食谱相同，则不予创建。对于相关原始食谱中未包含的内容类别，我们不会创建已删除食谱或重排食谱。

因此，表 A-1 展示了每个内容类别中创建的各类实验食谱数量。为避免第 6.3 节实验中各内容类别安排的使用次数出现偏差，我们从表 1 所列的所有内容类别安排中随机选取了 13 份实验文档，并将其用于第 6.3 节的实验。对于内容类别安排为“原因_A02”“原因_A03”和“份量_A03”的情况，由于其对应的实验食谱不足 13 份，因此使用了全部相应的实验食谱。因此，本分析共讨论了 444 对已删除食谱与原始食谱，或五种重排食谱。

300（参与者）× 20（步骤 1 或 2 的任务）/444 个任务= 16.89，因此，平均每个内容候选项由 15 名以上的参与者进行检查，平均而言，检查了 每一对内容候选项。

补充材料 C

Raw data (7,500 answers)

participant_ID	usefulness score	unknown words	understandability	skill level	Information needs (Reasons)	Information needs (Suitable Ingredients)	…
Id_001	1	1	8	8	8	7	
Id_001	2	0		7	8	7	
⋮	⋮	⋮	⋮	⋮	⋮	⋮	
Id_001	0	1	10	6	8	7	
Id_002	-2	0	9	8	9	9	
Id_002	1		8	6	9	9	
⋮	⋮	⋮	⋮	⋮	⋮	⋮	

Each participant repeated 25 tasks, and 300 participants answered.

There were some defects.

Data used for Analysis in Section 6.3 (5,574 answers)

participant_ID	usefulness score	unknown words	understandability	skill level	Information needs (Reasons)	Information needs (Suitable Ingredients)	…
Id_001	1	1	8	8	8	7	
⋮	⋮	⋮	⋮	⋮	⋮	⋮	
Id_001	0	1	10	6	8	7	
Id_002	-2	0	9	8	9	9	
⋮	⋮	⋮	⋮	⋮	⋮	⋮	

图 A-1. 第 6.2 节中获得的原始数据的详细信息，以及第 6.3 节中使用的数据。

补充材料 D

补充材料 D.1 模型选择结果

各模型的广泛适用信息准则得分见表 A-2。所有 $\hat{R}$ 均低于 1.05。

表 A-2. 每个模型的广泛适用信息准则值

	原因	合适的食材	备注	替换	变体	份量
原始模型	14744.9	14742.1	**14674.0**	14742.0	**14748.5**	14752.0
模型 A01	14746.7	14745.4	14681.3	14742.1	14749.0	14753.0
模型 A02	**14744.2**	**14741.3**	14687.2	14741.3	14750.6	**14751.3**
模型 A03	14746.5	14733.3	14680.7	14741.7	14747.1	14751.6
模型 A04	14745.5	14741.7	14681.3	14746.8	14749.7	14748.6
模型 A05	14746.6	14742.1	14682.3	**14741.2**	14750.1	**14751.3**

补充材料 D.2 结果：参数拟合

各内容类别中广泛适用信息准则（WAIC）最小的模型的结果详情见表 A-3–A-8。

表 A-3. 原因的模型 A02 估计结果

总体效应：					
	估计值	估计误差	95%置信区间下限	95%置信区间上限	Rhat
截距	0.37	0.60	-1.01	1.63	1.00
原始值（β_1）	0.00	0.01	-0.01	0.02	1.00
A01（$\beta_{,}$）	-0.01	0.01	-0.03	0.00	1.00
A02（β_3）	-0.03	0.78	-1.83	1.67	1.00
A03（β_4）	0.01	0.01	-0.01	0.04	1.00
A04 (β_5)	-0.01	0.01	-0.03	0.00	1.00
A05 (β_6)	-0.01	0.01	-0.02	0.01	1.00
家族特定参数：					
西格玛	0.91	0.01	0.89	0.92	1.00
组级效应					
替代项中存在未知词					
标准差（η_{3}）	0.63	0.99	0.01	3.43	1.01
标准差（γ_{3}）	0.81	1.04	0.01	3.78	1.00
相关系数（η_{3}，γ_3）	0.02	0.63	-0.96	0.98	1.00

表 A-4. 合适食材的模型 A02 估计结果

总体效应：					
	估计值	估计误差	95%置信区间下限	95%置信区间上限	Rhat
截距	0.32	0.61	-1.15	1.53	1.00
原始值 (β_1)	-0.02	0.01	-0.03	-0.00	1.00
A01 ($\beta_{,}$)	**0.01**	0.01	**0.00**	**0.03**	1.00

A02 (β_3)	-0.01	0.01	-0.02	0.01	1.00
A03 (β_4)	0.05	1.12	-2.36	2.55	1.00
A04 (β_5)	-0.00	0.01	-0.02	0.01	1.00
A05 (β_6)	-0.00	0.01	-0.02	0.01	1.00
家族特定参数:					
西格玛	0.91	0.01	0.89	0.92	1.00
组级效应					
替代项中存在未知词					
标准差（η_4）	0.71	1.02	0.01	3.45	1.00
标准差（γ_4）	1.12	1.28	0.08	4.73	1.00
相关性（η_4, γ_4）	-0.08	0.60	-0.98	0.95	1.00

表 A-5。 备注的原始模型估计结果

总体效应:					
	估计值	估计误差	95%置信区间下限	95%置信区间上限	Rhat
截距	0.31	0.63	-1.27	1.46	1.00
原始值 (β_1)	-0.05	0.81	-1.96	1.71	1.00
A01 (β,)	**0.03**	0.01	**0.01**	**0.04**	1.00
A02 (β_3)	-0.04	0.01	-0.05	-0.02	1.00
A03 (β_4)	**0.03**	0.01	**0.01**	**0.04**	1.00
A04 (β_5)	**0.02**	0.01	**0.01**	**0.03**	1.00
A05 (β_6)	0.01	0.01	-0.01	0.02	1.00
家族特定参数:					
西格玛	0.90	0.01	0.88	0.92	1.00
组级效应					
替代项中未知词的存在					
标准差（η_1）	0.66	1.00	0.01	3.42	1.00
标准差（γ_1）	0.83	1.01	0.04	3.81	1.00
相关系数（η_1, γ_1）	0.01	0.63	-0.98	0.98	1.00

表 A-6. 替换模型 A05 的估计结果

总体效应:					
	估计值	估计误差	95%置信区间下限	95%置信区间上限	Rhat
截距	0.32	0.71	-1.24	1.79	1.00
原始值（β_1）	-0.01	0.01	-0.02	0.00	1.00
A01（β, ）	-0.01	0.01	-0.02	0.01	1.00
A02（β_3）	-0.01	0.01	-0.03	0.00	1.00
A03（β_4）	-0.00	0.01	-0.02	0.01	1.00
A04（β_5）	-0.00	0.01	-0.02	0.01	1.00

A05 (β_6)	-0.05	0.72	-2.08	1.50	1.00
家族特定参数:					
西格玛	0.91	0.01	0.89	0.92	1.00
组级效应					
替代项中未知词的存在					
标准差（η_6）	0.70	1.07	0.01	3.67	1.01
标准差（γ_6）	0.75	1.14	0.01	4.02	1.00
相关系数（η_6, γ_6）	-0.04	0.63	-0.98	0.97	1.00

表 A-7。 变体原始模型的估计结果

总体效应:					
	估计值	估计误差	95%置信区间下限	95%置信区间上限	Rhat
截距	0.37	0.60	-1.03	1.65	1.00
原始值 (β_1)	-0.03	0.71	-1.57	1.54	1.00
A01 ($\beta_{,}$)	0.00	0.01	-0.01	0.02	1.00
A02 (β_3)	-0.01	0.01	-0.03	0.00	1.00
A03 (β_4)	-0.00	0.01	-0.02	0.01	1.00
A04 (β_5)	0.00	0.01	-0.01	0.02	1.00
A05 (β_6)	-0.00	0.01	-0.01	0.01	1.00
家族特定参数:					
西格玛	0.91	0.01	0.89	0.92	1.00
组级效应					
替代项中存在未知词					
标准差（η_1）	0.62	0.94	0.01	3.31	1.00
标准差（γ_1）	0.64	0.90	0.01	3.23	1.00
相关系数 (η_1, γ_1)	0.00	0.63	-0.98	0.98	1.00

表 A-8_1. 份量的模型 A02 估计结果

总体效应:					
	估计值	估计误差	95%置信区间下限	95%置信区间上限	Rhat
截距	0.38	0.53	-0.82	1.58	1.00
原始值（β_1）	-0.00	0.01	-0.01	0.01	1.00
A01 ($\beta_{,}$)	0.00	0.01	-0.01	0.02	1.00
A02 (β_3)	-0.01	0.66	-1.64	1.33	1.00
A03 (β_4)	0.00	0.03	-0.05	0.07	1.00
A04 (β_5)	0.01	0.01	-0.00	0.02	1.00
A05 (β_6)	0.00	0.01	-0.01	0.02	1.00
家族特定参数:					
西格玛	0.91	0.01	0.89	0.92	1.00
组级效应					

替代项中存在未知词					
标准差（η_3）	0.60	0.92	0.01	3.22	1.00
标准差（γ_3）	0.64	0.97	0.01	3.42	1.00
相关性（η_3, γ_3）	-0.04	0.62	-0.98	0.97	1.00

表 A-8_2。 份量的模型 A05 估计结果

总体效应:					
	估计值	估计误差	95%置信区间下限	95%置信区间上限	Rhat
截距	0.41	0.61	-0.81	1.72	1.00
原始值 (β_1)	-0.00	0.01	-0.01	0.01	1.00
A01 (β,)	0.00	0.01	-0.01	0.02	1.00
A02 (β_3)	0.00	0.01	-0.01	0.02	1.00
A03 (β_4)	0.01	0.03	-0.04	0.07	1.00
A04 (β_5)	0.01	0.01	-0.00	0.03	1.00
A05 (β_6)	0.01	0.63	-1.29	1.57	1.00
家族特定参数:					
西格玛	0.91	0.01	0.89	0.92	1.00
组级效应					
替代项中未知词的存在					
标准差（η_6）	0.61	0.88	0.01	3.18	1.00
标准差（γ_6）	0.66	0.92	0.01	3.35	1.01
相关系数（η_6, γ_6）	0.02	0.63	-0.97	0.98	1.00