

[3][1-2]

烹饪创造力：一种通过结构化表示增强 LLM 创造力的认知启发方法

Moran Mizrahi¹ Chen Shani² Gabriel Stanovsky¹
Dan Jurafsky² Dafna Shahaf¹

¹Hebrew University of Jerusalem ²Stanford University

{ moranmiz, dshahaf, gabriel.stanovsky }@cs.huji.ac.il { cshani, jurafsky }@stanford.edu

Abstract

大型语言模型 (LLMs) 在无数任务中表现出色,但在创造力方面却有困难。在本文中,我们引入了一种新方法,将 LLMs 与结构化表示和认知启发的操作相结合,以生成更有创意和多样化的想法。我们的创造力概念超越了表面的令牌级别变化;相反,我们明确地重组已有观点的结构化表示,使我们的算法能够有效地探索更抽象的想法领域。我们在烹饪领域通过 *DishCOVER*, 一个模型,来展示我们的方法。与 GPT-4o 的结果相比,我们模型的实验显示出更大的多样性。领域专家的评估显示,我们的输出结果,主要是连贯且可行的烹饪创作,在新颖性方面显著超过了 GPT-4o,因此在创造性生成方面表现更为出色。我们希望我们的工作能够激发更多关于 AI 结构化创造力的研究。

1 介绍

大型语言模型 (LLMs) 擅长生成流利的相干文本和涉及广泛世界知识的进行任务。然而,它们往往难以生成真正有创意的想法 (Franceschelli and Musolesi, 2024; Chakrabarty et al., 2024; Tian et al., 2024b; Zhao et al., 2024)。

在创造力研究中,创造性的成果 s 是通常被定义为那些既是新颖的(出乎意料的和原始)又是有价值的(有用的、相关、有意义或有效) (Mumford, 2003; Boden, 2004)。然而,由于其依赖于大量现有数据集,LLM (大型语言模型) 固守地遵循学习到的模式,使得它们容易产生重复或可预测的输出而缺乏真实的新颖性。讽刺的是,尝试对 LLM 进行明确地指导以“更有创造力地思考”经常,使得它们产生无效或虚构的解决方案,这可能会误导不知情的用户 (Wang et al., 2024a; Jiang et al., 2024)。这些限制使得 LLM 的创造性生成成为一个持续的挑战。

LLMs 的温度参数控制随机性的程度,常被称为创意参数。然而,创意包含远不止仅仅随机性;最近的一项研究 (Peeperkorn et al.,

2024) 发现,尽管更高温度 s 与增加的新颖性存在弱相关,他们的实际对总体创意的影响保持细微且有限。

许多最近的研究表明,将结构化知识(如知识图谱)与 LLMs 结合可以显著提高其性能 (Pan et al., 2024; Wang et al., 2024b; Feng et al., 2023; Sun et al., 2023)。特别是,有一系列研究利用 LLM 将文本解析成结构化表示,操纵这个表示,且(可选地)再次应用 LLM 将结果翻译成自然语言,特别是在推理和推断任务中使用 (Yang et al., 2023; Besta et al., 2024; Zelikman et al., 2023; Zhang et al., 2025)。

在这项工作中,我们展示了令人惊讶的是,融入结构也可以提高大语言模型的创造力和多样性。我们强调我们所指的创造力和多样性并不是词汇(标记)层面的;相反,我们希望模型能在更抽象的层面上,在概念的领域内(或者可以说是“思想的景观”中)展现创造力。

我们的方法如图 1 所示。与基于解析的方法 (Tian et al., 2024a; Zhao et al., 2023; Li et al., 2024) 类似,我们首先从文本输入中推导出结构化表示。接下来,我们操作这些结构化表示,以达到创意空间中的创意部分。我们从人类的创意过程汲取灵感,专注于 **重组**——创造力研究中的一个基本原则,认为新颖的想法常常通过以意想不到的方式融合现有概念而出现 (Guilford, 1967; Utterback, 1996; Ahuja and Morris Lampert, 2001)。例如,将披萨的制作方法 with 阿尔法霍雷斯饼干的风味结合,可能会产生一种全新的“阿尔法霍雷斯披萨”,而将沙发与书架结合可能会产生多功能家具。我们从重新组合的空间中抽样想法,并根据新颖性和价值对其进行评估 (Finke et al., 1996; Sawyer and Henriksen, 2024)。

我们在烹饪领域展示了我们的范式,推出了 *DishCOVER*, 这是一个用于创意生成食谱的模型。图 2 展示了由该模型生成的重组示例。

除烹饪的直接范畴之外,我们相信这一范式在扩大创造性和多样化生成到广泛领域方面具有潜力,从产品设计、叙事构建、科学发现到艺术创作。总之,我们的贡献是:

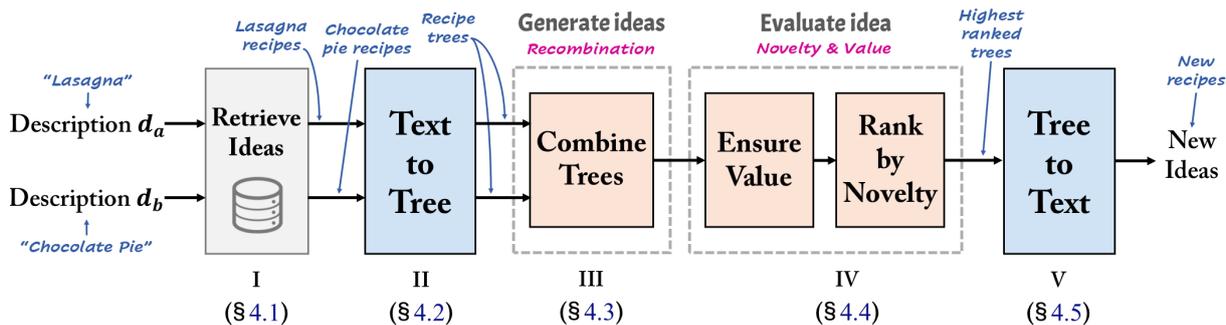


Figure 1: 用于创意食谱生成的 *DishCOVER* 管道（以蓝色阴影标识的为基于 LLM 的组件）将两个创意描述作为输入。每个描述被映射到一组特定的食谱 (§ 3.1)，这些食谱被解析成树结构表示 (§ 3.2)。然后，这些树使用最小编辑距离算法 (§ 3.3) 进行组合，进行价值评估，并根据新颖度评分进行排名 (§ 3.4)。最后，得分最高的树被翻译回自然语言食谱 (§ 3.5)。

- We introduce a novel paradigm to enhance LLM creativity by extracting structured representations from natural language and manipulating them (inspired by cognitive processes), going beyond mere token-level variability.
- We demonstrate our approach in the culinary domain with *DishCOVER*, a model that recombines recipes to generate creative ones.
- We curate a 5K-recipe dataset generated by *DishCOVER*, providing a valuable resource for future work on creative generation. We make both the code and data publicly available.¹
- Through systematic experiments, we show that *DishCOVER*'s generations are significantly more diverse compared to baseline SOTA LLM outputs. Most recipes generated by both models are deemed valuable (appropriate and coherent), although the baseline achieves better scores on an open-ended task. Most importantly, our outputs significantly surpass the baseline in terms of novelty, resulting in more creative culinary ideas. These findings are supported by both automated metrics and domain expert evaluations.

2 背景：人类创造力

人类创造力领域已经被广泛研究，识别出推动创新的众多原则。在设计我们的模型时，我们依赖以下原则：

Generation & Evaluation. 创造性思维的一个常见而有效的模型是两阶段过程 生成 & 评估，它表明创造力始于发散性思维（自由创意产生），然后是聚合性思维，在这其中最前景

的想法被选择和完善 (Finke et al., 1996; Sawyer and Henriksen, 2024)。我们将其作为我们模型的概念主干，实施一个生成组件来产生广泛的想法集合，随后是一个评价组件，以识别那些具有最大创造潜力的想法。

Recombination of Ideas. 我们的工作基于一个突出的创意生成方法：重组，即从现有创意中合并元素以创造新颖的概念 (Guilford, 1967; Koestler, 1964)。我们的模型有策略地重组现有种子创意对中的元素，以激发出乎意料的关系。

Creativity Assessment: Novelty & Value. 在生成了许多想法之后，挑战在于确定哪些是真正有创意的。众多研究已经考察了评估人类和计算系统创造力的复杂性 (Said-Metwaly et al., 2017; Lamb et al., 2018)。一个被广泛接受的创造力定义将其构架为新颖性与价值的交汇点 (Mumford, 2003; Boden, 2004, 2009; Lamb et al., 2018)。新颖性确保一个想法是令人惊讶或非传统的，而价值意味着它在其预期的上下文中有用的。

自动测量新颖性和价值。 新颖性可以通过识别一个想法在数据集中有多不常见来评估 (Heinen and Johnson, 2018; Kenett, 2019; Doboli et al., 2020)。然而，价值评估高度依赖于具体领域，通常被认为是计算创意的“圣杯” (Boden, 2004; Ritchie, 2007; Jordanous, 2012)。因此，我们将价值评估视为一个特定领域的任务。

创新通常涉及将现有的想法结合起来以创造新颖的想法。这个过程通常被称为“概念融合”或“创造性重组”，它是创新的核心，也是我们工作的重点。我们现在介绍我们公式的**关键要素**。给定一个可以以结构化格式表达想法的领域（例如，烹饪食谱、说明手册、计算机程序），令 \mathcal{I} 表示该领域内所有可能想法的理论集合——包括已存在的和尚未被发现的。 \mathcal{I}

¹<https://github.com/moranmiz/Cooking-Up-Creativity>

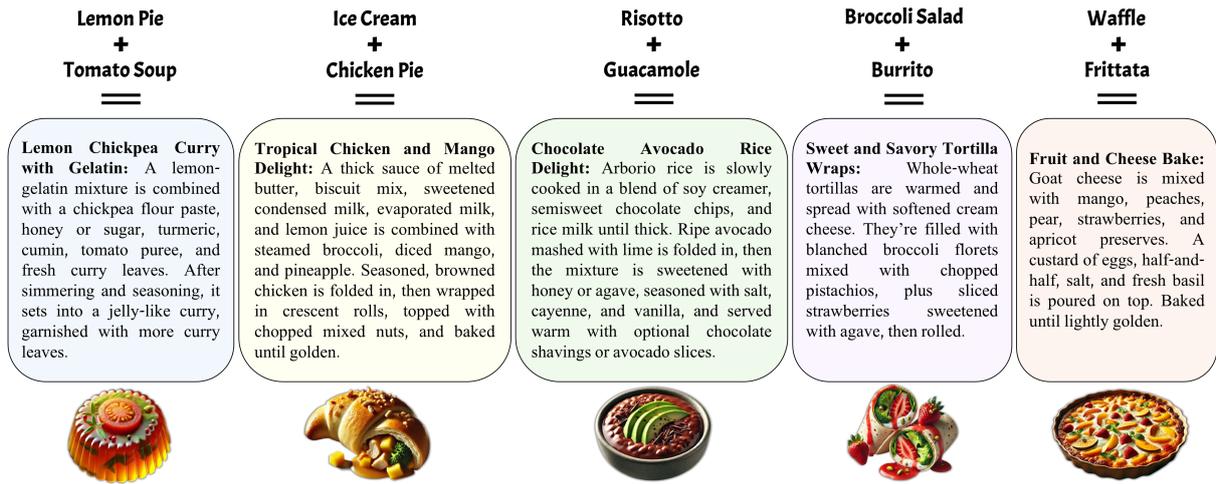


Figure 2: 由 *DishCOVER* 生成的新食谱创意示例。每个示例由一对输入菜肴及其最佳对应的生成食谱创意组成。生成的食谱以简明摘要的形式呈现，以节省空间。图像使用 OpenAI 的 DALL·E 生成。

表示整个概念空间，其中包含所有符合该领域结构和逻辑限制的想法，涵盖所有有效的可能性。此外，令 $I \subset \mathcal{I}$ 为在该领域内已记录或已知的一组想法。

Definition 1 (Recombination Function). 重组函数 C 以两个结构化的想法 $i_a, i_b \in I$ 作为输入，并生成一组新的组合 $I_{ab} \subseteq \mathcal{I}$ ，使得每个 $i \in I_{ab}$ 是 i_a, i_b 的不同混合。

“混合”的确切定义取决于表示。例如，当我们使用最小编辑距离方法从表示 i_a 过渡到表示 i_b 时，可以将中间步骤视为 i_a 和 i_b 的混合，在转换过程中以不同的比例混合两者的元素。

Definition 2 (Evaluation Function). 重组的结果是一组潜在的创新 I_{ab} ，可以用评估函数 $E: \mathcal{I} \rightarrow \mathbb{R}$ 进行评估。可以根据新颖性和实用性等标准对创新进行评估。

Definition 3 (Retrieval of Ideas from Descriptions). 思想经常以不同的抽象和细微程度来表达。令 m 为一个函数，它将想法描述 d 匹配到 I 、 $m(d) \subseteq I$ 中相关的已知想法。例如， m 可以将文本描述“lasagna”匹配到所有的烤宽面条食谱。

因此，正式的优化问题可以表述为：给定两个概念描述 d_a, d_b ，找出

$$\operatorname{argmax}_{i \in C(i_a, i_b) \mid i_a \in m(d_a), i_b \in m(d_b)} E(i)$$

图 2 展示了在烹饪食谱领域生成的创意示例，以及用于创造这些创意的描述。例如，将西兰花沙拉食谱和卷饼食谱结合，生成了一种以奶酪、西兰花、草莓和开心果填充的玉米饼食谱。

3 模型

在本节中，我们介绍了 *DishCOVER*，我们用于自动生成创新食谱的模型。¹ 图 1 展示了我们的方法论。输入由两个种子灵感（想法描述 d_a, d_b ）组成²。每个想法描述都被映射到一组特定的食谱（想法的实例，例如不同的千层面食谱；§ 3.1）。

这些配方首先在大型语言模型的帮助下被翻译为树状结构表示（图 1 的步骤 (I)，§ 3.2）。为了生成新的创意，我们重新组合这些树，并通过最小编辑距离算法产生新的候选创意（步骤 (II)，§ 3.3）。然后，利用新颖性 & 价值原则，我们优化候选创意以评估其价值，并根据其新颖性分数对其进行排名（步骤 (III)，§ 3.4）。最后，得分最高的树结构被翻译回自然语言配方，并使用大型语言模型进行完善（步骤 (IV)，§ 3.5）。接下来我们会提供关于每个步骤的更多细节。

3.1 采样种子想法（步骤 I）

我们从 *Recipe1M+* 数据集中选择了 100 道最受欢迎的菜肴（例如鸡肉沙拉、奶酪蛋糕），它们涵盖了不同的类别（例如开胃菜、甜点、主菜）。在该数据集中，每道菜的平均出现次数为 2,576.33 个食谱。

为了在我们模型的下一个阶段中保持使用 LLM 的财务成本在可控范围内，我们对每道菜抽取了 30 个食谱，总计得到 3000 个食谱样本。为了确保多样性和代表性，我们随机选择了 15 个食谱以捕捉每道菜的典型版本，并再选择 15 个以最大化多样性。多样化的实现可以依赖于表示方式和领域；在我们的案例中，我们使用 GMM 算法 (Ravi et al., 1994)，该算

²请注意，如果需要，可以使用更多的灵感。

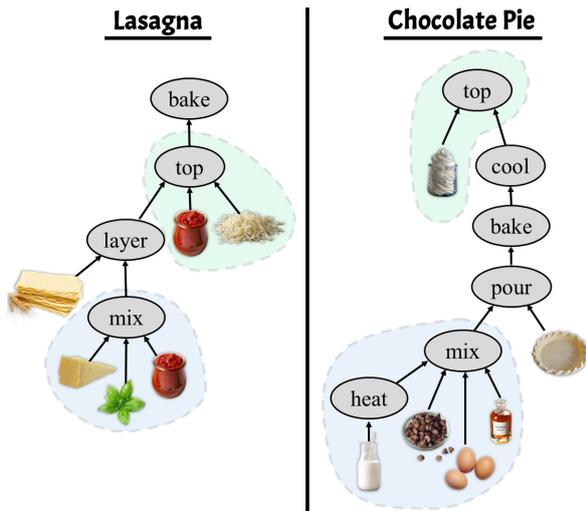


Figure 3: 树形表示法展示了千层面和巧克力派的食谱。类似的部分被突出显示，显示了在将一棵树转换为另一棵树时，最小编辑距离算法更可能保持的结构相似性。

法基于经过食谱数据微调的 Sentence-BERT 模型生成的食谱嵌入（详见附录 A）。

3.2 文本转树（步骤 II）

烹饪食谱像科学实验、装配手册和游戏说明一样，都是程序性文本。这些文本通常由一系列步骤构成，并附有执行这些步骤所需的对象。表示程序性文本的常见方法是使用树结构 (Jermsurawong and Habash, 2015; Maeta et al., 2015)，其中叶子节点对应所需的对象（我们的情况是配料），而内部节点表示对它们进行的动作。图 3 显示了以这种树格式表示的简单千层面和巧克力派的食谱。为了将食谱文本解析为树结构，我们提示了 GPT-4o，利用其广泛的世界知识和代码生成能力，加上连贯思维的方法。详细内容和相应的提示见附录 B & C。生成 3K 个食谱的树状表示的总成本大约是 \$ 40。初步检查显示，生成的树中有 1,347 个 (44.9%) 无效，原因包括孤立节点、单个节点有多条输出边或边方向不正确。为了解决这个问题，我们实施了一个纠正步骤，我们移除了有问题的边，并指示模型重新考虑它们。这将有效性提高到了 95% (2,850 棵树)。然后，我们使用 50 个随机食谱评估了最终的树。为每个食谱创建了一个黄金标准树，并在节点和边匹配方面自动比较了预测树和黄金树。对于节点，我们达到了 0.985 的准确率，0.956 的召回率，和 0.969 的 F1-分数。对于边，我们获得了 0.951 的准确率，0.909 的召回率，和 0.93 的 F1-分数。总体而言，这些结果展示了我们的方法在将程序性文本转化为结构化树表示方面的有效性，使它们适合进一步的操作和分析。

3.3 生成创意（步骤 III）

在本节中，我们使重组函数 C 变得可操作。我们通过结合最小编辑距离算法将食谱树进行混合，从而生成新颖的食谱。这种方法的关键思想是，通过研究两个概念之间的逐步转变，我们可以发现结合了两个概念特征的中间形式。

在配方生成的情况下，我们采用 Zhang-Shasha 算法，该算法计算树之间的最小编辑距离 (Zhang and Shasha, 1989; Bille, 2005)。给定两个配方树 i_a 和 i_b ，我们计算它们的最小编辑距离并记录所有需要的操作以将 i_a 转变为 i_b 。每个操作序列会产生代表新“合并”想法的中间树，我们从中随机选择一个作为我们的新配方。图 4 展示了一个示例序列，将简单的千层面树转化为简单的巧克力派树。一个中间变体可能是带罗勒的巧克力千层面；另一个可能是有外壳包裹的巧克力千层面。

最小编辑距离方法的一个关键优势在于其能够保留配料和烹饪步骤的结构角色。例如，在图 3 中，意大利千层面和巧克力派配方都包含一个“装饰”动作（用颜色标记）。使用最小编辑距离可以确保在番茄酱和奶酪旁边插入打发奶油更有可能，因为将打发奶油放在其他地方会增加整体编辑成本。实现细节请参见附录 D。

请注意，在转换过程中的不同点停止可以创造出独特的菜肴（见图 4）。此外，调整编辑顺序可以生成全新的中间想法。

3.4 评估想法（第四步）

现在我们通过重新配方来创建了一组候选创新，我们对每个生成的食谱进行评估。如在第 ?? 节中所述，评估函数 E 应该考虑新颖性和价值。具体而言，我们选择将价值视为一个约束条件，而将新颖性视为优化目标；也就是说，我们希望对所有通过价值门槛（即有意义）的候选者按新颖性进行排序。在烹饪领域，如果一个食谱在引入了意想不到的食材或技巧组合（新颖性）时仍能制作出美味且协调的菜肴（价值、实用性），则通常被认为是具有创意的。在下文中，我们将这些标准具体化。我们的实现专注于食谱，但我们相信这一原则可以推广到其他领域。

3.4.1 值约束

为了评估一个食谱的价值（味道），我们遵循 Varshney et al. (2019)，并检查其配料是否相互搭配良好。根据这项工作，如果两种配料共享较大比例的味道分子，则它们搭配良好；我们为每对原始配料计算一个搭配评分。如果这个评分低于这的阈值，我们认为这种搭配存在问题。我们反复删除低评分搭配数目最多

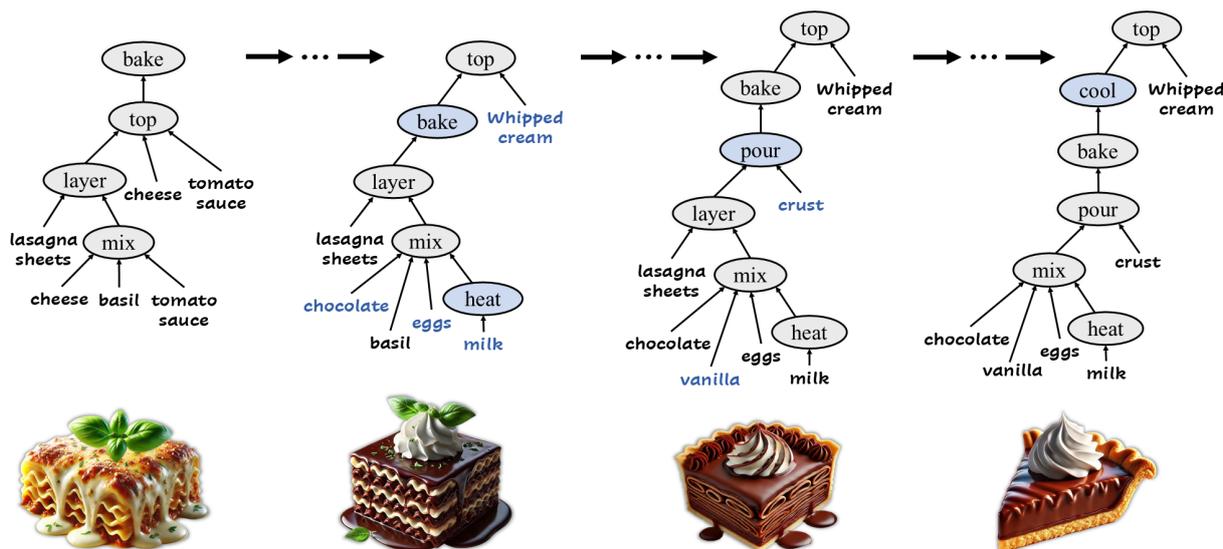


Figure 4: 我们基于树的编辑距离方法来生成新菜谱创意的示例，通过将简单的千层面树（左）转化为一个简单的巧克力派树（右）。在每个编辑步骤中都会生成中间的“合并”树，产生新颖的菜肴，如罗勒巧克力千层面或包裹在派皮中的巧克力千层面。菜谱图像是使用 OpenAI 的 DALL·E 模型生成的。

的配料，直到不再有进一步的冲突为止（或者，可以选择完全移除这些候选项）。详情见附录 E。

3.4.2 按照新颖性排序

我们的直觉是，新颖的食谱包含不常同时出现的成分和动作。因此，我们借鉴 tf-idf (Ramos et al., 2003) 中的逆文档频率 (idf) 概念，制定了一种惊讶度的度量。给定包含节点 E_T （成分和动作）的食谱树 T ，让 N_e 表示仓库中包含元素 e 的食谱数量。对于每个元素 $e' \in E_T \setminus e$ ， $df_e(e')$ 表示同时包含 e, e' 的食谱数量。然后我们定义³：

$$idf_e(e') = \log\left(\frac{N_e}{df_e(e')}\right)$$

较高的 $idf_e(e')$ 评分表明，相对于包含 e 的食谱， e' 更为独特。

为了计算 T 中一个元素的新颖性，我们计算所有 $idf_e(e')$ 分数并取前十相加。要计算 T 本身的新颖性，我们计算前十个元素的新颖性之和。例如，在图 2 中，鸡肉和芒果美食食谱由于像芒果、新月卷、坚果和西兰花这样的材料组合，以及例如蒸、焯水和展开这样的动作被评为新颖。

3.5 树到文本（第 V 步）

在识别出排名最高的树后，我们使用 LLM 将它们转换回自然语言食谱。与第 3.2 节类似，我

³为了防止由于拼写错误或极其罕见的成分导致分数膨胀，我们排除在整个数据集中出现极其不频繁的任何成分。

们采用了思维链方法，指示 GPT-4o 将树（以 DOT 格式编码）翻译成结构化的食谱，包括标题、食材列表和逐步指示。然后，我们提示模型对文本进行润色和修改，以确保连贯性、数量调整、成分一致性和整体可读性（完整提示见附录 F）。这种方法处理 1000 个食谱树的总成本约为 \$42。

4 创意食谱数据集

我们使用我们的模型生成了一个新的食谱数据集。我们首先识别了 100 道受欢迎的菜肴，涵盖了不同的类别。对于每道菜，我们采集了 30 份食谱并将它们转化为树结构。接着，我们采集了 1,000 对菜肴对，确保每对包含来自不同类别的菜肴。对于每对菜肴 d_a 和 d_b ， $m(d_a) \times m(d_b)$ 中的每对食谱被用于生成六个混合树，每对菜肴最多产生 5,600 棵树（~ 总计 5.5M）。我们从每组中选择五个最高排名的树，并将它们转化回自然语言食谱，最终生成一个包含 5K 食谱的数据集。

5 评估

我们现在转向评估 DishCOVER，通过研究以下研究问题：DishCOVER 生成的配方与 SOTA LLM (GPT-4o) 生成的配方相比如何？

为了回答这个问题，我们将考察两个关键方面。首先，输出在多样性方面的比较。具体来说，我们调查是否我们的方法缓解了在大语言模型 (LLMs) 中重复性的著名问题，这是创造性生成中的一个关键挑战。更重要的是，我们评估我们的输出在创造性方面如何与

GPT-4o 的输出相比。创造力要求输出既有价值（合乎逻辑）又新颖（出乎意料）。我们将我们的模型输出与 GPT-4o⁴ 在两个任务上的输出进行比较：

我们评估了 GPT-4o 和 DishCOVER 在生成结合特定菜肴对的创意食谱方面的能力。评估包括随机选择的 10 对菜肴。对于每一对菜肴，我们选择由 GPT-4o 生成的 5 个食谱，并将其与由 DishCOVER 生成的前 5 个食谱进行比较，从而每个模型总计生成 50 个食谱。

为了拓宽我们分析的范围，我们希望评估 GPT-4o 和 DishCOVER 可以普遍生成的最具创意的食谱，而不将它们限制在给定的输入对上。我们从每个模型中各使用了 100 个食谱。我们指示 GPT-4o 在一次对话会议中生成 100 个不同的创意食谱。对于 DishCOVER，我们从我们的 5K 食谱数据集中选择了 100 个食谱，该数据集为每个生成的食谱提供了一个新颖性评分。为了确保菜肴不是来自非常少量的输入，我们使用模拟退火 (Bertsimas and Tsitsiklis, 1993) 来最大化食谱的新颖性，同时对每道菜的最大出现次数施加约束。

请注意，实验 1 是一个更不常见的任务，模型在训练期间不太可能遇到这种任务。我们通过定性和自动化分析以及人工注释来评估这两个实验的输出。

5.1 实验细节

在这里，我们描述了选择 GPT-4o 提示参数的过程以及人工评估设置，这两者在两个实验中均一致实施。

Prompt and Temperature Selection. 大型语言模型 (LLM) 已知对提示语的改写很敏感 (Sclar et al., 2023; Mizrahi et al., 2024; Voronov et al., 2024)。为了优化我们的参数，我们首先通过从三个随机提示生成输出确定了温度值。在测试了各种设置后，我们选择了 $t = 1$ 作为生成连贯配方的最高温度，因为更高的值会导致输出混乱。在温度固定的情况下，并遵循 Mizrahi et al. (2024) 的方法，一名团队成员每次实验策划了 100 个提示，每个提示生成三个配方。然后对这 300 个生成的配方进行审查，并标记较为独特的输出。最终，选择了一个最终提示，因为它能够在所有输出中一致地生成具创意的配方（见附录 G 中的两个实验的选择提示）。

我们使用 Prolific 来招募和管理两个实验的共 48 名参与者。我们根据烹饪频率、调整食谱的自如程度以及判断创意结果的能力对参与者进行了初步筛选。参与者按道德补偿标准获

⁴版本: gpt-4o-2024-08-06 (撰写本论文时该模型的最新稳定版本)。

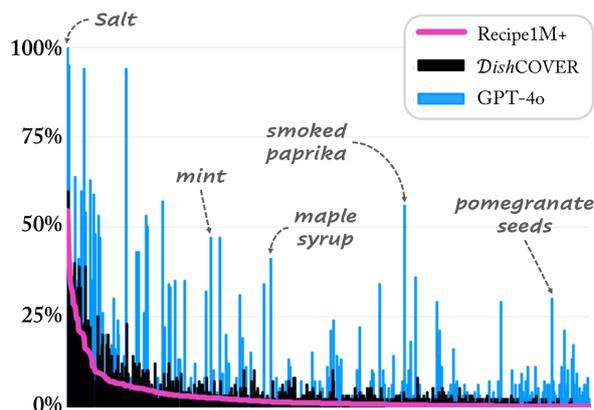


Figure 5: 表示在原始食谱库中的成分频率（用粉色表示）和模型生成的食谱中的成分频率。黑色直方图显示 DishCOVER 紧随库的分布。GPT-4o（蓝色）显示出尖峰，突出了对某些成分的偏向。

得每小时估价 9 美元的报酬。要求参与者在多种价值和新颖性方面对从两个模型中随机抽取的食谱进行评分（参见附录 H 中的问题列表）。为了减少认知负荷，他们查看了简短的食谱摘要（也使用 GPT-4o 生成），在需要时可以查看完整食谱。为减少评分噪声，每个食谱由五个注释者评分，最终评分采用中位数排名。新颖性和价值分数是其相应问题的平均值，价值分数还通过对三个相关价值问题设定阈值 4 进行二值化。平均而言，每位参与者对 31.25 个食谱进行了评分（标准差 = 33.827）。为确保质量，我们排除了平均在 45 秒内完成阅读食谱并回答其八个问题的参与者。

5.2 生成菜谱的多样性

多样性在创意生成中起着至关重要的作用，它反映了新食谱如何广泛和灵活地适应和转换原始概念。在我们评估的这一部分，我们使用定性和自动分析来比较 DishCOVER 的输出与 GPT-4o 生成的输出，研究每个模型如何整合来自不同菜肴的概念及其重复性程度。

食谱概念的深层与浅层合并。 中，DishCOVER 和 GPT-4o 都展示了将不同菜肴的创意合并的能力，但我们注意到它们的输出展现了不同类型的整合。DishCOVER 一贯产生更具凝聚力的整合，两个来源菜肴的食材被编织成一个统一的食谱。相比之下，GPT-4o 倾向于“浅层合并”，通常先分别准备每道菜，然后再在最后结合。例如，当被要求创造一个松饼和橙色沙拉的混合品时，GPT-4o 总是提出先烤松饼，准备一份橙色沙拉，然后将它们一起上桌，把沙拉放在顶部或旁边。相比之下，DishCOVER 生成食谱，如混合蔬菜烤，结合两道菜的元素，或含有橙酱的柑橘蛋

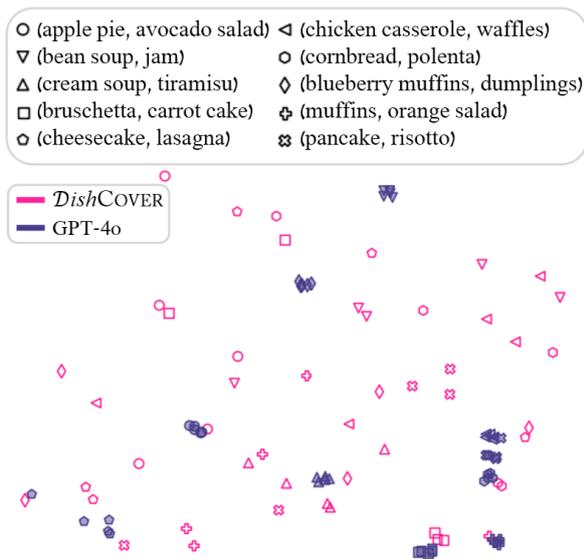


Figure 6: t-SNE 可视化用于菜谱嵌入实验 1。形状表示菜肴对，颜色表示模型。蓝色簇 (GPT-4o) 的相同形状显示紧密和定位集中，而粉红色 (DishCOVER) 则分散，显示出 DishCOVER 的输出具有更高的多样性。

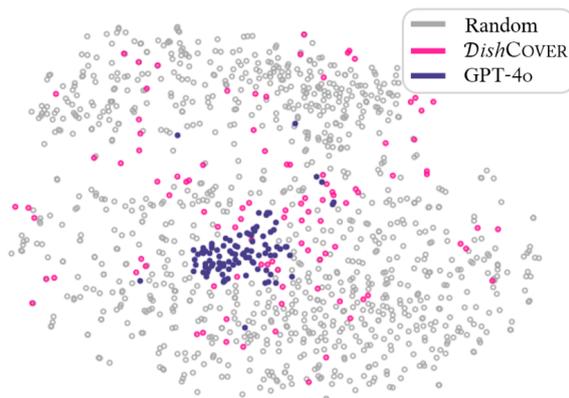


Figure 7: 对食谱嵌入的 t-SNE 可视化，实验 2。灰色点代表来自通用数据库的 1K 随机食谱。同样，DishCOVER 的输出更加多样化。

糕，以及整合了松饼食材的沙拉。

这种差异也反映在每个食谱的平均成分数量中：GPT-4o 的食谱中成分的数量几乎是 DishCOVER 的两倍：在第一次实验中，DishCOVER 平均使用了 12.3 种成分 (标准差 = 3.71)，而 GPT-4o 为 24.98 种 (标准差 = 5.2)；在第二次实验中，DishCOVER 平均使用了 13.32 种成分 (标准差 = 4.39)，而 GPT-4o 为 23.79 种 (标准差 = 3.25)。作为参考，食谱库中的平均成分数量较低，为 9.33 (标准差 = 4.31)。这种不平衡表明 GPT-4o 倾向于生成独立的子菜，每个子菜都有其自身的成分，然后将它们组合成一道最终菜肴，而不是将烹饪理念真正融合在一起。

对 GPT-4o 输出结果中的结构和成分的关注。

GPT-4o 表现出对某些结构和成分的强烈固执。当被要求合并两道菜时，它在多次尝试中往往使用相同的方法。例如，当被要求融合“扁豆汤”和“果酱”时，GPT-4o 反复生成变体，即在单独准备的果酱旁边提供扁豆汤，通常以相同的方式装盘 (例如，将一团果酱放在汤的中央)。这种固执在被要求合并每道菜的特定给定食谱时仍然存在，甚至在被要求合并更广泛的菜肴类型 (例如，一般汤和一般蘸酱) 时也是如此。

类似地，GPT-4o 反复使用相同的 (不常见的) 配料。例如，其 56% 的食谱包含烟熏辣椒粉 (而在库中仅为 0.375%)，47% 使用了薄荷，41% 使用了枫糖浆，29% 使用了石榴糖蜜。图 5 显示了 GPT-4o 和 DishCOVER 食谱中配料的频率与库基准的对比。虽然 DishCOVER 的分布与原始数据非常接近，但 GPT-4o 对某些低频配料有很强的偏向。

通过树距离量化多样性。

我们在将生成的食谱转换为层次树结构后，计算了它们之间的平均归一化树编辑距离 (Rico-Juan and Micó, 2003)。在两项实验中，GPT-4o 的输出比 DishCOVER 的树距离明显更低，表明相似度更高。在第一个实验中，我们分别计算了每对菜肴的平均编辑距离，发现 DishCOVER 的平均树距离为 132.14，而 GPT-4o 的为 89.35 (p-value = 3.6e-05, 配对 t 检验)。在第二个实验中，在所有输出中，DishCOVER 再次表现出更大的多样性，平均树距离为 140.25，而 GPT-4o 的为 129.55 (p-value < 1e-50, 双样本 t 检验)。

我们还使用一个专门为烹饪食谱微调的 Sentence-BERT 模型分析了生成食谱的嵌入 (见附录 A)。图 6 展示了第一次实验的 t-SNE 可视化，其中每种形状代表一个菜肴对，颜色表示模型 (DishCOVER 为粉色，GPT-4o 为蓝色)。GPT-4o 的食谱在每对中形成了紧密的簇，而 DishCOVER 的食谱则更加分散。在实验 2 中也出现了类似的模式 (图 7)。构建后，DishCOVER 的嵌入分布广泛。然而，有趣的是，GPT-4o 的嵌入倾向于紧密聚集。图 8 通过所有食谱嵌入的余弦相似性热图强化了这些发现。在第一次实验中，DishCOVER 的输出平均相似性为 0.387 (标准差 = 0.120)，而 GPT-4o 的更高，为 0.659 (标准差 = 0.121)。这种模式在第二次实验中得以保持 (DishCOVER : 0.402, 标准差 = 0.110; GPT-4o : 0.731, 标准差 = 0.078)。

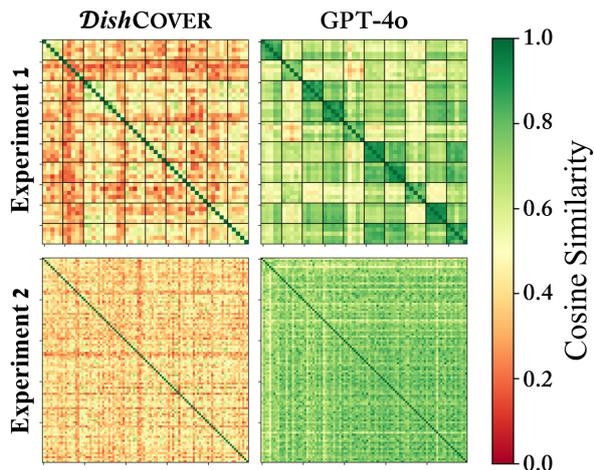


Figure 8: 四个热图展示了每个模型在两个实验中食谱嵌入的成对余弦相似性。GPT-4o 的热图（右）显示了更多的高相似性区域，表明 DishCOVER 的多样性更高。

5.3 人工注释

虽然多样性至关重要，食谱生成的真正创造力也取决于价值和新颖性。为了评估这些方面，我们基于烹饪专家在人工评估中的注释进行了研究。

就价值而言，这两种模型产生的输出大部分被认为是具有价值的。在第一个实验中，80% 的 DishCOVER 的配方被归类为有价值，其平均价值评分为 4.320，比较而言，GPT-4o 为 82% (4.326)。在第二个实验中，85% 的 DishCOVER 的输出被归类为有价值 (4.36)，而 GPT-4o 为 98% (4.51)。这很合理，因为我们注意到第二个实验（开放式提示）更类似于 GPT-4o 在训练过程中遇到的数据，并且它也倾向于输出相对相似的配方。

为了新颖性，我们只考虑被评为有价值的食谱；这反映了厨师浏览推荐食谱清单的一个用例：快速判断其合理性，仅深入研究那些有潜力的。在两个实验中，DishCOVER 在这方面显著优于 GPT-4o。在第一个实验中，DishCOVER 的平均新颖性得分为 3.53，而 GPT-4o 的为 3.146 ($p\text{-value}=0.0009$)。在第二个实验中，DishCOVER 的平均新颖性得分为 3.612，而 GPT-4o 的为 3.141 ($p\text{-value}=9.2E-08$)。

进一步检验第二个实验的有价值结果，DishCOVER 在新颖性评分的顶四分位中占主导地位，占最高评分食谱的 75.55%，而 GPT-4o 在较低的两个四分位中占优势，占较少新颖性结果的 71.74%。此外，在 37 个新颖性评分为 4 或更高的有价值食谱中，DishCOVER 贡献了 32 个，只有五个来自 GPT-4o。图 2 展示了该高新颖性集合中的五个 DishCOVER 食谱。这些结果强烈表明，虽然两个模型大多生成有

价值的食谱（GPT-4o 在开放式情况下更是如此），但 DishCOVER 在生成真正创造性的食谱方面具有明显优势。

6 相关工作

我们的方法基于最近的解析方法，这些方法指导 LLMs 将自然语言映射到结构化形式，提高了 LLM 在诸如短语结构解析 (Tian et al., 2024a) 和信息抽取 (Zhao et al., 2023; Li et al., 2024) 等任务中的表现。类似于将 LLMs 与知识图谱 (KGs) 集成以改善推理和论证 (Wang et al., 2024b; Feng et al., 2023; Jiang et al., 2023; Sun et al., 2023) 的研究，我们将领域特定的知识纳入结构化表示中，以提供给 LLMs 更清晰、富含上下文的信号。我们的工作也与将文本解析为结构化知识、操作这些表示并（可选地）将其转化为自然语言以增强 LLM 能力 (Yang et al., 2023; Besta et al., 2024; Zelikman et al., 2023; Zhang et al., 2025) 的模型一致。我们惊讶地发现类似的技术可以提高创造力和多样性。

我们专注于提高 LLM 输出多样性的工作，补充了通过人类反馈 (Chung et al., 2023)、上下文学习 (Zhang et al., 2024) 或基于知识图谱的干预 (Liu et al., 2021; Hwang et al., 2023; Liu et al., 2022) 来促进生成变化的努力。此外，我们的工作与旨在培养 LLM 创造性的更广泛研究方向一致。一项相关的研究利用幻觉产生新想法 (Jiang et al., 2024; Yuan and Färber, 2025)。另一项则汲取人类创造力研究的见解，结合诸如约束 (Lu et al., 2024)、联想思维 (Mehrotra et al., 2024)、角色扮演 (Chen et al., 2024) 和头脑风暴 (Summers-Stay et al., 2023; Chang and Li, 2025; Rana and Cheok, 2025) 等技术。同样地，我们专注于重组。

除了改进 LLMs 之外，研究人员还探索了将它们用作作家 (Mirowski et al., 2023; Yuan et al., 2022; Chakrabarty et al., 2023; Wan et al., 2024)、视觉艺术家 (Ko et al., 2023) 甚至幽默作家的创意辅助工具。然而，尽管这些工具可以提升用户的创造力感受，它们的多样性有限可能会使不同个人产生的想法趋于同质化 (Anderson et al., 2024)。

最近的研究已经探索了在烹饪领域中使用大规模语言模型，包括在菜谱生成方面的一些早期尝试。在大规模语言模型出现之前，计算菜谱生成主要关注提出新的食材组合，常常忽视完整的烹饪指令生成。

7 讨论与结论

我们引入了一种新颖的范式，通过在创意层面而非词元层面的结构化重组来增强大型语言模型中的创意生成。

Implications for AI Creativity and Structured Knowledge. 我们工作的一个核心见解是，结构化表示提供了一种在更高抽象层次上引入有意义变化的机制。增强 LLM 创造力的尝试通常集中在增加标记的随机性（例如，通过温度），这不足以产生创造力 (Peepkorn et al., 2024)。我们专注于重组，这是创造力研究中一个成熟的原则。通过在计算上实现这一原则，我们证明结合结构化知识使模型能够更有效地在思想空间中导航，从而产生新颖且有价值的输出。

我们的方法的另一个有前景的应用是在结构化采样和基于搜索的生成任务中。LLMs 中的采样方法通常依赖于在令牌层面上操作的随机技术（例如，核采样或最优 k 采样）来引入变异性。相比之下，我们通过抽象表示上进行采样而不是直接在原始令牌上进行采样来实现多样化。我们引入了一种控制但灵活地导航创意思维空间的方法。这种采样方法可以增强需要平衡新颖性和连贯性的多种应用。

除了烹饪，我们的框架可以推广到任何领域，在这些领域中，结构化知识可以指导创造性的重组，例如产品设计、科学发现或叙事建构。树结构的表示自然适用于许多这些领域，我们的方法也可以扩展到其他结构，如序列、通用图或其他形式结构。除了想法的生成，我们的框架还可以支持合成数据的模拟和扩充，解决数据稀缺的问题——这是训练低资源领域的 AI 模型的持续瓶颈。在这种情况下，新颖性评分确保了多样化的输出，而领域特定的约束有助于保持数据的有效性。

最终，我们的方法代表了向不仅生成文本而是真正参与创造性综合的 AI 系统迈出的重要一步。我们希望激发在结构化创造力和 AI 方面的进一步研究，强调结构化重组在增强人类创造力和扩展 AI 辅助创新的前沿上的潜力。

References

- Gautam Ahuja and Curba Morris Lampert. 2001. Entrepreneurship in the large corporation: A longitudinal study of how established firms create breakthrough inventions. *Strategic management journal*, 22(6-7):521–543.
- Barrett R Anderson, Jash Hemant Shah, and Max Kreminski. 2024. Homogenization effects of large language models on human creative ideation. In *Proceedings of the 16th conference on creativity & cognition*, pages 413–425.
- S Arora, C Lund, Ro MOTWANI, M Sudan, and M Szegedy. 1992. Proof verification and hardness of approximation problems. In *Proc. 33rd IEEE Symposium on Foundations of Computer Science, Pittsburgh, PA*, pages 14–23.
- Dimitris Bertsimas and John Tsitsiklis. 1993. Simulated annealing. *Statistical science*, 8(1):10–15.
- Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, and 1 others. 2024. Graph of thoughts: Solving elaborate problems with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17682–17690.
- Philip Bille. 2005. A survey on tree edit distance and related problems. *Theoretical computer science*, 337(1-3):217–239.
- Margaret A Boden. 2004. *The creative mind: Myths and mechanisms*. Routledge.
- Margaret A Boden. 2009. Computer models of creativity. *Ai Magazine*, 30(3):23–23.
- Tuhin Chakrabarty, Philippe Laban, Divyansh Agarwal, Smaranda Muresan, and Chien-Sheng Wu. 2024. Art or artifice? large language models and the false promise of creativity. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–34.
- Tuhin Chakrabarty, Vishakh Padmakumar, Faeze Brahman, and Smaranda Muresan. 2023. Creativity support in the age of large language models: An empirical study involving emerging writers. *arXiv preprint arXiv:2309.12570*.
- Hung-Fu Chang and Tong Li. 2025. A framework for collaborating a large language model tool in brainstorming for triggering creative thoughts. *Thinking Skills and Creativity*, page 101755.
- Jing Chen, Xinyu Zhu, Cheng Yang, Chufan Shi, Yadong Xi, Yuxiang Zhang, Junjie Wang, Jiashu Pu, Rongsheng Zhang, Yujiu Yang, and 1 others. 2024. Hollmwood: Unleashing the creativity of large language models in screenwriting via role playing. *arXiv preprint arXiv:2406.11683*.
- John Joon Young Chung, Ece Kamar, and Saleema Amershi. 2023. Increasing diversity while maintaining accuracy: Text data generation with large language models and human interventions. *arXiv preprint arXiv:2306.04140*.
- Simona Doboli, Jared Kenworthy, Paul Paulus, Ali Minai, and Alex Doboli. 2020. A cognitive inspired method for assessing novelty of short-text ideas. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.
- Chao Feng, Xinyu Zhang, and Zichu Fei. 2023. Knowledge solver: Teaching llms to search for domain knowledge from knowledge graphs. *arXiv preprint arXiv:2309.03118*.
- Ronald A Finke, Thomas B Ward, and Steven M Smith. 1996. *Creative cognition: Theory, research, and applications*. MIT press.
- Giorgio Franceschelli and Mirco Musolesi. 2024. On the creativity of large language models. *AI & SOCIETY*, pages 1–11.
- Naomi K Fukagawa, Kyle McKillop, Pamela R Pehrson, Alanna Moshfegh, James Harnly, and John Finley. 2022. Usda’s fooddata central: what is it and why is it needed today? *The American journal of clinical nutrition*, 115(3):619–624.
- Neelansh Garg, Apuroop Sethupathy, Rudraksh Tuwani, Rakhi Nk, Shubham Dokania, Arvind Iyer, Ayushi Gupta, Shubhra Agrawal, Navjot Singh, Shubham Shukla, and 1 others. 2018. Flavordb: a database of flavor molecules. *Nucleic acids research*, 46(D1):D1210–D1216.
- Joy Paul Guilford. 1967. The nature of human intelligence.
- David JP Heinen and Dan R Johnson. 2018. Semantic distance: An automated measure of creativity that is novel and appropriate. *Psychology of Aesthetics, Creativity, and the Arts*, 12(2):144.
- EunJeong Hwang, Veronika Thost, Vered Shwartz, and Tengfei Ma. 2023. Knowledge graph compression enhances diverse commonsense generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 558–572.
- Jermsak Jermsurawong and Nizar Habash. 2015. Predicting the structure of cooking recipes. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 781–786.
- Jinhao Jiang, Kun Zhou, Zican Dong, Keming Ye, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Structgpt: A general framework for large language model to reason over structured data. *arXiv preprint arXiv:2305.09645*.

- Xuhui Jiang, Yuxing Tian, Fengrui Hua, Chengjin Xu, Yuanzhuo Wang, and Jian Guo. 2024. A survey on large language model hallucination via a creativity perspective. *arXiv preprint arXiv:2402.06647*.
- Anna Katerina Jordanous. 2012. *Evaluating computational creativity: a standardised procedure for evaluating creative systems and its application*. University of Kent (United Kingdom).
- Yoed N Kenett. 2019. What can quantitative measures of semantic distance tell us about creativity? *Current Opinion in Behavioral Sciences*, 27:11–16.
- Hyung-Kwon Ko, Gwanmo Park, Hyeon Jeon, Jaemin Jo, Juho Kim, and Jinwook Seo. 2023. Large-scale text-to-image generation models for visual artists’ creative works. In *Proceedings of the 28th international conference on intelligent user interfaces*, pages 919–933.
- Arthur Koestler. 1964. The act of creation.
- Carolyn Lamb, Daniel G Brown, and Charles LA Clarke. 2018. Evaluating computational creativity: An interdisciplinary tutorial. *ACM Computing Surveys (CSUR)*, 51(2):1–34.
- Yinghao Li, Rampi Ramprasad, and Chao Zhang. 2024. A simple but effective approach to improve structured language model output for information extraction. *arXiv preprint arXiv:2402.13364*.
- Chenzhengyi Liu, Jie Huang, Kerui Zhu, and Kevin Chen-Chuan Chang. 2022. Dimongen: Diversified generative commonsense reasoning for explaining concept relationships. *arXiv preprint arXiv:2212.10545*.
- Ye Liu, Yao Wan, Lifang He, Hao Peng, and S Yu Philip. 2021. Kg-bart: Knowledge graph-augmented bart for generative commonsense reasoning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 6418–6425.
- Yining Lu, Dixuan Wang, Tianjian Li, Dongwei Jiang, and Daniel Khashabi. 2024. Benchmarking language model creativity: A case study on code generation. *arXiv preprint arXiv:2407.09007*.
- Hirokuni Maeta, Tetsuro Sasada, and Shinsuke Mori. 2015. A framework for procedural text understanding. In *Proceedings of the 14th International Conference on Parsing Technologies*, pages 50–60.
- Pronita Mehrotra, Aishni Parab, and Sumit Gulwani. 2024. Enhancing creativity in large language models through associative thinking strategies. *arXiv preprint arXiv:2405.06715*.
- Piotr Mirowski, Kory W Mathewson, Jaylen Pittman, and Richard Evans. 2023. Co-writing screenplays and theatre scripts with language models: Evaluation by industry professionals. In *Proceedings of the 2023 CHI conference on human factors in computing systems*, pages 1–34.
- Moran Mizrahi, Guy Kaplan, Dan Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky. 2024. State of what art? a call for multi-prompt llm evaluation. *Transactions of the Association for Computational Linguistics*, 12:933–949.
- Moran Mizrahi and Dafna Shahaf. 2021. 50 ways to bake a cookie: Mapping the landscape of procedural texts. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 1304–1314.
- Michael D Mumford. 2003. Where have we been, where are we going? taking stock in creativity research. *Creativity research journal*, 15(2-3):107–120.
- Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. 2024. Unifying large language models and knowledge graphs: A roadmap. *IEEE Transactions on Knowledge and Data Engineering*, 36(7):3580–3599.
- Max Peeperkorn, Tom Kouwenhoven, Dan Brown, and Anna Jordanous. 2024. Is temperature the creativity parameter of large language models? *arXiv preprint arXiv:2405.00492*.
- Juan Ramos and 1 others. 2003. Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*, volume 242, pages 29–48. Citeseer.
- Sharif Uddin Ahmed Rana and Adrian David Cheok. 2025. Generative innovation: Leveraging the power of large language models for brainstorming. In *The Economics of Talent Management and Human Capital*, pages 175–192. IGI Global.
- Sekharipuram S Ravi, Daniel J Rosenkrantz, and Giri Kumar Tayi. 1994. Heuristic and special case algorithms for dispersion problems. *Operations Research*, 42(2):299–310.
- Juan Ramón Rico-Juan and Luisa Micó. 2003. Comparison of aesa and laesa search algorithms using string and tree-edit-distances. *Pattern Recognition Letters*, 24(9-10):1417–1426.
- Graeme Ritchie. 2007. Some empirical criteria for attributing creativity to a computer program. *Minds and Machines*, 17:67–99.
- Sameh Said-Metwaly, Wim Van den Noortgate, and Eva Kyndt. 2017. Approaches to measuring creativity: A systematic literature review. *Creativity: theories-research-applications.-Warsaw, Poland, 2014, currens*, 4(2):238–275.
- R Keith Sawyer and Danah Henriksen. 2024. *Explaining creativity: The science of human innovation*. Oxford university press.
- Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2023. Quantifying language models’ sensitivity to spurious features in prompt design or: How

- i learned to start worrying about prompt formatting. In *The Twelfth International Conference on Learning Representations*.
- Douglas Summers-Stay, Clare R Voss, and Stephanie M Lukin. 2023. Brainstorm, then select: a generative language model improves its creativity score. In *The AAAI-23 Workshop on Creative AI Across Modalities*.
- Jiashuo Sun, Chengjin Xu, Lumingyuan Tang, Saizhuo Wang, Chen Lin, Yeyun Gong, Lionel M Ni, Heung-Yeung Shum, and Jian Guo. 2023. Think-on-graph: Deep and responsible reasoning of large language model on knowledge graph. *arXiv preprint arXiv:2307.07697*.
- Yuanhe Tian, Fei Xia, and Yan Song. 2024a. Large language models are no longer shallow parsers. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7131–7142.
- Yufei Tian, Tenghao Huang, Miri Liu, Derek Jiang, Alexander Spangher, Muhao Chen, Jonathan May, and Nanyun Peng. 2024b. Are large language models capable of generating human-level narratives? *arXiv preprint arXiv:2407.13248*.
- James M Utterback. 1996. *Mastering the dynamics of innovation*. Harvard Business School Press.
- Lav R Varshney, Florian Pinel, Kush R Varshney, Debarun Bhattacharjya, Angela Schörgendorfer, and Y-M Chee. 2019. A big data approach to computational creativity: The curious case of chef watson. *IBM Journal of Research and Development*, 63(1):7–1.
- Anton Voronov, Lena Wolf, and Max Ryabinin. 2024. Mind your format: Towards consistent evaluation of in-context learning improvements. *arXiv preprint arXiv:2401.06766*.
- Qian Wan, Siying Hu, Yu Zhang, Piaohong Wang, Bo Wen, and Zhicong Lu. 2024. "it felt like having a second mind": Investigating human-ai co-creativity in prewriting with large language models. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW1):1–26.
- Dawei Wang, Difang Huang, Haipeng Shen, and Brian Uzzi. 2024a. A preliminary, large-scale evaluation of the collaborative potential of human and machine creativity. *arXiv preprint*.
- Yu Wang, Nedim Lipka, Ryan A Rossi, Alexa Siu, Ruiyi Zhang, and Tyler Derr. 2024b. Knowledge graph prompting for multi-document question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19206–19214.
- Zhun Yang, Adam Ishay, and Joohyung Lee. 2023. Coupling large language models with logic programming for robust and general reasoning from text. *arXiv preprint arXiv:2307.07696*.
- Ann Yuan, Andy Coenen, Emily Reif, and Daphne Ippolito. 2022. Wordcraft: story writing with large language models. In *Proceedings of the 27th International Conference on Intelligent User Interfaces*, pages 841–852.
- Shuzhou Yuan and Michael Färber. 2025. Hallucinations can improve large language models in drug discovery. *arXiv preprint arXiv:2501.13824*.
- Eric Zelikman, Qian Huang, Gabriel Poesia, Noah Goodman, and Nick Haber. 2023. Parsel: Algorithmic reasoning with language models by composing decompositions. *Advances in Neural Information Processing Systems*, 36:31466–31523.
- Jiahuan Zhang, Tianheng Wang, Hanqing Wu, Ziyi Huang, Yulong Wu, Dongbai Chen, Linfeng Song, Yue Zhang, Guozheng Rao, and Kaicheng Yu. 2025. Sr-llm: Rethinking the structured representation in large language model. *arXiv preprint arXiv:2502.14352*.
- Kaizhong Zhang and Dennis Shasha. 1989. Simple fast algorithms for the editing distance between trees and related problems. *SIAM journal on computing*, 18(6):1245–1262.
- Tianhui Zhang, Bei Peng, and Danushka Bollegala. 2024. Improving diversity of commonsense generation by large language models via in-context learning. *arXiv preprint arXiv:2404.16807*.
- Bowen Zhao, Changkai Ji, Yuejie Zhang, Wen He, Yingwen Wang, Qing Wang, Rui Feng, and Xiaobo Zhang. 2023. Large language models are complex table parsers. *arXiv preprint arXiv:2312.11521*.
- Yunpu Zhao, Rui Zhang, Wenyi Li, Di Huang, Jiaming Guo, Shaohui Peng, Yifan Hao, Yuanbo Wen, Xing Hu, Zidong Du, and 1 others. 2024. Assessing and understanding creativity in large language models. *arXiv preprint arXiv:2401.12491*.

A 微调后的 Sentence-BERT 模型

我们最初使用标准句子级的 Sentence-BERT (SBERT) 模型进行的实验表明, 该模型倾向于将重视文本指令的食谱分组, 而忽略了食材。因此, 它无法区分广泛的类别 (例如, 沙拉与汤), 并且在细粒度的区分上也存在困难 (例如, 胡萝卜蛋糕与芝士蛋糕)。

为了更好地处理食谱相似性, 我们对 Sentence-BERT 模型⁵进行了微调。我们的微调数据集由 3 万个食谱对其新的相似度评分组成, 这些评分平等地权衡了原始 Sentence-BERT 评分和通过 Mizrahi and Shahaf (2021) 计算的两个食谱配料表的基于 Ruzicka 的相似性。数据集被分为三个相等的子集: (1) 1 万个食谱对, 每对代表同一道菜的实例, (2) 1 万个来自同一类别 (例如, 胡萝卜蛋糕和奶酪蛋糕) 但不同菜肴的食谱对, (3) 1 万个来自完全不同类别 (例如, 甜点和沙拉) 的食谱对。

为了保持原始模型的能力并避免过拟合, 我们对其进行了一个 epoch 的微调, 保留了 5% 的对用于验证, 实现了 92-95% 的验证准确率。遵循 Sentence-BERT 微调指南, 我们采用了 10% 个预热步骤, 使用余弦相似损失作为损失函数, 并将最大序列长度设置为 512 个标记。

这种微调过程使我们能够获得更准确的食谱嵌入, 既考虑了文本指令也考虑了配料重叠。图 9 显示了 40 个食谱在微调前后的余弦相似度热图: 10 个与胡萝卜蛋糕高度相似的食谱, 10 个甜点松露的食谱, 10 个饺子的食谱, 以及 10 个披萨 (主菜) 食谱。如图所示, 原始的 Sentence-BERT 模型在所有菜肴对中表现出较高的相似性分数, 包括胡萝卜蛋糕和披萨食谱之间意外的高相似性。相反, 经过微调的模型在同类菜肴间保持高相似性得分, 同时在同类别菜肴间和跨类别之间提高了区分度。

B 文本到树解析器的详细信息

在第一个子任务中, 我们指示模型解析配料。对于每一行配料, 模型提取了 (1) 配料名称, (2) 配料是对菜肴的结构核心有贡献 (例如千层面的千层面片) 还是对味道有贡献 (例如柠檬派中的柠檬), 以及 (3) 配料的简化基本形式 (例如“罗勒”“香草”, “核桃”“坚果”)。为了评估准确性, 我们随机抽取了 200 行配料, 然后由我们中的一人检查解析的信息是否与真实情况匹配。我们获得了配料名称解析的 95%% 准确率, 结构与味道参考的 97.5%% 准确率, 以及配料基本形式转换的 96%% 准确率。这些结果表明, GPT-4o 在解析配料方面表现可靠。

⁵all-distilroberta-v1。

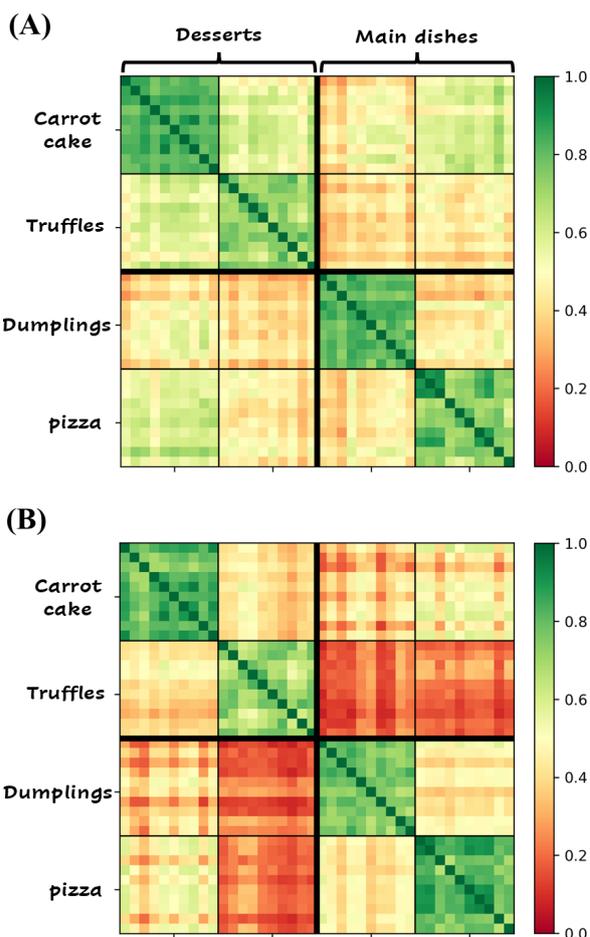


Figure 9: 40 个食谱在微调前后的余弦相似度热图: (A) 微调前的 Sentence-BERT, (B) 微调后的 Sentence-BERT。原始的 Sentence-BERT 模型在所有菜肴对中显示出一致的高相似性, 而经过微调的模型在同类菜肴中保持高相似性, 并改善了类别间的区分。

简化说明。 接下来, 我们要求模型在不丢失重要内容的情况下简化食谱说明中的每一个句子。具体来说, 我们指示它去掉关于数量、尺寸和描述性元素的细节, 确保每个简化后的句子正好包含一个动作 (放在开头), 并将模糊的指令转化为主动形式 (例如, “bring to a boil” “boil”)。例如, 将指令 “Sprinkle salt over the basil and mozzarella and return to the broiler for 1 to 2 minutes, until the cheese is melted and bubbling.” 简化为 “Sprinkle salt over basil, mozzarella. Broil until cheese melts”。我们在 50 个采样的食谱上测试了这个阶段, 产生了 451 个指令句子。其中一名研究者审查了这些句子, 发现 423 个 (93.79%) 被正确简化。常见的错误包括使用信息不足的动词, 遗漏动词或成分, 或意外地把两个动作合并成一个句子。

最后, 我们利用 GPT-4o 的编码能力生成了 DOT 中的有向树表示, DOT 是一种图形描述

语言。我们为模型提供了一个单例示例，其中包含简化的食谱文本（菜名、一小部分配料清单和简化的说明），以及相应的 DOT 代码，并附有注释，以指导模型构建树形表示。为了优化输出，我们实施了一个校正步骤，自动去除有问题的边，并指导模型重新考虑这些边。

C 文本到树解析器提示

本节包括我们用来提示 GPT-4o 将食谱文本解析为树形表示的提示。所有在此展示的提示的系统消息是：“你是一个烹饪食谱解析器”。

解析原料提示： 给定一个食谱标题、编号和成分，对于每种成分，确定：(1) 缩写：最简明的描述。(2) 参考类型：确定该成分是用于菜肴的结构 (“structure”) 还是口味 (“taste”)。同时影响两者的成分标记为 “taste”。(3) 核心成分：布尔值，表明该成分是否对菜肴的身份至关重要（例如，巧克力蛋糕中的巧克力为 True）。(4) 抽象：将成分简化为其基本形式（例如，将 “basil” 简化为 “herb”，“walnuts” 简化为 “nut”，“eggs” 简化为 “egg”)。请仅以以下 JSON 格式返回结果：{ “recipe_id”: [(缩写, 参考类型, 核心成分, 抽象), ...], ... }。输入：recipe_title, recipe_id, ingr_list, recipe_title, recipe_id, ingr_list, ... 输出：

简化指令提示： 针对以下烹饪说明，请尽可能简化和缩短。去除数量、大小和描述。确保每个动词开启一个新句子，并且一个句子不包含两个动词。将允许性或含糊的指令转换为主动形式（例如，“let cool” 转换为 “cool”，“alternate layers” 转换为 “layer”)。以 JSON 格式返回输出，键为 “recipe_id”，值为简化后的完整文本。输入：{ recipe_id: <指令文本>, ... } 输出：

将配方转换为树形提示： 标题：... 成分：... 说明：... 代码：... # 代码结束 (1 次示例)
标题：<dish_name> 成分：<ingredient_abbreviation_list> 说明：[i1] <1st_direction> [i2] <2nd_direction> ... 代码：

树校正提示： 你将获得一个菜谱的标题、成分和制作步骤，以及表示该菜谱树结构的部分 Dot 代码。该 Dot 代码缺少一些边。此外，你将收到这些连接缺失的节点名称。对于每个提供的节点名称，从该节点到使用它的动作节点（如果它是一个成分）或处理其结果的动作节点（如果它是一个动作）添加恰好一条边。请只返回这些特定边的 Dot 代码，包括必要的注释，并排除任何额外的文本。标题：<dish_name> 成分：<ingredient_abbreviation_list> 步骤：[i1] <1st_direction> [i2] <2nd_direction> ... 部分

Dot 代码：<dot_code> 缺少边的节点名称：<node_names> 输出：

D 树编辑距离实现

在这个附录中，我们描述了计算配方树之间最小编辑距离的方法。正如在 3.3 节中提到的，我们采用了 Zhang-Shasha 算法 (Zhang and Shasha, 1989)，该算法扩展了著名的字符串编辑距离方法到有序标记树。具体来说，标记树是指每个节点从一个固定有限字母表中分配一个符号，有序树是指每组兄弟节点具有定义的从左到右的顺序。虽然对于无序树计算树编辑距离是 NP-hard，甚至是 MAX SNP-hard (Arora et al., 1992)，但 Zhang-Shasha 算法为有序情况提供了多项式时间解决方案。为了使我们的标记配方树与这种方法兼容，我们通过根据其标签按字典顺序排序来对兄弟节点施加顺序。

我们进一步调整编辑成本以鼓励在食谱之间匹配类似的节点。具体来说，我们允许一个节点被另一个节点替换，但仅当两个节点共享相同类型时（即，都是成分节点或者都是操作节点）。此外，如果两个节点具有相同的标签，则替换之间的成本为零；如果标签共享相同的抽象含义，则为一个固定的小成本。例如，两种均被分类为“草药”的成分可以以低廉的成本替换对方，而“草药”和“液体”成分则成本较高。同样，两个被分类在“加热应用”下的操作节点更可能相互替换，而一个被分类在“加热应用”下的操作节点与另一个被分类在“调味增强”下的操作节点则不如前者容易替换。

为了判断两个配料节点是否共享相同的抽象，我们使用从解析配料得到的配料抽象（见附录 B）。为了判断两个动作节点是否共享相同的抽象（例如，“烘焙”和“微波”的“加热应用”），我们收集了食谱中最常见的 250 个动作动词，并创建了一个层次结构，将这些动词分组为加热应用、准备、定位、味道增强等类别。

形式上，设 T_1 和 T_2 为两个由成分和动作节点构成的配方树。我们允许插入、删除和更新的操作。插入或删除的成本设置为 100，而更新的成本取决于两个节点是否具有相同的类型和相同的标签或抽象。如果它们具有相同的类型和相同的标签，成本为 0；如果它们具有相同的类型但仅有相同的抽象，成本为 5；否则，成本为 ∞ ，表示不可行的替代。该成本方案鼓励编辑距离算法更倾向于替换类似的部分而不是进行插入和删除，从而实现更具语义意义的转换。

如在第 3.3 节中所述，在转化过程中停止于不同的步骤可以创造出独特的菜肴（见图 4

)。此外，打乱编辑顺序可以生成全新的中间想法。为了产生更连贯的结果，我们对打乱后的操作施加部分顺序。我们优先插入和更新目标食谱中的关键风味成分（例如，在柠檬派中的“柠檬”），使它们在转化中更早出现。同时，我们推迟删除或更新来源食谱中的结构成分（例如，在千层面中的“千层面皮”），以保留其核心结构。我们在解析阶段确定哪些成分对风味及哪些对结构有贡献（见附录 B）。这一方法有助于保持两个菜肴的基本特征，整合目标菜肴中不同的风味成分，同时保留来源菜肴的结构完整性。值得注意的是，反向转化（菜肴 B 菜肴 A）会导致不同的编辑序列，产生截然不同的新食谱。

为了确保这些菜肴确实是重新组合的，我们舍弃任何缺少至少一个原始菜肴中的必要成分的食谱（如果存在这样的成分）。我们将一种成分定义为在某道菜的食谱中频繁出现但不是所有食谱中普遍常见的成分（例如，意大利宽面条在意大利千层面中的应用）。此外，我们移除与起始食谱太相似的想法。为了确保组合后的树保留跨菜肴的灵感源，我们要求其元素（节点和边）至少有 30% 来自每个原始食谱。

E 识别冲突成分

我们去除在味道上可能冲突的成分。具体来说，我们寻找在食谱库中很少同时出现的成分对，认为它们的配对是不常见的，并尝试确定这是否可能是一个创造性的成功或失败。

为此，我们依赖两个外部数据集。首先，我们使用 flavorDB (Garg et al., 2018)，它为包括水果、蔬菜和鱼在内的广泛原材料目录化味觉分子。受到此数据集所有者关于若两种原料共享大比例味觉分子则适合搭配的主张的启发，我们定义了一个基于 Jaccard 的原料搭配评分。由于 flavorDB 不涵盖加工成分（例如，由面粉、水和鸡蛋组成的千层面片），我们还使用 FoodData Central (Fukagawa et al., 2022) 来推断它们的原料成分。我们定义两种复合成分之间的搭配评分为其组成原材料对之间最低得分。在实验之后，我们选择 0.3 作为值的门槛。若得分低于 0.3，我们则认为这种搭配有问题。

F 从树到文本的提示

本节包括我们用来提示 GPT-4o 将结构化树表示转换回自然语言的提示。这里呈现的所有提示的系统消息是：“您是烹饪专家”。

将树翻译为原始食谱提示： 给定以下 DOT 代码，它通过定义成分节点、动作节点及其互连来图形化地表示食谱，将结构转换为自然语言食谱。DOT 代码将每种成分映射到特定动

作，并概述这些动作的顺序以展示烹饪过程。DOT 代码：“<recipe_idea_dot_code>” 请将此结构化表示转换为自然语言中的详细烹饪食谱。要求：(1) 输出应仅包括标题、配料及其数量和顺序指示。(2) 避免任何解释性评论或修饰。输出：

查找问题并修正食谱提示： 步骤 I: 查看下面提供的用自然语言撰写的食谱。识别并列其中任何潜在问题，不包括与不常见配料组合相关的顾虑。请仅提供潜在问题列表，无需修改食谱。食谱：“<GENERATED RECIPE>”

步骤 II: 请编辑食谱以解决识别的问题。确保食谱仍然是一个单一的统一组件。仅输出修正后的食谱版本。输出：

总结食谱提示： 请用几句话总结以下菜谱：(1) 以超简洁的方式对菜品进行描述，仅关注其最终结果。(2) 然后，对菜谱进行总结，包括其主要成分、操作和所有使用的材料。使用描述性语气来写这一部分，避免使用祈使句。菜谱：“<full_recipe>”

回顾成分提示： 给定一个创意食谱的描述。创意食谱描述：“<creative_recipe_description>” 你的任务是保留食谱中的创意成分，同时建议去除或替换可能对菜肴口味产生负面影响的成分。你应当：(1) 识别出为菜肴的创意做出贡献的独特和不寻常的成分。(2) 系统性地比较菜肴中的所有成分对，识别由于口味冲突而明显、严重冲突的成分。要仔细检查，确保包括所有可能的成分对，有严重冲突的都要包括。(3) 基于识别出的严重冲突，提出去除和替换成分的建议，以避免冲突，同时保留菜肴的创意性。只返回以下 JSON 输出格式：{ “dish_ingredients”: <字符串列表: 菜肴中完整的成分列表>, “creative_ingrs”: <字符串列表: 为菜肴创造性做出贡献的成分列表>, “flavor_clashes”: <字符串对列表: 冲突的成分>, “removals”: <字符串列表: 需要去除的成分列表>, “substitutions”: <字符串对列表: 需替换的成分 - (ingr1, ingr2) 表示”用 ingr2 替换 ingr1”> }

提高可读性提示： 给定以下食谱：(1) 去掉以下成分：<bad_ingredients>。(2) 进行以下原料替换：<required_substitutions>。(3) 将成分和步骤分成不同的部分以提高可读性（例如，“混合干料”，“组装”等）。你可以更改行的顺序，但保持内容不变。“<full_recipe>”

G 实验选择的提示

在本节中，我们展示了用于指导 GPT-4o 生成两个实验的配方的提示。

实验 1 提示： 你是一家融合餐厅的厨师，该餐厅擅长创造经典菜肴的美妙和意想不到的组合。你今天的任务是设计一个创新的食谱，将 { dish1 } 的复杂层次与 { dish2 } 的丰富奢华相融合。开发一个全面的食谱，包括：(1) 一个体现这种融合菜肴精髓的独特名称。(2) 详细的食材清单。(3) 分步的烹饪和组装说明，突出创新的烹饪技巧或不寻常的食材互动。鼓励大胆试验风味和质感，创造出既令人惊讶又令人满意的菜肴。

以下提示：设计另一个不同的创新食谱，将 { dish1 } 的复杂层次与 { dish2 } 的丰富奢华相融合。

实验 2 提示： 你能创造出最富有创造性和突破常规的食谱是什么？

以下提示：创造一个新鲜、独特的食谱，需与之前的食谱不同，但要匹配它们的创造性水平。

H 人类实验问卷

1. 这些食谱中的指示合理吗？一个不太合理的食谱通常包含技术问题。这些问题可能是小问题（例如，搅拌已经混匀的沙拉）或是大问题（比如没有将生鸡肉煮熟）。{ 评分标准 (1-5)：1：直接丢掉食谱，需要太多改动，2：需要更改大部分食谱，3：需要大量更改，4：几乎完美，只需一些小改动，5：完全合理，可以按原样烹饪 }
2. 这个配方中的成分组合是否合理？{ 标度：与 } 相同
3. 这个食谱与您见过或使用过的其他食谱有多相似？{ 评分 (1-5)：1：和许多食谱非常相似，2：相似但不太常见，3：有些类似于其他食谱，4：与大多数食谱完全不同，5：非常不同，极为不寻常 }
4. 与典型的食谱相比，这个食谱中说明的组合方式有多新颖？{ 量表：相同 }
5. 在这个食谱中，配料的组合与典型食谱相比有多新颖？{ 尺度：相同 }
6. 假设一名厨师遵循这个食谱，那么人们愿意品尝它吗？{ 标度 (1-5)：1：绝不会，2：只有必要时才会，3：只有真的很饿时

才会，4：他们可能会尝试，5：是的，绝对会 }

7. 假设一位厨师在进行了所需的修改后按照这个食谱操作，人们会想尝尝吗？{ 量尺度：相同 }
8. 整体而言，您觉得这个食谱有多原创？(在做出的必要修改之后) { 评分 (1-5)：1：完全没有原创性，2：略有原创性，3：有些原创性，4：相当有原创性，5：极具原创性和创造性 }