本文已提交至 IBPSA-USA 第十二届全国会议并获接收。会议于 2026 年 5 月 20 日至 22 日在明尼苏达州明尼阿波利斯举行。

BEMEval-Doc2Schema：建筑能耗建模中结构化数据抽取的大语言模型基准测试

Yiyuan Jia[1*], Xiaoqin Fu[2], Liang Zhang[2,3]

1. Big Ladder Software，美国科罗拉多州丹佛 2. 亚利桑那大学，美国亚利桑那州图森 3. National Laboratory of the Rocikes，美国科罗拉多州戈尔登 *通讯作者：yiyuan.jia@bigladdersoftware.com

摘要

基础模型（包括大语言模型（LLM））的最新进展为自动化建筑能耗建模（BEM）带来了新的机遇。然而，由于缺乏公开可用、面向特定任务的数据集和标准化性能指标，系统性评估一直面临挑战。我们提出 BEMEval，这是一个用于评估基础模型在各项建筑能耗建模任务中表现的基准测试框架。该套件中的首个基准测试 BEMEval-Doc2Schema，聚焦于从建筑文档中抽取结构化数据，这是实现建筑能耗建模流程自动化的基础步骤。

BEMEval-Doc2Schema 引入了键值重叠率（KVOR），这一指标用于量化大语言模型生成的结构化输出与真实模式参考之间的一致性。我们使用该框架，在三个数据集——HERS L100、NREL iUnit 和 NIST NZERTF——上，采用零样本和少样本提示策略评估了两种领先模型（GPT-5 和 Gemini 2.5）。结果表明，Gemini 2.5 的表现始终优于 GPT-5，且少样本提示能够提升两种模型的准确率。模型表现也因模式而异：EPC模式的 KVOR 得分显著高于 HPXML，这反映出其层级结构更简单、层级深度更浅。通过结合精选数据集、可复现的指标和跨模型比较，BEMEval-Doc2Schema 建立了首个由社区推动的基准测试，用于评估大语言模型执行建筑能耗建模任务的能力，并为未来有关人工智能辅助 BEM 工作流的研究奠定了基础。

引言

建筑约占全球能源消耗的 30% 和二氧化碳排放的 26%（IEA，2023年），因此成为全球实现脱碳、减少温室气体排放工作中的重点（Xiang、Ma 等，2022）。建筑能耗建模（BEM）是一项关键工具可用于分析和优化建筑性能，助力实现能效提升和减碳目标。

建筑能耗建模流程通常由一系列相互关联的任务组成，将设计意图和文档转化为仿真就绪的输入数据格式。流程始于信息收集与准备阶段，能源建模人员在此阶段收集项目文件，例如设计简报、审计报告、图纸和规范，并确定相关的规范要求和参考数据。下一阶段是数据整理与提取，将这些异构且往往非结构化的信息转换为结构化的仿真就绪数据。

然而，在当前实践中，很少有建模工作流采用统一的“建筑能耗建模模式”来存储结构化建筑信息。相反，建模人员通常依靠定制电子表格或项目专用数据库，整理仿真就绪输入数据，例如围护结构参数、内部负荷、日程安排、暖通空调分区及系统细节。随后，这些内部数据表会通过人工或半自动方式映射到 EnergyPlus/Openstudio、DesignBuilder、IES-VE 或 eQuest 等建模软件所需的输入格式。因此，工具之间的结构化数据交换仍然有限，互操作性方面的挑战依然存在。

在进行结构化数据抽取的同时，还会进行几何定义：根据图纸或建筑信息模型重建建筑几何，并将其简化为适用于模拟的热区。后续任务包括暖通空调系统建模和基准生成，以确保符合适用的性能标准（如 ASHRAE 90.1）；随后进行模拟和验证，检查模型行为与结果是否一致且合理。最后，通过后处理和报告，将仿真输出汇总为性能指标、可视化内容和合规文档。

使用标准化模式以机器可读且可互操作的方式表示建筑数据这一概念

这种格式开始逐渐得到认可，尤其是在住宅领域。其中一项代表性工作是由建筑性能研究所（BPI）开发的住宅性能可扩展标记语言（HPXML），它为住宅能源信息定义了一套通用数据字典与传输标准（BPI，2013年）。这类基于模式的方法为系统地将非结构化或半结构化的建筑文档转换为结构化表示奠定了基础，从而实现更一致的数据交换，并为建筑能耗建模流程中的自动化创造了机会。

大语言模型（LLM）的最新进展为自动化完成其中许多任务带来了新机遇。它们能够理解自然语言、对异构数据进行推理并生成结构化表示，这与BEM工作流信息丰富、以文本为主的特点高度契合。Zhang等人（2024年）开发了一种自动生成EnergyPlus输入文件的工作流程，将语言模型与基于规则的启发式方法相结合，但该方法缺乏标准化指标和可复现数据集。Gang等人（2024年）进一步展示了大语言模型如何支持建筑能耗模型的自动元数据提取和模式映射，凸显了其减少模型准备中人工工作量的潜力。
因此，大语言模型已开始成为自动化文档解析、模型生成和数据验证的有力工具。与此同时，自然语言处理领域的研究也发展出了更先进的提示和推理策略。思维链提示（Wei等人，2022年）和自我一致性（Wang等人，2022年）均提升了通用任务中的结构化推理表现。智能体工作流（Schick等人，2023年）也越来越受关注，因为它能够执行比提示更复杂的任务。这些进展共同为将大语言模型的推理能力应用于特定领域的建模与仿真流程奠定了基础。

尽管大语言模型日益广泛地应用于各个领域，包括建筑能耗建模（BEM），但由于幻觉问题，其准确性和一致性始终面临挑战（Ji等人，2023年）。这一问题对于工程应用尤为关键，因为此类应用要求高度的可重复性和精确性。因此，通过系统化基准测试对大语言模型的表现进行定量评估，对于评判和提升其可靠性至关重要。目前已有许多研究用于评估大语言模型的表现。在建筑、工程与施工领域，Arch-Eval（Wu等人，2025年）通过特定领域的问答评估大语言模型的建筑学知识。DrafterBench（Li等人，2025年）则提出了一个用于解读土木工程文档的框架。相关领域的近期研究表明，人们对模型评估的兴趣日益浓厚。

尽管人工智能长期以来一直应用于能源预测与优化，但针对建筑能耗建模任务中的语言模型开展正式基准测试，直到最近才刚刚起步。目前的研究仍较为零散，缺乏严谨的基准测试框架，无法整合针对特定任务的基准数据集、标准化的性能指标，并全面覆盖端到端建模工作流。没有这样的框架，就难以比较不同方法、量化进展，或确定有针对性的改进机会。

为弥补这一空白，我们推出了BEMEval，这是一个开放的基准测试框架，旨在评估基础模型在建筑能耗建模任务中的表现。该框架采用渐进式扩展设计：套件中的每项基准测试均针对一项不同任务，并配有精选数据集、真实值参考和定量指标。随着基础模型的能力从文本扩展到视觉、多模态推理和物理仿真，BEMEval的范围也将相应扩大。它填补了研究领域缺乏标准化基准测试数据集和评估协议的空白，如同计算机视觉领域的ImageNet（Deng等人，2009年），有助于确保不同工具和建模平台之间的互操作性与可复现性。
作为该套件的第一步，我们将最基础且最耗时的任务作为初始重点：将非结构化建筑描述映射为结构化的建筑信息。这项任务直接发挥了语言模型的核心优势：语境理解、语义提取和结构化数据生成。

在本研究中，我们介绍了 BEMEval-Doc2Schema，这是 BEMEval框架中的首个基准测试，用于评估大语言模型和提示工作流将原始住宅建筑文档转换为适用于能耗建模的标准化机器可读输入的准确性。住宅建筑是开发和评估基于大语言模型的建筑能耗建模自动化的理想起点。与商业或工业设施相比，住宅建筑在几何形状、围护结构组件和暖通空调系统类型方面具有更高程度的标准化。这种一致性降低了建模差异，并使自动化输出与参考数据之间能够进行更直接的比较。此外，住宅领域拥有相对成熟的模式基础设施，以及不断发展的开放数据集生态系统。HPXML（住宅性能可扩展标记语言）（BPI,2013年）等标准提供了

特征，能够在能源审计、改造和仿真工具之间实现[illegible]完备的住宅测试设施，包括 HERS 测试案例 L100、美国国家标准与技术研究院净零能耗住宅测试设施（NZERTF）以及 NREL iUnit，提供了详细的文本和定量建筑描述，以及部分结构化的仿真就绪数据。

在实际的建筑能耗建模工作流程中，BEMEval-Doc2Schema 对应于项目早期阶段：能源建模人员从设计文件、审计报告和规范中提取建筑特征，并在模拟前将其填入结构化输入模板或数据库。基于大语言模型的结构化数据提取并非旨在取代仿真引擎或建模专业知识，而是为了减少人工转录工作、提高一致性并加快模型设置。BEMEval 提供了一种标准化的评估方式，用于检验此类工具是否达到最低工程要求，以便将其集成到生产工作流程中。

本文结构如下。方法部分介绍 BEMEval 框架及其范围，以及其首个组成部分 BEMEval-Doc2Schema，包括基准测试的结构及其关键组成部分。案例研究部分介绍使用 BEMEval-Doc2Schema 对特定语言模型和提示策略进行的评估。讨论部分总结案例研究的主要结果和见解。最后，结论部分重点阐述本文的主要贡献，展示 BEMEval 的实用价值，并概述构建自动化建筑能耗建模工作流程稳健评估生态系统的长期愿景。

方法

BEMEval框架与范围

BEMEval 是一个模块化基准测试框架，旨在评估基础模型在不同建筑能耗建模任务中的性能。BEMEval 并不将 BEM 自动化视为一个单一的整体性问题，而是将建模工作流分解为多个面向具体任务的基准测试，每个基准测试都具有明确定义的范围、精选数据集和定量评估指标。这种模块化结构能够在建筑能耗建模流程的不同阶段，对模型能力进行受控评估，同时支持基准测试套件的逐步扩展。

BEMEval框架旨在随着时间推移不断扩展，涵盖更多符合相同核心筛选标准的建筑能耗建模任务：对从业者影响重大、适合作为基准测试，且适用于基础模型评估。目前已确定两个近期候选任务。第一个是规范合规推理，即根据建筑描述，使用 ASHRAE 90.1附录G 等标准生成符合规范的基准配置。由于相关规则公开记录、具有确定性且以文本形式呈现，并且可以直接根据标准推导出真实值，因此这项任务非常适合用于基准测试。第二个是质量保证和错误检测，重点在于审查仿真输出，以识别异常、不合理结果和建模错误。这项任务本质上由推理驱动，其基准测试真实值可以根据专家标注的案例或刻意植入错误的错误模型构建。图1展示了 BEMEval 框架。

随着基础模型的能力扩展到视觉和多模态推理，BEMEval 还可以面向涉及建筑图纸解读、空间推理和几何抽象的任务。开发这些未来的基准测试，需要与实际从事工作的能源建模人员密切合作，以整理具有代表性的测试案例，并根据专家判断验证评估指标。

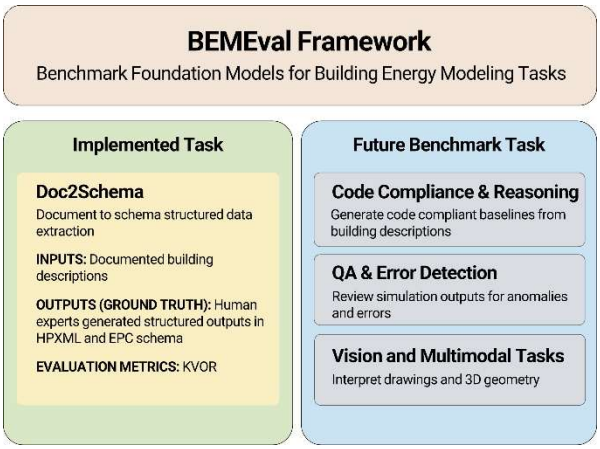

图*1 BEMEval*框架

BEMEval-Doc2Schema 是在 BEMEval 框架内实施的首个基准测试。选择这项任务基于三项标准。首先，这项任务对从业者影响重大，因为结构化数据抽取是当前建筑能耗建模实践中最耗费人力且重复性最高的任务之一。其次，这项任务适合作为基准测试，因为现有文档完善的建筑案例和标准化模式能够提供清晰且可审计的真实值。第三，这项任务非常适合大语言模型，因为它主要依赖文本理解、语义解释和领域推理。

BEMEval-Doc2Schema 基准的目标是提供一套标准化测试与评分框架，用于评估大语言模型和不同提示机制将非结构化建筑能耗建模文档转换为结构化且符合模式要求的数据的效果。该基准测试评估模型解读建筑特定信息、运用建筑能耗建模知识，以及正确填充目标模式且不引入缺乏依据的假设的能力。

BEMEval-Doc2Schema 由三个核心部分组成：输入、参考输出或真实值，以及评估指标。以下各节将如图1所示，详细介绍每个部分。

输入：真实建筑（L100（PSD Consulting 2014年）、NZERTF（NIST 2014年）和 iUnit（DOE 2024年））的非结构化建筑能耗建模文档，格式包括 PDF、电子表格等。

输出（真实值）： 采用现有广泛使用的开放模式（HPXML 和/或 EPC）格式的结构化数据集。部分 1 和部分 2 被视为输入-输出对或问答对。

评估指标：定义评估指标，用于量化部分 2 中的正确答案与大语言模型生成的输出之间的接近程度。

NZERTF 建筑描述基准测试

美国国家标准与技术研究院净零能耗住宅测试设施（NZERTF）是一栋配备大量仪器设备的独栋住宅式建筑位于马里兰州盖瑟斯堡美国国家标准与技术研究院园区内的住宅。设计旨在模拟典型的郊区两层住宅（居住空间约 2,700 ft2 / 252 m2）住宅设有地下室和阁楼（各 ~1,450 ft2 / 135 m2），总体积为 44,900 ft3（1,270 m3）。该设施被设计为一座“实验室住宅”，用于展示和研究一栋规模与形式均为人们所熟悉的住宅建筑，能否通过现场可再生能源，在一年内产生与其消耗量相当的能量（即实现净零能耗）。它还可作为研究平台，用于验证建筑能耗模型、评估高性能围护结构和系统技术、监测室内环境质量，并探索建筑系统之间的相互作用，包括暖通空调、通风和光伏。



图*2* 美国国家标准与技术研究院净零能耗住宅测试设施

在 BEMEval-Doc2Schema 框架中，NZERTF 是独栋住宅类型的一个代表性数据点。它提供了丰富且记录完善的设计与施工数据，并包含高效暖通空调系统和可再生能源系统的详细设备文档

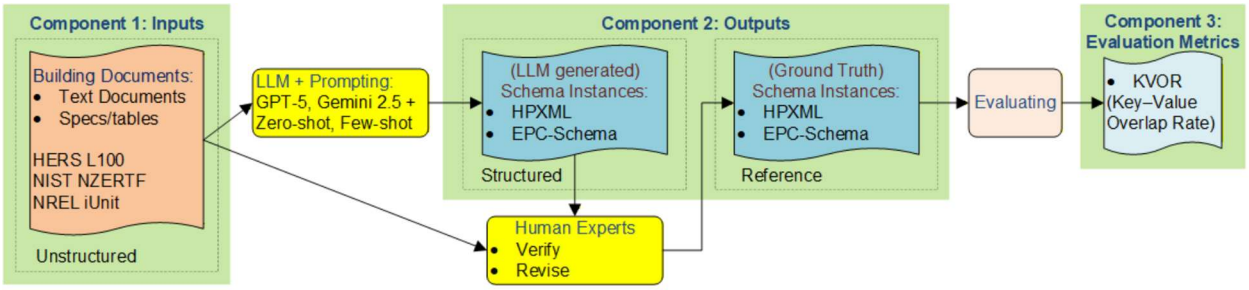

图*2 BEMEval-Doc2Schema*

iUnit

NREL iUnit 是一款模块化、工厂建造的公寓，由美国国家可再生能源实验室（NREL）与丹佛-由开发商 iUnit 打造。iUnit 被设计为一款高性能的电网互动原型，代表了多户住宅领域中典型的微型公寓布局。该原型的建筑面积约为 38 m2

(~400 ft2)，在场外建造后安装于美国国家可再生能源实验室位于科罗拉多州戈尔登的户外研究区。

iUnit 融合了先进的围护结构设计、高效的机械系统和现场可再生能源技术，旨在探索实现低能耗工厂化生产住宅的途径。在 BEMEval-Doc2Schema 框架中，iUnit 是公寓及多户住宅建筑类型的代表性数据样本。

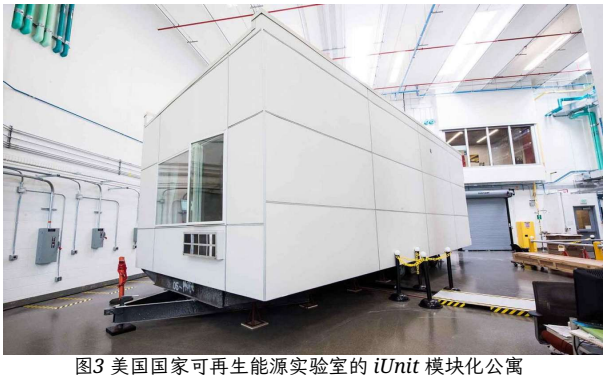

图*3* 美国国家可再生能源实验室的 *iUnit* 模块化公寓

L100

该基础建筑是实施住宅能源评级系统（HERS）建筑能耗模拟测试（BESTEST）的基本参考模型。BESTEST 是一种用于评估住宅能源评级系统所用建筑能耗软件可信度的方法。

该建筑是一个标准化的住宅建筑，用于评估和比较不同软件工具的能耗模拟结果。大多数其他案例都以此案例为基准进行比较，以确定评级工具能否准确识别建筑配置变化所导致的能耗差异。

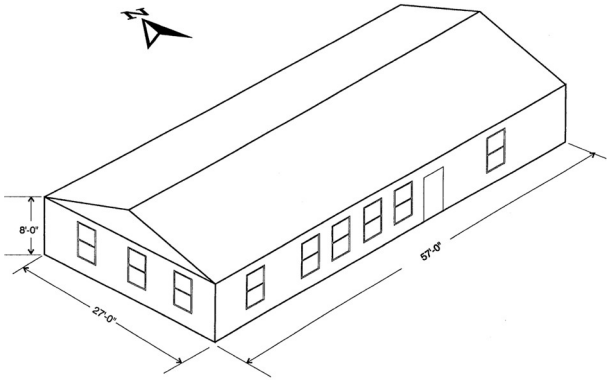

图*4* 美国国家可再生能源实验室用于 *HERS BESTTEST* 的 *L100* 基础建筑

基准建筑 L100 是一栋面积为 1539 平方英尺的单层木框架住宅，设有完全通风的爬行空间，并安装了总面积为 270 平方英尺的单层玻璃窗（北向和南向各 90 平方英尺，东向和西向各 45 平方英尺）。墙体采用 R-11 保温材料，天花板和地板采用 R-19 保温材料。它有一个供暖空调区（主层）、一个未供暖空调阁楼和一个通风爬行空间。其中存在与暖通空调无关的内部产生负荷，例如设备、照明、人员、动物等。对于基准建筑，我们仅考虑其机械系统产生的纯负荷输出，并假设所有设备（包括风管）的效率均为 100%。此外，还作出三项假设：1）空气系统完全（100%）采用对流；2）恒温器仅响应空气温度；3）恒温器为非比例式。测试提供了两个地点：科罗拉多州科罗拉多斯普林斯（晴冷气候）和内华达州拉斯维加斯（炎热干燥气候）

表*1* 模式概要

	描述	创建者	特征	分层？（是/否）	定制？（是/否）	链接至网站
HPXML	HPXML（住宅性能 XML）是一种数据传输标准，适用于住宅能源审计和改造项目，旨在确保之间的互操作性软件工具，承包商和项目管理人员在住宅能源能效领域。	BPI NREL DOE	侧重于详细的住宅审计，改造和能源模型数据交换；支持基于可扩展标记语言的互操作性；符合 RESNET 和能源之星住宅升级程序。	Yes	No	https://hpxml.nrel.gov/

EPC	EPC（能源 性能计算器） 模式是一种面向模拟的 数据模式 用于实现规范性能耗 计算 用于评估建筑能源性能的标准（CEN/ISO 13790、EN 15603 等） 和 认证，适用于 住宅和 商业建筑。	佐治亚理工学院 High 性能 建筑实验室	对参数进行编码 由 ISO/CEN 定义 热能标准 能源、交付能源和 一次能源，以及 排放；支持 基于模拟的 能源性能 评估。	No	Yes	https://github.com/BEMEval/doc2schema

模式描述

HPXML

住宅性能 XML (HPXML)模式是一种行业标准数据规范，旨在促进住宅能耗建模、审计和评级工具之间的信息交换。HPXML 由建筑性能研究所 (BPI)、美国能源部 (DOE) 和美国国家可再生能源实验室 (NREL) 合作开发，提供分层、机器可读的结构，以一致且可互操作的格式描述住宅建筑特征，包括围护结构组件、暖通空调和生活热水系统、设备效率、燃料类型、人员占用时间表及其他运行细节。在 BEMEval-Doc2Schema 中，HPXML 作为参考模式，用于评估语言模型将非结构化住宅建筑描述映射为结构化仿真就绪数据的能力。由于它编码了详细 EnergyPlus 或 OpenStudio 模型所依据的相同物理属性和系统属性，因此能够真实、标准化地呈现住宅建模领域。因此，在 BEMEval-Doc2Schema 中纳入 HPXML，可使基准测试立足于行业认可且广泛采用的框架，确保评测任务符合当前实践，并能直接用于下游生产工作流程。

EPC-Schema

能源性能计算器（EPC）是一种降阶建筑能耗建模框架，由佐治亚理工学院的前研究生在 Godfried Augenbroe 教授指导下开发。该电子表格工具依据 CEN/ISO 规范性计算标准（ ISO 13790 、 EN 15603 等），用于评估建筑能源性能（Lee 等，2011）。EPC 不采用传统模型中的详细模拟方法，而是采用规范性计算范式，遵循系统化的准稳态方法来估算能耗性能，该方法基于标准化假设和简化的物理关系。这种方法弥合了高保真模拟与经验基准评估之间的差距，提供了一种快速、透明且可复现的建筑性能表征方式（Lee 等，2013）。

EPC模式使用 TOML（Tom 的显而易见的极简语言）作为数据序列化格式，对 EPC 仿真引擎的输入进行标准化（TOML，2021年）。TOML 是一种轻量、易读的配置格式，在简洁性与严格类型之间实现了平衡，非常适合表示规范性计算方法所需的结构化输入参数。每个 EPC 输入文件都将建筑围护结构特征、系统效率、时间表和气候数据等模型参数组织为扁平的键值格式。该格式同时支持可解释性和自动化，确保 EPC 模型能够轻松进行参数化、版本控制和程序化解析。

建筑描述、文档、模式定义和评估资源均已在开源 BEMEval-Doc2Schema 存储库中公开，地址为 https://github.com/BEMEval/doc2schema。该存储库包含具有代表性的住宅建筑案例（例如 HERS L100、NIST NZERTF 和 NREL iUnit）、以 HPXML 和 EPC-Schema 格式提供的相应结构化参考输出、模式规范与元数据，以及实现键值重叠率（KVOR）指标的自动化评估脚本。此外，还提供了完整的文档和入门教程，以支持可复现性、可扩展性和社区采用。

评论

选择 HPXML 和 EPC 的原因在于，它们在详细程度和侧重点上有所不同。HPXML 强调详细的住宅改造和审计数据交换，从而实现承包商、项目管理人员和软件工具之间的互操作性（BPI 2013年，HPXML 指南 2023年）。另一方面，EPC 侧重于规范性计算

方法以及商业和住宅领域的建筑能源性能评级。

同样，这三个建筑案例的选择旨在代表本研究涵盖的不同住宅建筑类型。这种多样性确保基准测试覆盖具有代表性的住宅配置和数据结构。

未来的工作将扩展建筑组合和模式覆盖范围，纳入多户住宅建筑、工厂建造住宅以及更多开放模式格式。

评估指标

本研究采用的评估指标是键值重叠率（KVOR），用于量化大语言模型生成的结构化输出与真实参考的匹配程度。由于 BEMEval- Doc2Schema 侧重于将非结构化建筑文档映射为符合模式的格式，例如 HPXML 和 EPC-Schema，该指标能够直接且直观地衡量模型的提取准确率。

$$KVOR = \frac{Number\ of\ correctly\ matched\ key{-}value\ pairs}{Total\ number\ of\ keys\ in\ the\ reference\ sche} \quad (1)$$

此公式中，分子表示模型正确识别键及其对应值，且二者均与真实值一致的键值对数量。分母表示给定数据集中参考模式定义的键总数。所得比率介于 0 到 1 之间，数值越高，表示大语言模型生成的输出与预期结构化数据越一致。

即使大语言模型的输出通过了模式对齐验证，也仍可能因语义误解、上下文歧义或领域知识局限，将非结构化文档中的某个值分配给错误的字段。键值重叠率通过惩罚键和值之间的错误关联来衡量此类差异，从而为结构和语义准确性提供定量指标。

键值重叠率越高，说明模型理解建筑信息并将其准确映射到指定模式字段的能力越强。反之，较低的比率则表明模型难以保持文档内容与模式结构之间对应关系的一致性。因此，该指标为比较自动化建筑能耗建模工作流程中不同大语言模型和提示策略的表现，提供了清晰且可复现的依据。

案例研究

本节中，我们在自行开发的 BEMEval 框架下，对两种大语言模型和两种提示工作流进行了评估，以检验它们将 L100、iUnit 和 NZERTF 的原始建筑描述转换为 HPXML 和 EPC 结构化模式的能力。

我们使用的两种基础模型是 GPT-5 和 Gemini 2.5。GPT-5 由 OpenAI 开发并于 2025年 发布，它在 GPT系列 的基础上进行了扩展，增强了推理能力，延长了上下文处理能力，并改进了结构化数据生成，因此适用于涉及模式对齐和技术文本解读的任务。Gemini 2.5 于 2025年 由 Google DeepMind 推出，融合了多模态理解、符号推理和事实依据。该模型旨在实现高度一致的分析推理，尤其擅长将复杂文档转换为结构化的机器可读格式。

两种提示策略分别是：1）零样本提示；2）少样本提示。零样本提示是指模型仅接收任务指令和模式文档，不提供任何先前示例。它完全依靠内部知识和推理来完成映射。少样本提示是指模型还会获得一组输入输出示例，用于展示正确的模式格式化和映射行为。这些示例会引导模型在处理新输入之前学习映射约定。

下面的列表展示了零样本提示模板：

系统／角色指令：
你是建筑能耗建模、数据标准化以及{schema_name(e.g., HPXML)}模式领域的专家

任务：
你的任务是将非结构化建筑能耗建模信息——这些信息可能来自PDF文件、Excel文件、Word文档、纯文本或扫描版描述——转换为结构化{schema_name(e.g., HPXML）格式的{file_format(e.g., .xml)} 文件。

非结构化建筑信息以an 附件 调用了{name_of_unstructured_data.pdf}

HPXML模式在附件中有所描述名为：{schema_documentation.pdf}

请严格遵循以下规则：1. 提取与映射规则

- 输出应严格遵循 HPXML 格式

- 不要假设或推断输入中未明确说明的任何值
- 忽略建筑几何相关的输入，例如坐标

2. 每个 HPXML 字段都必须可追溯——在每个字段上方添加 XML 注释 (<! -- -->)，说明：

- 找到该值的确切来源文本或单元格，以及
- 将其放入该特定 HPXML 元素的原因。
- 如果缺少某个值，则完全省略该字段（不要填入默认值）

清单 *1.* 零样本提示模板

以下清单展示了添加到清单 1 中、用于少样本提示的额外提示内容：

以下是[N] 组非结构化输入及其对应的HPXML输出示例，供参考。模型在处理新输入之前，应先从这些示例中学习格式、注释风格和映射行为。示例输入1：{Example_input_1} 示例输出1：{Example_output_1} ..... 示例输入n：{Example_input_n} 示例输出n：{Example_output_n}

清单 *2*。少样本提示模板

结果表 *2*。模型性能摘要（键值重叠率，*KVOR*）

得分：KVOR	**L100（HPXML）**	**iUnit (EPC)**	**NZERTF (HPXML)**
零样本 - ChatGPT	0.000	0.480	0.005
零样本 - Gemini	0.040	0.851	0.166
少样本 - ChatGPT	0.010	NA	NA
少样本 - Gemini	0.082	NA	NA

结果表明，模型能力与提示策略均存在明显差异：

模型比较：在所有数据集上，Gemini 2.5 的表现始终优于 GPT-5，展现出更强的模式理解能力和更一致的值映射。Gemini 在 iUnit EPC 数据集上的得分尤其高（KVOR = 0.851），这表明它更善于解读结构化的键值关系，并在扁平 EPC 模式格式中保持上下文一致性。

提示策略：两种模型在少样本提示下准确率均有所提高，证实了基于示例的指导对模式对齐的有效性。这一改进在Gemini的L100案例中尤为明显（从0.040提升至0.082）。

模式复杂度：EPC模式的表现总体上高于HPXML，主要是因为EPC的结构更简单、归一化程度更高。HPXML更深的层级结构、更高的属性多样性以及内嵌的领域惯例给模型带来了更多挑战，导致重叠率较低。

建筑案例难度：在基于HPXML的数据集中，NZERTF的重叠率最低，反映出其更高的系统复杂度以及更高密度的相互依赖参数（例如暖通空调、通风、可再生能源系统）。

总体而言，结果表明基于大语言模型的模式抽取仍然对数据结构、模式深度和提示设计较为敏感。尽管Gemini展现出有前景的泛化能力与准确率，但要实现持续改进，仍需将领域导向微调、单位归一化以及混合验证流水线整合到未来版本的BEMEval中。

讨论

本研究结果凸显了使用大语言模型自动将非结构化建筑描述转换为标准化能源建模模式的潜力与当前局限。

在所有案例中，Gemini 2.5 的表现均优于 GPT-5，展现出更强的模式理解和上下文推理能力。它在 iUnit EPC 数据集上的 KVOR 显著较高（0.851），这表明它更能适应简单、扁平的数据结构。HPXML 更广泛的层级结构以及跨系统相互引用的标识符增加了关系复杂性，使一致映射更加困难。在 HPXML 案例中，NZERTF 的得分最低，这反映出其系统细节丰富且结构高度参数化。

两种模型均受益于少样本提示，这证实了基于示例的指导有助于模式对齐。然而，由于可用

示例有限，评估只能采用一种少样本配置，因此只能部分反映模型的上下文学习能力。

为了恰当地解读这些结果，必须注意 KVOR 指标的范围与局限性。KVOR 对所有模式键一视同仁，不会根据参数对仿真结果的相对影响进行加权。虽然这简化了解释并支持与模式无关的比较，但未来工作将探索敏感性感知加权方案、参数分组以及下游仿真影响指标。当前实现有意保持保守并以任务为中心，优先考虑抽取正确性而非性能归因。

尽管本研究的主要目标是建立一个基准测试框架，而非一个完全自动化的系统，但研究结果为未来工作流和数据模式如何演进以更好地适应基于人工智能的自动化任务提供了有价值的见解。在不同的模型配置中，包括零样本、少样本提示工作流，出现了一个一致的趋势：当模式具有更高的抽象程度和更低的结构复杂度时，模型的表现更为有效。这些结果表明，在语言模型更有能力处理复杂数据结构之前，应更加重视设计能够更好地支持部分自动化的“对人工智能友好的”模式。简化、语义清晰且文档完善的模式有可能帮助弥合自然语言输入与结构化模型表示之间的现有差距。虽然模式设计并非本研究的核心重点，但它已成为未来探索的一个潜在方向。一个可能的方向是建立模式评估指标，用以衡量不同模式设计特征如何促进自动化性能的提升。

另一个值得注意的观察是，在执行结构化数据生成任务时，大语言模型往往倾向于自动化，而非手动推理。当被要求将非结构化建筑描述转换为基于模式的表示时，模型经常会编写代码或脚本来自动完成数据提取和整理，而不是直接自行进行映射。这种行为表明，模型将此类任务视为程序合成问题，把任务看作生成可完成任务的流程的机会，而非进行明确的语义推理。虽然这体现了模型试图以资源效率更高或计算成本更低的方式完成任务，但也暴露出一种偏见或走捷径的思维倾向：模型假定可以通过代码解决问题，而不是理解其背后的领域语义。

BEMEval-Doc2Schema 的一项核心设计原则是确保可复现性，并能切实融入现实世界的BEM工作流。所有基准数据集均源自记录完善的住宅测试案例，具有明确定义的输入、模式参考和评估标准，从而能够在不同模型、提示机制和实验设置之间进行一致的比较。该基准与常见的BEM实践保持一致，明确针对中间层的仿真就绪数据表示（例如 HPXML 和 EPC-Schema），这些表示可直接供下游工具使用，包括 EnergyPlus 或基于 OpenStudio 的工作流程。通过关注模式级正确性，而非特定工具的输入文件，BEMEval-Doc2Schema 将语言模型评估与个别仿真引擎解耦，同时保留与现有建模流水线的兼容性。这一设计使 BEMEval-Doc2Schema 既可作为受控的研究基准，也可作为自动化或半自动化BEM生产工作流中的验证层。

BEMEval-Doc2Schema 作为一个面向社区的开源框架发布，旨在推动透明评估、可扩展性和广泛采用。基准资源包括数据集、模式定义、评估脚本和文档，均已公开，以支持独立复现结果，并促进新增任务、模式和建筑案例的贡献。通过采用开放标准和模块化设计，BEMEval-Doc2Schema 使研究人员和从业者能够将基准测试扩展到更多建筑类型、建模阶段和语言模型架构。随着时间推移，这一开放基准测试生态系统旨在推动人工智能辅助建筑能耗建模不断取得进展，使社区能够跟踪性能改进、识别持续存在的失效模式，并为未来的方法学进步建立共享基线。

局限性与未来工作

尽管本研究初步探索了如何评估大语言模型在建筑能耗建模中的应用，但仍需承认若干局限性。

当前版本的 BEMEval-Doc2Schema 基准仅包含数量有限的数据点。尽管这些示例涵盖了具有代表性的建筑类型和模式，但尚未体现住宅建筑能耗建模实践的全部多样性。例如，HERS L100 案例可作为标准化参考，但未包含明确的暖通空调系统信息。NIST NZERTF 案例提供了详细的围护结构和系统数据，但它代表的是一栋经过高度仪器化的原型住宅

，针对特定气候进行了优化，这限制了其在更广泛住宅环境中的普适性；而 NREL iUnit 案例采用了简化的内部负荷和时间安排，主要是为校准研究而制定的。未来的工作将着重扩展数据集，纳入更多种类的建筑类型、建筑年代和系统配置，从而提升其代表性和普适性。

数据点数量有限也制约了评估方法。在少样本设置中，可用示例进一步受到限制：仅进行了一种测试配置，即以 HERS L100A 为目标案例，并使用 NZERTF 输入–输出对作为少样本示例。未测试反向配置（以 NZERTF 为目标案例，并使用 L100A 示例）；而采用 EPC-Schema 的 iUnit 案例缺少用于少样本评估的相应示例。因此，少样本结果仅部分反映了模型的上下文学习潜力。

本研究评估的模型是未经任何领域特定微调的通用大语言模型。尽管它们展现出广泛的语言和推理能力，但缺乏对建筑物理、暖通空调系统和模式约定的专业理解。因此，其表现反映的是基线泛化能力，而非通过定向适配所能达到的上限。未来工作应探索使用建筑设计和仿真语料库以及模式文档进行微调或指令优化的模型，使其更好地适应 HPXML 和 EPC 风格数据集的语义。将这些策略与扩充数据集及更均衡的少样本配置结合起来，将有助于深入评估领域适配语言模型如何提升自动化建筑能耗建模工作流程的准确性、一致性和可解释性。

为定量评估所提取数据的准确性，我们将评估率定义为大语言模型答案中正确关键值的数量与模式中预期关键值数量之比。正确值的定义取决于数据类型。对于基于字符串的属性，采用若干字符串相似度指标来评估文本对应关系，包括基于编辑距离的度量，如莱文斯坦距离（Levenshtein 1966年）和 Jaro–Winkler 相似度（Winkler 1990年），基于词元的度量，如 Jaccard 和 Sørensen–Dice 系数（Jaccard 1912年，Sørensen 1948年），以及基于向量的度量，如余弦相似度（Salton 等 1983年）。这些指标使用 RapidFuzz 和 TextDistance 等 Python 库实现（RapidFuzz 文档 2025年；TextDistance 文档 2025年），可提供高效且可复现的相似度计算。相似度阈值（例如，归一化字符串指标为 ≥0.9，或数值偏差为 ≤5%）用于判定所提取的值是否被视为正确。在未来工作中，将开发自动化验证工具链，以整合这些相似度指标、单位归一化流程和可配置阈值，从而系统性地验证结果正确性。

然而，这些传统相似度指标主要捕捉表层词汇或句法对应，往往无法考虑建筑相关文本中的语义等价或语境含义（Reimers 等人（2019））。例如，“机械通风”和“暖通空调系统”等表述可能指向相似的概念，但字符串相似度分数却很低。此外，数值比较可能忽略单位换算不一致或依赖语境的容差，而这些情况在工程数据集中很常见。

未来的改进可能包括采用 Transformer 编码器（如 Sentence-BERT ）纳入基于嵌入的语义相似度度量，并使用领域专用归一化规则，以更准确地反映提取值与参考值之间的真实等价性（Zhang 等人（2020））。此外，BEMEval 的未来扩展将引入面向鲁棒性的评估维度，例如通过多随机种子测试评估输出稳定性、系统地测量幻觉率，以及分析不同模型之间的成本—延迟—准确率权衡。我们将开发自动化验证工具链，整合这些相似度指标、单位归一化流程、鲁棒性指标和可配置阈值，以便在真实部署条件下系统地验证结果正确性。

除了拓展技术范围外，基准测试的长期成功从根本上取决于社区能否持续协作。营造一个活跃、包容且得到充分支持的开源社区，对于确保基准测试持续提升其相关性与严谨性至关重要。完善的文档、易于获取的教程以及开放的评估资源，有助于降低参与门槛，并使研究人员、从业者和开发者能够共同为基准测试作出贡献、验证并加以扩展。

结论

本研究的主要贡献是提出并初步实现了 BEMEval，这是一个用于评估大语言模型在建筑能耗建模任务中表现的框架和数据集集合。通过采用一致的评估协议、透明指标和可复现数据集，BEMEval 为比较模型在各类建筑能耗建模任务中的性能奠定了标准化基础。

该基准测试的首个实现 BEMEval- Doc2Schema 将这一框架应用于评估语言模型将住宅建筑数据映射为结构化模式表示的表现。结果表明，当模式相对简单且提供示例指导时，模型能够完成部分任务；但随着模式复杂度增加，其性能会下降。这些发现凸显了系统化基准测试的重要性：它能够揭示当前的局限，并为开发更稳健、可靠的 AI 辅助建模工作流程提供指导。

BEMEval有望成为连接人工智能与建筑模拟领域的桥梁。通过将建模任务设定为可衡量、可复现的挑战，它鼓励各方共同参与，并促进透明协作的文化。该基准测试旨在为研究人员、从业者和软件开发者之间的持续合作奠定基础，使人们能够在建模标准和人工智能技术不断发展的过程中跟踪进展。

展望未来，本研究的长期目标是为建筑能耗建模构建稳健且可扩展的开源评估生态系统。随着基础模型和领域数据集不断发展，BEMEval可以逐步成为社区资源，用于跟踪进展并推动人工智能更深入地融入建筑建模实践，从而在未来实现更高效、透明且易于使用的工作流程。

参考文献

BPI（建筑性能研究所），2013年。HPXML：住宅性能可扩展标记语言（数据字典与传输标准）。参见：https://www.hpxmlonline.com/overview/（访问日期：2025年10月22日）。Christensen, C., Horowitz, S., Polly, B., and Andrews, R. 2016年。使用NREL iUnit和NZERTF测试设施对建筑能耗模拟进行实证验证。美国国家可再生能源实验室（NREL）技术报告 NREL/TP- 5500-65924。Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L. 2009年。ImageNet：大规模分层图像数据库。IEEE计算机视觉与模式识别会议（CVPR）论文集，美国佛罗里达州迈阿密，248–255。DOE（美国能源部），2024年。使用FRP iUnit和NZERTF对能耗模拟进行实证验证。参见：
h
ttps://www.energy.gov/eere/buildings/empirical-validation-energy-simulation-frp-iunit-and-nzertf（访问日期：2025年10月22日）。Gang, W., Zhao, J., Chen, X. 和 Li, N. 2024年。使用大语言模型自动进行建筑信息提取与能耗模型生成。《能源与建筑》，325，113611。Hogeling, J.；Janssens, A.；等。2019年。关于EPB系列标准应用的IEA/EBC建筑能效规范工作组报告。IEA EBC附件58工作文件，2019年11月。HPXML 指南 2023。HPXML实施指南与工具箱。网址：
h
t
t
p
s://
hpxml-guide.readthedocs.io/en/latest/overview.html（访问日期：2025年10月22日）。国际能源署（IEA）。2023年。建筑——能源系统。网址：https://www.iea.org/energy-system/buildings（访问日期：2025年10月21日）。Jaccard, P. 1912年。高山带植物区系的分布，《新植物学家》，11(2)，37–50。Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y., Madotto, A., Fung, P. 2023年。自然语言生成中的幻觉综述。《ACM计算调查》，55(12)，1–38。ISO 13790:2008。建筑能效——空间供暖与制冷能耗的计算。国际标准化组织。Jensen, P.；Smith, R.；Thompson, K. 2024年。将住宅数据标准扩展至多户住宅和非住宅领域。《建筑性能标准期刊》，12(1)，45-59。Lee, S. H.；Zhao, F.；Augenbroe, G. 2013年。规范性能耗计算在建筑性能评级系统之外的应用：《建筑性能模拟期刊》，6(4)，282-292。DOI:10.1080/19401493.2012.720712。Levenshtein, V.I. 1966年。能够纠正删除、插入和逆序错误的二进制码，《苏联物理学文摘》，10(8)，707–710。Li, J., Zhang, S., Hu, X. 和 Wang, Y. 2025年。DrafterBench：面向土木工程文档理解的大语言模型基准测试。《建筑自动化》，164，105942。Long, N.；Fleming, K.；CaraDonna, C.；Mosiman, C. 2021年。BuildingSync：用于商业建筑能耗审计数据交换的模式。《建成环境发展》，7，100054。

Merket, N. 2014年。NREL助力住宅能效公司实现高效数据交换。NREL/FS-5500-61007。美国国家可再生能源实验室。美国国家标准与技术研究院（NIST）。2014年。净零能耗住宅测试设施（NZERTF）。网址：https://www.nist.gov/laboratories/tools-instruments/net-zero-energy-residential-test-facility-nzertf（访问日期：2025年10月22日）。PSD Consulting。2014年。用于住宅能源评级系统验证的HERS BestEST和L100测试案例。网址：https://psdconsulting.com/software/treat/hers-bestest-doe-approved/ 和https://github.com/NREL/OpenStudio-HPXML/tree/master/workflow/tests/ASHRAE_Standard_140（访问日期：2025年10月22日）。RapidFuzz文档。2025年。RapidFuzz：Python中的模糊字符串匹配。网址：https://maxbachmann.github.io/RapidFuzz/（访问日期：2025年10月21日）。Reimers, N., Gurevych, I. 2019年。Sentence-BERT：使用孪生BERT网络的句子嵌入。2019 年自然语言处理实证方法会议（EMNLP）论文集，3982–3992。Salton, G., McGill, M.J. 1983年。《现代信息检索导论》，麦格劳-希尔，纽约州纽约市。Schick, T., Dwivedi-Yu, J., Sun, S., Lewis, M. 和 Zettlemoyer, L. 2023 年。Toolformer：语言模型能够自学使用工具。arXiv 预印本 arXiv:2302.04761。Sørensen, T. 1948年。基于物种组成相似性建立等幅度群组的方法及其在丹麦公地植被分析中的应用，《Biologiske Skrifter》，5，1–34。TextDistance文档。2025年。TextDistance：使用Python计算序列间距离。网址：https://github.com/life4/textdistance（访问日期：2025年10月21日）。Wang, X., Wei, J., Schuurmans, D., Bosma, M., Chi, E. H., Le, Q. V. 和 Zhou, D. 2022年。自洽性提升语言模型中的思维链推理。arXiv预印本 arXiv:2203.11171。Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E. H., Le, Q. V. 和 Zhou, D. 2022年。思维链提示引发大语言模型。神经信息处理系统进展（NeurIPS 2022）。Winkler, W.E. 1990. Fellegi–Sunter 记录链接模型中的字符串比较器指标与增强决策规则，《美国统计学会调查研究方法分会论文集》，354–359。Xiang, X., Ma, M., Ma, X. 2022. 利用大语言模型推进建筑能耗建模。《应用能源》，309，118417。可通过以下网址获取：https://doi.org/10.1016/j.apenergy.2021.118417（访问于2025年10月21日）。Wu, Y., Chen, Y., Liu, Z., and Zhang, H. 2025. Arch- Eval：评估大语言模型的建筑知识理解能力。《建筑性能模拟期刊》，即将发表。Zhang, T., Feng, C., Hong, T., and Chen, Y. 2024. 利用大语言模型和基于规则的启发式方法实现自动化 EnergyPlus 模型生成。《能源与人工智能》，18，100368。Zhang, T., Zhao, J., Saleh, M., Liu, P. 2020. PEGASUS：基于提取间隔句的预训练，用于抽象式摘要。第37届国际机器学习大会（ICML）论文集，11328–11339。