评估统计语言模型作为语用推理器 rs

Benjamin Lipkin[1] **Lionel Wong**[1] **Gabriel Grand**[2] **Joshua B. Tenenbaum**[1,2]
BCS，麻省理工学院 [2]计算机科学与人工智能实验室，麻省理工学院{lipkinb, zyzzyva, gg, jbt}@mit.edu

arXiv:2305.01020v1 [cs.CL] 1 May 2023

摘要

交际语言与意图之间的关系通常具有概率性，并且对语境十分敏感。人们尝试采用多种策略来估计这种映射，其中往往会利用交际的递归贝叶斯模型。与此同时，大语言模型（LLM）日益被应用于语义解析任务，负责从自然语言中推断逻辑表示。尽管现有的大语言模型研究大多局限于字面语言的使用，但在本研究中，我们评估了大语言模型推断语用话语含义的能力。具体而言，我们探讨了对可分级形容词“*strong*”进行阈值估计的情形：在力量先验的语境条件下，进一步扩展到与限定、否定、极性反转和类别比较的组合。我们发现，大语言模型能够针对若干复杂的语用话语，基于语境推导出类似人类的解释分布，但在与否定进行组合时存在困难。这些结果揭示了统计语言模型的推理能力，以及它们在语用和语义解析应用中的使用价值。所有相关代码均已公开发布[1]。

关键词： 语言模型；语义解析；语用学

引言

自然语言理解是在语境中展开的，反映的不仅仅是字面解释。人们认为，这一过程由一系列推断所介导，这些推断共同审视语言结构与心理表征之间的映射，并同时考量由此产生的解释是否合理。像*“Mia is tall”* 这样简单的句子本身可能含义宽泛，但听者认为合理的身高范围会随语境而变化，例如*“Miaplays in the WNBA”* 或*“Mia isa three-year old child.”*这类语境推断广泛见于语言语用学研究（Wittgenstein，1953；Searle，1969；Austin，1975年；Levinson，1983；Grice，1989；Clark，1996）。

近年来，关于大规模训练 Transformer 语言模型的研究，催生了各类工程成果；这些成果在一系列自然语言处理（NLP）基准测试集上表现出色。尽管这类系统的训练目标明确是优化下一个词元预测，但它们隐式地再现了传统 NLP 流程中的大量环节，从词性标注和句法分析，到语义角色标注和共指消解（Tenney、Das 和 Pavlick，2019年；Bo mmasani 等，2021年）。事实上，越来越多当代研究利用大语言模型，从自然语言（NL）输入中合成类似程序的表征，以用于下游应用，范围涵盖行动规划到定理求解（Acquaviva 等，2021年；Gao 等，2022年；Collins、Wong、Feng、Wei 和 Tenenbaum，2022年；Mishra 等，2022年；Zelikman、Huang、Poesia、Goodman 和 Haber，2022年；Wong 等，即将发表）。这类研究将大语言模型用作语义解析器，将其视为语言形式与意义表示之间映射关系的形式化描述。然而，这类评估大多局限于字面语言的使用与翻译。相比之下，语用意义估计通常需要结合语境，考虑多个解释构成的分布，因此更为复杂（Fried、Tomlin、Hu、Patel 和 Nematzadeh，2022年；Hu、Floyd、Jouravlev、Fedorenko 和 Gibson，2022年；Ruis 等，2022年；Hu、Levy、Degen 和 Schuster，2023年）。

现有的语用推理模型通常依赖显式的概率计算，往往采用理性言语行为 （RSA）交际框架。在这一框架中，语用听者会对信息型说话者进行推理，以推断其意图表达的含义（Frank & Goodman，2012；Goodman & Stuhlmüller，2013；Goodman & Frank，2016）。我们提出的问题是：统计语言模型能否摊销 常见的语用推断，从而恢复语言与受语境调节的含义之间近似等价的分布？

为回答这一问题，本文探讨了在虚构游戏中描述玩家时，对可分级形容词 *“strong”* 的解释。给定描述可能世界上的生成模型的上下文（该模型除其他变量外，还对*“strength”*设定了数值先验），我们的范式要求对关于新玩家力量的文本描述进行数值解释估计。我们收集了由大语言模型估计以及由人类测量得到的此类话语解释分布，并探索了与其他复杂性维度的组合。我们发现，大语言模型能够出色地推断复杂语用话语的、考虑上下文且类似人类的分布，例如*“very strong fora beginner”*。与此同时，我们观察到，这些模型难以将所推断的意义与否定（例如*“not strong”*）或极性反转（例如*“weak”*）进行组合，这为其潜在缺陷提供了启示。

作为概率程序的意义

在表达语言意义的形式化表征时，一种方法是从模型

[1]https://github.com/benlipkin/probsem/tree/CogSci2023

论语义学框架（Kripke，1963；Montague，1973；Partee、ter Meulen 和 Wall，1990；Kratzer 和 Irene，1998）出发，并结合不确定性量化（Van Eijck 和 Lappin，2012；Cooper、Dobnik、Lappin 和 Larsson，2015），最终将概率编程语言（PPL），例如 Church（Goodman、Mansinghka、Roy、Bonawitz 和 Tenenbaum，2012），作为一种实用的基础。Goodman 和 Lassiter（2015）尤其提出了一个框架——本文正是在此基础上展开——将自然语言（NL）视为对概率程序进行信念更新。从描述某个领域的可能世界生成模型出发，句子被逐步表达为条件化语句，并通过执行这些语句来更新对世界状态的后验信念。

Goodman 和 Lassiter 通过讨论一个拔河（ToW）虚构游戏的例子，阐述了这一框架。在这个经典儿童游戏的简化版本中，两支队伍相互竞争，每队有一名或多名玩家，最终由队员施加力量最大的一队获胜（Goodman & Tenenbaum, 2010；Gerstenberg & Goodman, 2012；Goodman, Tenenbaum, & Gerstenberg, 2014）。在此基础上，Goodman 和 Lassiter 构建了由概率编程语言（PPL）支持的语境语义分析示例。例如，*"Team Ahasmore than 3players"* 可以表示为 (condition (>(l ength team-a) 3))；当问题是*"Team A"* 是否可能战胜*"Team B"*（后者或许只有 2 名玩家）时，评估这场比赛的结果分布就会考虑这一信息。为了将这种方法拓展到超越字面语言使用的情境，即自然语言（NL）呈现出非确定性解释的场景，Goodman 和 Lassiter 提议借助 RSA 进行显式概率计算。这一框架的难点之一，是需要手动编写能够表达所评估自然语言语义的程序。借鉴语义解析和程序合成领域的成功方法，这一流程越来越适合借助大语言模型实现自动化。

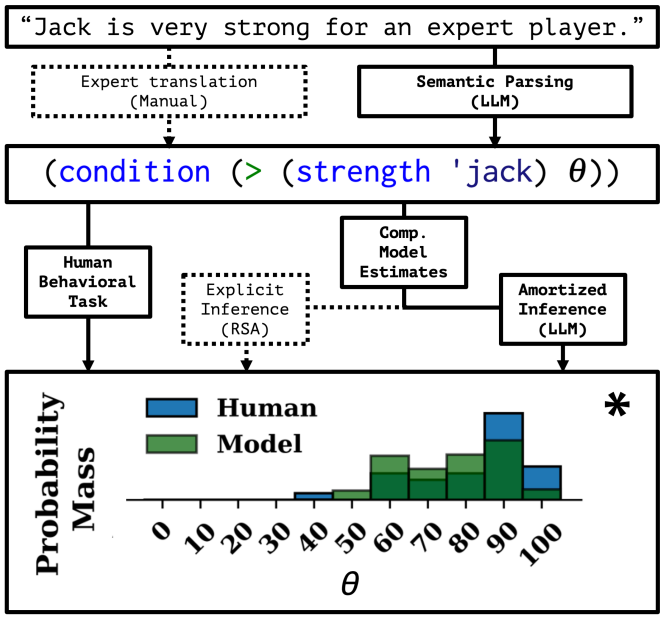


图1：示意图概览。大语言模型取代传统的语用学流程，通常能够针对复杂结构的多种解释得出类似人类的估计。

本研究

Goodman 和 Lassiter（2015年）强调了概率程序语言在表达句子意义的逻辑表征方面具有出色的能力，但并未说明这类程序最初是如何推导出来的。与此同时，现代语义解析研究将大语言模型描绘为能够实现此类转换的系统。然而，当任务超越字面语言的使用时，问题便不明朗了：a）大语言模型是否适合实现此类复杂推断；b）模型的估计是否符合人类的预期。为回答这些问题，我们以拔河领域模型为基础，并将可分级形容词作为一种富有表现力的测试平台。

可分级形容词，例如"强"，由于缺乏精确的类别边界，具有模糊性。人们已提出多种方法来表达可分级形容词的语义；一种常见方法是引入一个自由阈值变量，将"强"定义为具有"力量" $>\theta$ (Cresswell, 1976; Klein, 1980; Kennedy, 2007; Lassiter & Goodman, 2017; Tessler, Tsvilodub, Snedeker, & Levy, 2020)。尽管可以通过 RSA 的递归概率推断，在不同程度上推导出关于 θ的分布或其他表述（Qing & Franke, 2014; Tessler & Goodman, 2022），但我们在此探讨，大语言模型能否直接通过单次前向传播估计关于 θ的分布。图 1 概述了相关内容。

在拔河玩家的语境中，假设领域描述中对"力量"设定了先验，我们先对"强"及其极性相反的对应词"弱"进行基本评估，然后扩展到否定的情形，例如"不强"；限定语的情形，例如"相当强"；以及比较类别的情形，例如"对新手玩家来说很强"。为了考察大语言模型推断的合理性，我们收集了人类参与者针对相同刺激材料的行为规范数据，以量化模型在何处、以何种方式捕捉或未能体现人类直觉。积极的一面是，我们发现大语言模型能够对一系列包含字面推断和语用推断的推断进行相当复杂的上下文摊销，并优雅地解析复杂的语用表达；这一过程以表达生成式世界模型的文本为条件，而该世界模型被表示为概率程序。然而，消极的一面是，大语言模型可能难以处理逻辑上更简单的否定或极性反转属性，在这些情况下偏离人类的解释。这些结果有助于我们理解大语言模型的推断能力，也因此为围绕统计语言学习器能力的争论提供参考（例如参见 Piantadosi（2023））。

方法

为探讨迄今提出的问题，我们首先以更正式的方式在 Church 中定义拔河领域模型，概述本研究后续所探讨语义的先验与约束。我们定义了一个评分函数，大语言模型可据此在给定领域模型语境和自然语言查询的条件下，估计特定文本解释的概率。为评估该评分函数的有效性，我们设计了一组测试材料，用于评估人类和大语言模型对可分级形容词的解释；并让我们的建模框架和 60 名人类参与者完成这项任务的两种变体，其中一种主要关注限定语，另一种关注类别比较。否定和极性反转也作为限定语实验的一部分进行了探讨。最后，我们将模型估计的解释分布与在人类参与者中实测的分布进行比较。

领域模型与大语言模型上下文

```
;; This Church program models a tug-of-war game between teams of players.

;; Each player has a strength, with strength value 50 being about average.
(define strength (mem (lambda (player) (gaussian 50 20))))

;; Each player has an intrinsic laziness frequency.
(define laziness (mem (lambda (player) (uniform 0 1))))

;; The strength of the team is the sum of the player strengths.
;; When a player is lazy in a match, they pull with half their strength.
(define (team-strength team)
    (sum (map
        (lambda (player)
            (if (flip (laziness player))
                (/ (strength player) 2) (strength player)))
        team)))

;; The winner of the match is the stronger team.
;; Returns true if team-1 won against team-2, else false.
(define (won-against team-1 team-2)
    (> (team-strength team-1) (team-strength team-2)))

;; Now, let us translate some user-defined statements.
;; Each statement begins with either `Condition` or `Query`.
;; `Condition` statements provide facts about the scenario.
;; `Query` statements are questions that evaluate quantities of interest.

;; Condition: Jack is strong.
(condition (> (strength 'jack) 50))
```

图2：针对单个查询传递给大语言模型的完整文本示例。拔河领域模型（蓝色）和任务指令（绿色）在所有试验中保持一致。对于每个待评估的句子（黄色），都会评估每个程序（红色）的概率，以返回给定解释的分数。

作为可分级形容词实验的背景，我们考虑前文介绍过的拔河领域。图2展示了该领域作为 Church 程序的精确定义。给定描述该领域的上下文、当前任务的说明（将自然语言翻译为 Church），以及一个特定的自然语言查询，大语言模型便可以作为程序表达式上的生成模型，既能从提示词开始采样后续词元，也能根据模型为预定义的程序赋予概率。

至关重要的是，我们看到力量的先验分布 $\sim \mathcal{N}(50,20)$。虽然该领域模型的其他元素并非我们的下游任务所必需，但我们仍纳入完整上下文，以便在完整的世界模型中评估有效性和稳健性。

任务描述

为了将我们的世界模型设定为*“Jack 很强壮”*，其表达形式为（condition (> (strength 'jack) θ)），θ 取什么值才合理？尽管采用 RSA 是一种策略，但要准确估计所有可分级形容词的此类范围，很快就会变得难以处理；如果再考虑与其他约束的可能组合，例如“有些强壮”，组合空间的问题会更加严重。因此我们要问：大语言模型能否以一种具有语用敏感性、并且与人类推断一致的方式，对 θ上的这一分布进行摊销推断？为了评估这一问题，我们设计了一组刺激材料，每项材料都使用可分级形容词来描述一位名为*“Jack”*的虚构运动员的力量。这些材料被分配到两个实验中。

E1：限定语在 E1 中，我们首先评估了有关 Jack 力量的句子被解释为以下形式程序的概率：(condition (> (strength 'jack) θ))，其中θ的取值范围为 $0-100$，步长为 10。这些句子既包括 Jack 在不同程度上“强壮”的情形，也包括“不弱”的情形。对于每个句子，大语言模型估计了$P_{model}(\theta)$，而$P_{human}(\theta)$则根据一组人类参与者给出的点估计测得。除这些测试句外，我们还加入了一个形式如下的控制句：*“Jack 的力量至少达到平均水平”*。该句没有模糊性，旨在恢复位于 $\theta = \mu_{strength} = 50$ 处的大部分概率质量。随后，为检验对极性反转的稳健性，我们设计了一组平行材料来评估 Jack 的弱点，改为考察以下形式的程序：(condition (< (strength 'jack) θ))。这些材料与 E1 第一部分中的材料直接对应，仅将“强壮”和“不弱”分别替换为“不强壮”和“弱”。完整的 18 项材料见图3。

E2：比较类在 E2 中，我们引入比较类来有条件地细化解释，从而扩展了这项评估。我们修改了大语言模型提示词中力量的定义，通过注入以下条件语句，将玩家所属的联赛作为一个新变量纳入考量（完整的更新后提示词可在论文仓库中找到）：（条件

((equal? league '初学者) (gaussian 30 20)) ((equal? league '中级者) (gaussian 50 20)) ((equal? league '专业选手) (gaussian 70 20))) 在这里，我们测试大语言模型能否利用对玩家的文字描述，同时推断其所属联赛以及在该联赛中的相对

力量。我们从 E1 的一个子集中取材，开发了一组融入这些比较类别的新材料。具体而言，我们保留了控制形式：*"Jack* 的力量至少达到平均水平" 以及图3中偏离均值最多的形式：*"Jack* 非常强壮"。我们针对每个联赛，按三个抽象程度修改每个句子：完全匹配、同义表达和暗指。例如，对于第一个联赛，我们分别评估 Jack 对"初学者"、"新手"和"刚接触这项运动的人"而言的力量。完整的 18 项材料见图4。

人类参与者评估

为了评估每个刺激材料中的 $P_{human}(\theta)$，研究开展了两项行为实验。从 Prolific 招募了 60 名参与者，E1 和 E2 各 30 名。参与者均提供了知情同意，并获得了约每小时 15 美元的报酬。实验要求参与者根据独立阅读刺激句子后的判断，移动滑块以标示一名名叫 *"Jack"* 的虚构运动员在力量方面的阈值（θ）。E1 中有一名参与者自述存在理解困难，因此被剔除。分析仅纳入其余参与者。实验源文件（包括说明和刺激材料）已随论文代码库一并发布。

大语言模型评分函数

为了评估每个刺激材料中的 $P_{model}(\theta)$，研究定义了一个针对 θ 不同取值的程序评分函数。研究使用 OpenAI 的 code-davinci-002 大语言模型（Chen等人，2021）对语言模型进行参数化，使其能够为任意字符串 $x_i \in \mathcal{X}$ 分配条件概率。为了将每个程序 $y_i \in \mathcal{Y}$ 的得分解释为相对于所考察的受限假设空间的归一化概率，研究将大语言模型对所考察程序给出的对数概率输入 softmax函数，并使用温度参数 α。该参数针对每个刺激句分别通过留一交叉验证（LOOCV）独立选定，具体方法将在下一节详述。

$$P(y_i) = \frac{\exp(\alpha \log P(x_i))}{\sum_{j=1}^{n} \exp(\alpha \log P(x_j))} \tag{1}$$

在程序仅在 θ 上存在差异的情况下，$P_{model}(\theta_i)$ 近似为 $P(y_i)$。这些离散的程序概率构成了后续分析的基础。

比较 $P_{human}(\theta)$ 和 $P_{model}(\theta)$

对于 36 个刺激句中的每一个，研究人员测量了人类参与者在 θ 上的 29（E1）或 30（E2）个点估计值。根据这些点估计值，通过对每个刺激进行计数并归一化，在定义域 $0-100$ 上以 10 为间隔计算出一个离散经验分布。

$$P_{human}(\theta_i) = \frac{C(\theta_i)}{\sum_{j=1}^{n} C(\theta_j)} \tag{2}$$

对于相同的刺激句子，在同一领域内为每个程序计算一个权重。这些权重按照方程 1 进行归一化，并通过最小化每项实验中其余 $N-1$ 个刺激材料对应的 $P_{human}(\theta)$ 与 $P_{model}(\theta)$ 之间的 Jensen-Shannon距离（JSD；方程 4）之和，为每个刺激材料选取 α。优化采用 Nelder-Mead下山单纯形法（Nelder和Mead，1965）。

$$\arg\min_{\alpha} JSD\left(P_{human}(\theta), P_{model}(\theta)\right) \tag{3}$$

定义 $P_{human}(\theta)$ 和 $P_{model}(\theta)$ 后，使用 Jensen-Shannon距离计算二者的相似度。Jensen-Shannon距离是两个概率分布 P 和 Q之间的度量距离，其中M 是 P 和 Q逐点取平均得到的分布，而 KL 是 Kullback-Leibler 散度（Lin，1991）。

$$JSD(P \parallel Q) = \sqrt{\frac{KL(P \parallel M) + KL(Q \parallel M)}{2}} \tag{4}$$

为评估这一相似度指标的统计显著性，研究采用了非参数置换检验。为生成零分布，将$P_{human}(\theta)$ 和$P_{model}(\theta)$ 的值在 θ 上随机打乱 $N = 10{,}000$ 次迭代，并对每种变体测量 JSD。通过将小于真实 JSD 的零样本数量除以N，计算p值。使用错误发现率（FDR）校正对原始p值进行多重比较校正，校正范围为每项实验中的检验数量（Benjamini & Hochberg，1995）。

结果

为评估大语言模型能否有效利用上下文，准确推断语言意义上的分布，我们开展了多项实验。

E1：限定语

针对对 Jack 力量的描述，我们在 θ 上评估了形式为 (condition (> (strength 'jack) θ)) 的程序。$P_{model}(\theta)$ 在图 3A中以绿色呈现。

大语言模型能够结合语境，并对程度形容词和限定语进行符合语用的推断。 对于每种变化，θ 上呈现的分布在定性上平滑且易于解释。对于*"Jack* 很强"，大部分概率质量落在 $> \mu^{strength}$；而当*"Jack* 非常强"时，分布会进一步偏移。对于对照句*"Jack* 的力量至少达到平均水平"，概率质量则正确地落在 $\theta = \mu^{strength}$。

大语言模型的推断大多符合人类的判断，但在处理否定时仍有困难。 在同一张图3A中，我们看到$P_{human}(\theta)$以蓝色呈现。值得注意的是，$P_{human}(\theta)$ 和 $P_{model}(\theta)$ 通常高度重叠，即使限定语的组合较为复杂也常常如此。事实上，对于所有不含否定的句子，这些分布都表现出显著相似性（图 3A）。然而，在包含否定的句子中，只有一半的解释与之吻合。

大语言模型在处理极性反转时更加吃力。为了进一步评估这一框架的稳健性，我们开展了一项后续实验，探究概念极性的反转。针对一组描述 Jack 的

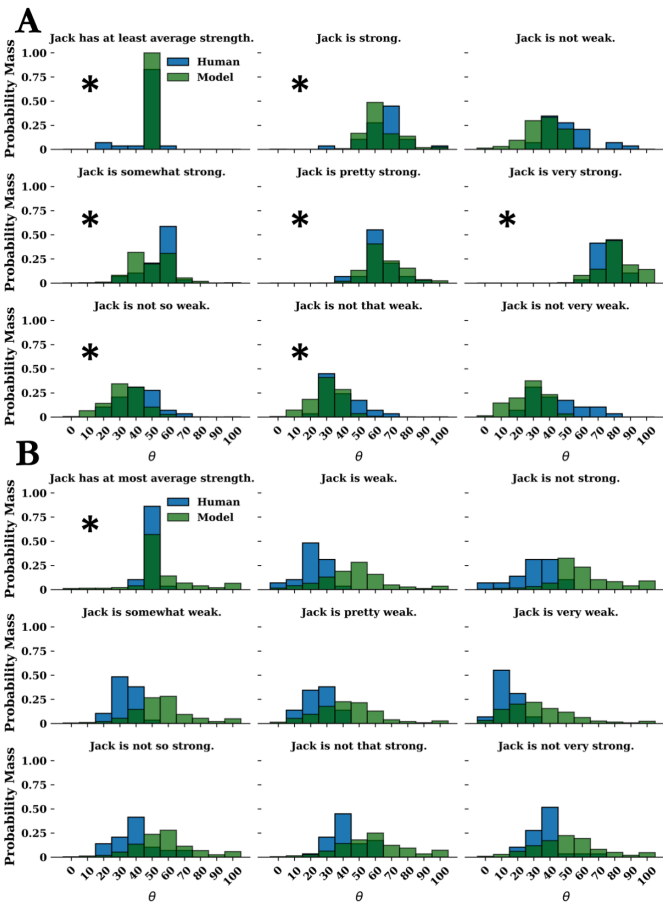


图3：模型估计和人类测得的$P(\theta)$分布。面板A考察以下形式的程序：(condition (> (strength 'jack) θ))，面板B则考察：(condition (< (strength 'jack) θ))。每个子图对应一个独特的句子，其中$P_{model}(\theta)$以绿色显示，$P_{human}(\theta)$以蓝色显示。星号表示$P_{model}(\theta)$与$P_{human}(\theta)$之间存在显著相似性（$p < 0.05$；经错误发现率校正），其定义为相较于零假设置换分析，Jensen-Shannon距离（JSD；公式4）有所降低。

弱点的句子，我们在 θ 的范围内评估了形式为 (condition (< (strength 'jack) θ)) 的程序。$P_{model}(\theta)$ 在图3B中以绿色呈现。同样，各分布在定性上看起来平滑，并呈现出一些符合直觉的特征。例如，*"Jack* 非常弱*"*的数值低于*"Jack* 很弱*"*，而控制句*"Jack* 的力量至多为平均水平。*"*中的均值也被正确解析。然而，与人类参与者的判断一致性呈现出不同趋势。在本例中，当所评估概念相对于提示词中呈现的变量具有负极性时， θ 往往会被模型持续高估。除控制句外，对于所有句子，模型都无法检测出$P_{model}(\theta)$ 与$P_{human}(\theta)$ 之间存在显著相似性。

E2：比较类

选择控制条件"至少平均"，而与$\mu^{strength}$ 偏差最大的条件是图3A中的"非常强壮"。随后编拟了一组新句子，用于描述 *"Jack"* 的力量如何取决于他所属的不同"联赛"，并为每个联赛设定了各自的力量先验。提示词明确列出了"初学者"、"中级者"和"职业"联赛，其平均力量分别为30、50和70。

大语言模型能够准确解析条件混合，甚至能从间接描述中推断群体成员身份。我们为每种力量描述和每个联赛都提供了*"Jack……*属于*……*"形式的句子，包括提示词中明确描述的联赛（图4A），以及此前未见过、作为同义词的联赛描述（图4B），甚至是间接指涉（图4C）。模型对这些句子的解析和解读都取得了出色的效果：在评估的18个句子中，有17个与人类的预期显著一致，其中包括所有控制句，以及所有复杂程度属于直接匹配或同义词的句子。

讨论

我们以将语用语言理解视为一种推断过程的框架开启了这项研究，随后提出将语言意义表征视为概率程序。我们选择可分级形容词作为测试对象，设计了一项任务，以评估大语言模型在复杂语义解析练习中的语用推理能力。我们以表达一个生成式世界模型的代码为背景，该模型定义了一场拔河比赛的语义；我们评估了若干描述虚构玩家力量的句子，其中许多句子结合了语用上复杂的现象。我们使用大语言模型，为每个目标句子估算$P_{model}(\theta)$，并开展人类行为实验，以实证测量与之对应的$P_{human}(\theta)$。

根据我们的初步评估（E1；图3），我们发现，大语言模型能够有效地摊销对 θ 上平滑分布的推断，同时使其以提示词的语义为语境基础，并对可分级形容词和限定语表现出语用敏感性。对于描述玩家有多"强"的所有情况，此类模型估计都与人类测量结果一致；但在玩家"弱"、"不弱"或"不强"的大多数情况下，模型未能再现人类分布的复杂细节。这些结果表明，尽管模型能够为上述每种情况估计出某种近似分布，但在词汇空间中组合否定（例如"强"与"不强"），或在概念空间中进行极性反转（例如"强"与"弱"）时，模型推断出与人类完全相似的分布的能力会下降。这与先前关于大语言模型普遍难以处理否定的研究结果一致（Kassner & Schütze，2019年；Hosseini 等，2021年；Cres well、Shanahan 和 Higgins，2022年）。这也与关于概念习得的儿童发展研究形成了耐人寻味的关联；该研究指出，相较于对应的正极性概念，儿童掌握负极性概念（例如"矮"）存在可观察到的滞后，方娱乐网址。

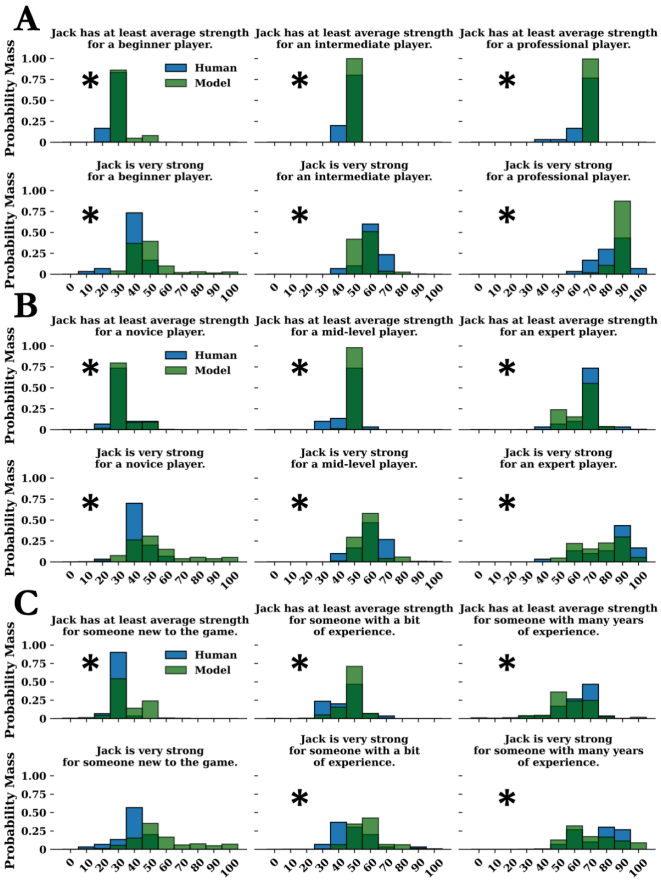


图4：结合比较类别后，模型估计和人类测量的 $P(\theta)$ 分布。面板A使用提示词中的精确类别，面板B使用同义词，面板C使用间接指涉。与图3一样，$P_{model}(\theta)$ 以绿色表示，$P_{human}(\theta)$ 以蓝色表示；星号表示 $P_{model}(\theta)$ 与 $P_{human}(\theta)$ 之间具有显著的（p <0.05；错误发现率）相似性。

部分，例如“高”，或许突显了概念复杂性中更普遍的不对称性（Klatzky、Clark 和 Macken，1973年；Brewer 和 Stone，1975 年；Barner 和 Snedeker，2008 年）。通过对类别比较的评估（E2；图4），我们进一步凸显了此类模型的上下文敏感性：它们能够恰当地解析条件混合，即使上下文附近存在错误指称，也表现出令人印象深刻的鲁棒性。当查询与上下文变量之间并非完全匹配，而是需要根据同义词或间接暗示进行估计时，这些结果更具说服力。这些结果支持了这样一种观点：在此方法下，大语言模型具有词汇语义鲁棒性。与某些基于组合范畴语法（CCG）的传统语义解析器相比，这是一种便捷的情形；后者往往需要更复杂的变通方法（Artzi、Das 和 Petrov，2014年；Kwiatkowski、Zettlemoyer、Goldwater 和 Steedman，2011年；Steedman，2001年）。

总体而言，这些结果描绘出这样一幅图景：大语言模型在这些复杂测试案例中，能够有效地恢复某种合理的分布，但也凸显出其与人类推断之间的一些差异。如果我们能够完美地恢复人类的分布，就可能引出一系列解释。对此类发现的一种解释是：正如大语言模型似乎会隐式表征其他形式的语言结构一样，它们在此处也会隐式地进行推断，这一点与其他关于摊销的研究所提出的观点相呼应（White、Mu 和 Goodman，2020年；Wu 和 Goodman，2022年）。另一种解释是，在实践中，训练期间文本中的统计规律足以让模型在测试时恢复这些分布，而无需显式地基于世界模型进行计算。这种观点或许能为人类语言处理的资源理性框架提供启示，可能意味着，原则上，部分语用计算可以通过启发式方法近似，甚至直接提取，而不必在每个实例中都重新显式计算（Gershman 和 Goodman，2014年；Gershman、Horvitz 和 Tenenbaum，2015年；Dasgupta 和 Gershman，2021年）。尽管我们的数据并未显示大语言模型在所有案例中都能完美估计人群分布，但我们认为，这些数据至少在一定程度上支持第二种假设。确实，解决我们任务所需的计算中，可能有些可以摊销，但并非全部，因此模型在某些情况下能够产生类似人类的分布，却会在其他分布外案例中给出不正确的近似。例如，否定与其他成分的组合或许要求人类参与者在测试时进行更显式的计算，从而使其分布相对于大语言模型的启发式估计发生偏移。未来的研究应从计算效用框架出发，更直接地检验这一点。

局限性 尽管本文所呈现的结果描绘了一幅总体积极的图景，表明大语言模型能够在复杂的语义解析任务中优雅地处理语用推理，但迄今为止，我们仅探讨了单一范围内的少量示例。为了确认这些结果的结论具有普适性，还需要在更多任务背景下，对更广泛的语用现象进行评估。

未来方向 一个特别令人振奋的未来方向，是将大语言模型对概率程序语言（PPL）程序所进行的推断，与这些程序的实际执行及其生成分布的评估联系起来。如果我们提出这样的问题：“一个实力很强的初学者吉尔能打败实力还不错的中级玩家简吗？”，就可以将其转化为神经符号编程问题。借助大语言模型推理，可以推导出关于每位玩家实力阈值的分布。接下来，可以在 PPL 解释器中显式执行这些程序，从而得到每位玩家实力的分布。在此基础上，很容易查询这场比赛的获胜者：（query (won-against ’(jill) ’(jane)））。如果进一步考虑更难的情况，例如涉及否定的情况，则可以考虑结合类似 RSA 的方法与大语言模型辅助的方法。例如，使用大语言模型的估计值初始化序贯蒙特卡罗（SMC）假设，再根据概率程序推断对其进行更新。

致谢

我们感谢匿名审稿人提出的宝贵反馈和建议。BL 获得了 MIT 总统奖学金的支持，GG 则获得了美国国家科学基金会研究生研究奖学金的支持（资助编号：2141064）。LW 和 JBT 获得了 MIT 智能探索计划、美国空军科学研究办公室资助（编号：FA9550-19-1-0269）、MIT-IBM Watson AI Lab、美国海军研究办公室人工智能科学项目以及 DARPA Machine Common Sense 的支持。

参考文献

Acquaviva, S., Pu, Y., Kryven, M., Sechopoulos, T., Wong, C., Ecanow, G. E., . . . Tenenbaum, J. B.（2021年）。向人类和机器传达自然程序。*arXiv*预印本 *arXiv:2106.07824*。Artzi, Y., Das, D., & Petrov, S.（2014年）。学习用于组合范畴语法语义解析的紧凑词汇表。Austin, J. L.（1975年）。《如何以言行事》。牛津大学出版社。Barner, D., & Snedeker, J.（2008年）。形容词习得中的组合性与统计规律：4岁儿童根据新名词指称对象的大小分布来理解 tall 和 short。《儿童发展》,*79*(3), 594–608。Benjamini, Y., & Hochberg, Y.（1995年）。控制假发现率：一种实用且有力的多重检验方法。《皇家统计学会杂志：*B*辑（方法论）》, *57*(1), 289–300。Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., . . . others（2021年）。基础模型的机遇与风险。*arXiv*预印本 *arXiv:2108.07258*。Brewer, W. F., & Stone, J. B.（1975年）。空间反义词对的习得。《实验儿童心理学杂志》,*19*(2), 299–307。Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H. P. d. O., Kaplan, J., . . . others（2021年）。评估基于代码训练的大语言模型。*arXiv*。Clark, H. H.（1996年）。《语言的使用》。剑桥大学出版社。Collins, K. M., Wong, C., Feng, J., Wei, M., & Tenenbaum, J. B.（2022年）。结构化、灵活且稳健：对大语言模型进行基准测试并加以改进，使其在分布外推理任务中表现得更像人类。*arXiv*预印本 *arXiv:2205.05718*。Cooper, R., Dobnik, S., Lappin, S., & Larsson, S.（2015年）。概率类型论与自然语言语义学。《语言技术中的语言学问题》,*10*。Cresswell, M. J.（1976年）。程度语义学。载于《蒙塔古语法》（第261–292页）。爱思唯尔。Creswell, A., Shanahan, M., & Higgins, I.（2022年）。选择-推断：利用大语言模型实现可解释逻辑推理。*arXiv*预印本 *arXiv:2205.09712*。Dasgupta, I., & Gershman, S. J. (2021). 记忆作为一种计算资源。认知科学趋势，*25*(3), 240–251. Frank, M., & Goodman, N. D. (2012). 预测语言博弈中的语用推理。《科学》，*336*(6084), 998–998. Fried, D., Tomlin, N., Hu, J., Patel, R., & Nematzadeh, A. (2022). 接地语言学习中的语用学：现象、任务与建模方法。*arXiv*预印本 *arXiv:2211.08371*. Gao, L., Madaan, A., Zhou, S., Alon, U., Liu, P., Yang, Y., . . . Neubig, G. (2022). Pal：程序辅助语言模型。*arXiv*. Gershman, S., & Goodman, N. (2014). 概率推理中的摊销推断。载于认知科学学会年会论文集（第36卷）。Gershman, S., Horvitz, E. J., & Tenenbaum, J. B. (2015). 计算理性：脑、心智与机器智能研究中逐渐汇聚的范式。《科学》，*349*(6245), 273–278. Gerstenberg, T., & Goodman, N. D. (2012). 教堂里的乒乓球：人类概率推断中概念的创造性运用。载于认知科学学会年会论文集（第34卷）。Goodman, N. D., & Frank, M. C. (2016). 将语用语言解释视为概率推断。认知科学趋势，*20*(11), 818–829. Goodman, N. D., & Lassiter, D. (2015). 概率语义学与语用学：语言与思维中的不确定性。当代语义理论手册（第二版）。*Wiley-Blackwell*. Goodman, N. D., Mansinghka, V., Roy, D., Bonawitz, K., & Tenenbaum, J. B. (2012). Church：一种生成模型语言。*arXiv*. Goodman, N. D., & Stuhlmüller, A. (2013). 知识与会话含义：将语言理解建模为社会认知。认知科学专题，*5*(1), 173–184. Goodman, N. D., & Tenenbaum, J. B. (2010).认知的概率模型（第一版）。http://v1.probmods.org/. Goodman, N. D., Tenenbaum, J. B., & Gerstenberg, T. (2014). 概率思维语言中的概念。大脑、心智与机器中心（*CBMM*）备忘录，*010*. Grice, P. (1989). 言语之道研究。哈佛大学出版社。Hosseini, A., Reddy, S., Bahdanau, D., Hjelm, R. D., Sordoni, A., & Courville, A. (2021). 通过理解“非”来理解：对语言模型中的否定进行建模。*arXiv*预印本 *arXiv:2105.03519*. Hu, J., Floyd, S., Jouravlev, O., Fedorenko, E., & Gibson, E. (2022). 人类与语言模型语用语言理解能力的细粒度比较。*arXiv*预印本 *arXiv:2212.06801*. Hu, J., Levy, R., Degen, J., & Schuster, S. (2023). 预期——

未说出的备选项能够预测语用推论。*arXiv*预印本 *arXiv:2304.04758*. Kassner, N., & Schütze, H. (2019). 面向预训练语言模型的否定与错误启动探测：鸟会说话，但不会飞。*arXiv*预印本 *arXiv:1911.03343*. Kennedy, C. (2007). 模糊性与语法：相对与绝对等级形容词的语义。语言学与哲学，*30*(1), 1–45. Klatzky, R. L., Clark, E. V., & Macken, M. (1973). 极性形容词习得中的不对称：语言层面还是概念层面？《实验儿童心理学杂志》，*16*(1), 32–46. Klein, E. (1980). 原级形容词与比较级形容词的语义学。语言学与哲学，*4*(1), 1–45. Kratzer, A., & Irene, H. (1998). 生成语法中的语义学（第1185卷）。Blackwell Oxford. Kripke, S. A. (1963). 正规模态命题演算中的模态逻辑语义分析 I。《数学逻辑季刊》，*9*(5-6), 67–96. Kwiatkowski, T., Zettlemoyer, L., Goldwater, S., & Steedman, M. (2011). 用于语义解析的组合范畴语法归纳中的词汇泛化。载于*2011*年自然语言经验方法会议论文集（第1512–1523页）。Lassiter, D., & Goodman, N. D. (2017). 贝叶斯解释模型中的形容词模糊性。*Synthese*，*194*(10), 3801–3836. Levinson, S. C. (1983). 语用学。剑桥大学出版社。Lin, J. (1991). 基于香农熵的散度度量。《*IEEE*信息论汇刊》，*37*(1), 145–151. Mishra, S., Finlayson, M., Lu, P., Tang, L., Welleck, S., Baral, C., . . . others (2022). Lila：数学推理统一基准测试集。*arXiv*预印本 *arXiv:2210.17517*. Montague, R. (1973). 普通英语中量化的恰当处理。载于《自然语言研究方法：*1970*年斯坦福语法与语义研讨会论文集》（第221–242页）。Nelder, J. A., & Mead, R. (1965). 一种用于函数最小化的单纯形法。《计算机杂志》，*7*(4), 308–313. Partee, B. B., ter Meulen, A., & Wall, R. (1990). 语言学中的数学方法（第30卷）。施普林格科学与商业媒体。Piantadosi, S. T. (2023). 现代语言模型驳斥了乔姆斯基的语言研究方法。*Lingbuzz*预印本, *ling-buzz/007180*. Qing, C., & Franke, M. (2014). 可分级形容词、模糊性与最优语言使用：一种说话者导向模型。载于语义学与语言学理论（第24卷，第23–41页）。Ruis, L., Khan, A., Biderman, S., Hooker, S., Rocktäschel, T., & Grefenstette, E. (2022). 大型语言模型并非零样本交流者。*arXiv*预印本 *arXiv:2210.14986*。Searle, J. R.（1969）。言语行为：语言哲学论文集 （第626卷）。剑桥大学出版社。Steedman, M.（2001）。句法过程。麻省理工学院出版社。Tenney, I., Das, D., & Pavlick, E.（2019）。BERT重新发现经典自然语言处理流水线。*arXiv*预印本 *arXiv:1905.05950*。Tessler, M. H., & Goodman, N. D. （ 2022 ）。暖意（适用于冬季）：推断交际中的比较类。《认知科学》，*46*(3), e13095。Tessler, M. H., Tsvilodub, P., Snedeker, J., & Levy, R. P.（2020）。信息目标、句子结构与比较类推断。载于认知科学学会年度会议论文集。Van Eijck, J., & Lappin, S.（2012）。自然语言的概率语义学。逻辑与交互理性（*LIRA*），*2*，17–35。White, J., Mu, J., & Goodman, N. D.（2020）。通过摊销语用推理，学会提供有信息量的指称。*arXiv*预印本 *arXiv:2006.00418*。Wittgenstein, L.（1953）。《哲学研究》。Wong, L., Grand, G., Lew, A., Andreas, J., Goodman, N. D., Mansinghka, V., & Tenenbaum, J. B.（待刊）。从自然语言转换为思维语言。预印本。Wu, M., & Goodman, N.（2022）。用于近似概率推断的基础后验。*arXiv*预印本 *arXiv:2205.09735*。Zelikman, E., Huang, Q., Poesia, G., Goodman, N. D., & Haber, N.（2022）。Parsel：用于算法推理的统一自然语言框架。*arXiv*预印本 *arXiv:2212.10561*。