大语言模型推理机制探索
——分析、比较与提升

Exploring the Inference Mechanism of Large Language Models: Analysis, Comparison and Enhancement

张立英 /ZHANG Liying

（1. 中国科学院大学人文学院，北京，100049；2. 中国科学院哲学研究所，北京，100190）
(1. School of Humanities, University of Chinese Academy of Sciences, Beijing, 100049;
2. Institute of Philosophy, Chinese Academy of Sciences, Beijing, 100190)

摘　要：本文首先结合一些测试结果展示并分析了大模型的弱推理表现；指出大模型作为语元关联度预测模型，其“推理”的本质是模式匹配。不过，大模型基于输入问题生成输出结果的过程仍可看作一种推理，尽管其与人类推理机制并不相同。人类的推理中暗含推理规则、抽象－替换机制、缺省前提、语义理解等，是自上而下的，大模型推理则主要依靠人类语言规模的语料，在语言表层进行语元关联频率计算，是自下而上的。本文认为，这两种模式并不冲突，要想提升大模型的推理能力，可寻找两者在推理运行机制底层架构上的共性，如搭建分类－比较架构；引入自然逻辑的推理规则；添加外部的推理评判标准等。

关键词：语元关联度预测　类比　符号主义　联结主义　自然逻辑

Abstract: In this paper, the weak inference performance of large language models (LLMs) is presented and analyzed based on some test results. It is pointed out that as token correlation degree prediction models, the essence of LLMs inference is pattern matching. However, the process by which a LLM generates an output based on an input problem can still be regarded as a kind of inference, although it is not the same as the human reasoning mechanism. Human reasoning implies inference rules, abstract-substitution mechanisms, default premises, semantic understanding, etc., which is top-down, while LLMs inference mainly relies on human language-scale corpora and calculates token correlation frequencies at the language surface, which is bottom-up. This paper holds that these two modes do not conflict, and in order to improve the inference ability of LLMs, one possible way is to find the commonality between them in the underlying structure of the reasoning operation mechanism; such as building a classification-comparison framework, adding the inference rules of natural logic to the LLMs, and introducing external criteria for inference.

Key Words: Token correlation degree prediction; Analogy; Symbolism; Connectionism; Natural logic

中图分类号：N031; O141.4　DOI: 10.15994/j.1000-0763.2026.05.002　CSTR: 32281.14.jdn.2026.05.002

当下，AI 大语言模型（Large Language Models, LLMs）的表现具有双面性。一方面，它似乎具备了与人交流对话的能力，且能够迅速完成很多人类可能需要花更多时间才能完成的文字生成任务；另一方面，它时不时冒出假话、胡话，而且，它的基础数学和推理能力仍经不起考验。从基底上来看，大模型的主体是基于语元之间关联度的预测模型，其开发并没有依赖经典逻辑学。如果想让大模型与人类形成真正的“交流”，且兼顾基础推理能力和强

基金项目：国家社会科学基金重大项目“不确定性推理的逻辑及其应用研究”（项目编号：24&ZD227）。
收稿日期：2025 年 2 月 10 日
作者简介：张立英（1978–）女，河北隆尧人，中国科学院大学人文学院、中国科学院哲学研究所教授，研究方向为哲学逻辑。
Email: zhangliying@ucas.ac.cn

大的语元关联算力，需要思考如何为大模型构建能够与人类基础推理能力相匹配的逻辑基底，以此提升大模型的逻辑推理能力。

本文首先结合近几年来对大模型展开的推理测试实例，综述大模型的弱推理表现；继而结合大模型的语元关联度预测机制，分析大模型为什么会有弱推理表现；在此基础上，本文将基于大模型的“推理”与人类推理的比较，重新审视推理的本质，指出大模型的“推理”也可被看成一种与人类推理机制有所不同的新型推理；最后，基于对两种推理的比较，本文指出，可通过还原认知的分类和比较要素来搭建大模型和人类推理所共同依赖的基底，并引入自然逻辑的规则作为大模型推理增强的参考。

一、大模型的弱推理表现

近年来，尽管AI大模型似乎具备了与人交流对话的能力，但它们的基础数学和推理能力仍有很大的提升空间。有诸多从不同角度展开的测试[1]-[4]显示，大模型在计数、符号推理、奇偶性、算术推理、子集求和、几何推理等方面存在不同程度的问题。

与此同时，为了提升大模型的数学推理能力，OpenAI发布了GSM8K的数据集。经过近两年的训练和调整，目前，诸大模型们在面对GSM8K的测试时，性能已经有了显著的提高。[5]但一种质疑是，由于这个数据集被拿来反复使用，很可能出现数据污染——用于测试的例子同时也被包含在了模型的训练数据中。

2024年10月，苹果公司的研究人员给出了基于GSM8K的微调测试系统GSM-Symbolic。该系统相较原数据集增加了三种微调方式：替换题目中的专有名词；改变其中的数字；添加无关信息，从而千变万化出更多难度相当的题目对大模型进行评估。测试表明，大模型对专有名词的变化表现出一定的稳健性，但对数值的替换非常敏感；而无关信息的添加会导致最先进的大模型的性能大幅下降高达65%。苹果公司的研究人员基于这些测评得出结论：大模型既不理解这些问题中的数学概念，也不能进行逻辑推理，而仅仅是将面对的问题和训练数据中的问题相比较而已。[6]

另一个与“GSM8K测试结果变好”背道而驰的例证是2024年6月德国学者们提出的爱丽丝漫游奇境（AIW）推理测试。测试问题如下：爱丽丝有N个兄弟，她还有M个姐妹。爱丽丝的兄弟有多少个姐妹？（N和M的数字可变化）对人类而言，这个问题并不复杂：答案是M+1，即爱丽丝的姐妹数量（M）再加上爱丽丝自己（1）。但该测试却让GPT-3.5/4、Claude、Gemini、Llama、Mistral等大模型几乎全线崩塌。[7]

值得一提的是，对强调推理能力开发的DeepSeek-R1[①]进行的测试显示，其计数和空间推理能力仍不理想。

二、大模型为什么会有弱推理表现

为什么大模型的推理表现经不起考验？这要从大模型的原理说起。

1. 大模型是语元关联度预测模型

大模型技术主要由预训练（Pretraining）、监督细调（Supervised Finetuning）和强化学习（Reinforcement Learning）三个板块组成，其中预训练作为大模型的主体和特色所在，其底层机制是由海量文本数据训练出来的语元（token）关联度预测模型。语元是指字、词、标点符号，或者由字词和标点符号组成的字符串等，例如“我要上网，请打开浏览器”这句话可分解为“我”“要”“上网”“，”“请”“打开”“浏览器”“。”等语元。[8]大模型基础模型预训练的目标就是结合（与之前出现的）语元的关联度来预测下一个语元，进而逐步生成整段的文字[②]。大模型生成的文字之所以很像

①DeepSeek在原理上已属于大模型+深度学习的综合改良模型，本文主要探讨的是基于相对纯粹的语元关联度预测技术的大模型及其推理机制。

②在具体的设置中，在预测输出下一个语元时，通常并非选择与前面的文字关联度概率最大的语元，而是选择关联概率较大的，具体设置各个大模型有所不同。

人话，是由于大模型背后有海量的代表人类日常语言沟通习惯的语料的喂养。注意，这些大模型的原始运算并没有附加额外的知识库、预设和推理规则等，这种极简的语元关联模式能够在最大程度上发挥计算的优势。不过，这也意味着语元关联度预测的底层设计上没有依赖于任何逻辑，因此，大模型对于上一节中所提到的推理问题的解答，并不是基于与人类相似的推理过程，而是面对输入的信息，通过关联度预测，经模式匹配得出答案。

2. 大模型目前没有抽象－替换能力

人类在进行推理时，会应用推理规则，除了掌握规则外，人们在使用规则时，还体现出两种能力：一是抽象的能力，二是替换的能力。例如，人们对≤关系的理解通常起步于简单的例子，如：从2≤3和3≤5，可以得到2≤5；再从例子抽象出具有广泛适用性的传递性规则“从A≤B和B≤C，可以得到A≤C”；再通过替换，把传递性应用于更多实例，如已知10^2≤101和101≤2^7，应用传递性规则，可得10^2≤2^7。类似的抽象规则还有很多，如数学运算中加法和乘法的交换律、结合律、分配律等；逻辑推理中的假言三段论（也对应某种传递性）、逆否命题转换、德摩根律等。由于学习经历、学习能力等的不同，每个人掌握规则的数量和熟练程度可能不同，抽象和替换能力也会有一定差异；但人类拥有抽象－替换能力——能够发现、总结或理解自然语言背后的一般性规则，并通过替换在自然语言层面应用这些一般性规则——这一点是无疑的。

但大模型是基于海量人类语料中的语元关联素材来预测生成语元关联序列，体现的是各个语元在自然语言层面的表层关联。基于海量数据，大模型的输出几乎不出语法问题，但这种方法暂时无法体现自然语言背后的抽象－替换推理过程①，大模型每一次生成的答案都相当于是个例，并没有通过一次次的生成总结出一般规律，当然也不会有一般规律的应用。上一节提到，大模型在计数、符号推理、奇偶性、算术推理、子集求和、几何推理等方面都存在不同程度的问题，而这些测试在不同的程度上都需要用到抽象－替换的能力，这也就能解释为何大模型推理给出的答案不尽理想。苹果公司的GSM-Symbolic测试，其微调模式中的前两种——替换题目中的专有名词；改变其中的数字——涉及抽象－替换能力；微调的方式的第三种，分辨一个条件与题目是否相关，则涉及对整段话的语义理解和分析。

3. 大模型不预设共识

对于爱丽丝漫游奇境（AIW）的推理测试，其中除了数字的抽象（M、N）和替换外，还涉及一些隐藏较深的常识知识，如：爱丽丝是女孩；爱丽丝的姐妹都是女孩；爱丽丝的兄弟都是男孩；男孩不是女孩；就爱丽丝和他的兄弟姐妹这个讨论范围中的每个人而言，每个女孩都是（除自己以外的）其他人的姐妹等。由于这些知识对于人类而言太过“基础”，反而在日常语料中出现得较少，此时，基于语料库出答案的大模型“崩塌”非常可以理解。

4. 大模型缺乏判定推理结论是否正确的机制

人类对于数学推理是否有效有相对统一的评判标准，且根据前提的情况，可以推不出结论。但作为概率预测和模式匹配，大模型一定会有输出。在爱丽丝漫游奇境（AIW）测试中，各大模型不仅几乎全线崩塌，还展现出“迷之自信”，宣称自己的“思考过程”非常合逻辑。[7]不仅如此，“大模型在面对超出其知识范围的问题时，还会基于其训练数据进行推测和‘幻觉’，生成看似合理但不真实的内容”。[9]

对以上分析的总结见图1。

三、重新思考推理

比照人类推理的方式来看，大模型似乎并没有推理，不过是基于语料、通过计算进行模式匹配。但如果依据人类所给出的推理界定重

①注意，这里讨论的是语元关联度预测方法不涉及抽象－替换，这不等同于计算机或其他人工智能方式没有抽象－替换，如苹果公司开发的GSM-Symbolic测试系统，本身就使用了抽象－替换。

图1 测试题目中体现的推理能力

新来思考这一问题，也许答案并没有那么想当然。

1. 大模型“推理”可以看作一种推理

在逻辑学领域，关于推理的定义有很多，如“依据一定的规则，由若干命题得出一个命题的思维过程”[10]等，为了在更一般化的范围下的讨论推理，我们不妨把依赖于逻辑理论的“命题”替换为更具一般性的前提/结论，同时弱化仍旧未有清晰界定的概念“思维”，则推理的界定可一般化为：从若干前提出发，依据一定的规则来推出结论的（思维）过程。基于这一界定，大模型基于提问（有时还附加提示）给出答案的过程也可被看作某种类型的推理，只是这种推理与人类推理的机制不同而已。

如果说人类推理所依赖的是逻辑规律，大模型的推理则没有遵循任何现有的（刻画演绎推理的）逻辑，是现有数学模型和逻辑系统不能刻画的一种新机制。要想理解大模型的推理机制，不妨先从人类推理出发，对人类的推理和大模型的推理进行比较。

2. 重新审视人类推理

人类的推理可分为演绎推理和非演绎推理两大类。

演绎推理是能够保证如果推理的前提都真则结论一定真的推理，前文提到的测试中涉及的数学推理，都属于演绎推理范畴。对于演绎推理，逻辑学家们给出了传统词项逻辑、命题逻辑、谓词逻辑等进行刻画，这些逻辑系统所总结的推理规律是具有抽象形式的一般规则，可通过替换反复使用。尽管是抽象规律，但这些规则背后也暗含着人类处理概念的模式，如传统词项逻辑依赖于属种关系；谓词逻辑依赖于个体–集合–性质–空位关系等，这些分析背后都包含着分类。

除了演绎推理，在日常生活和科学研究中，人们还大量使用归纳、类比等推理，这类推理无法确保从真前提一定得到真结论，又被归结为非演绎推理。非演绎推理虽然没有演绎推理那样的保真性，却恰恰是发现和创新的源泉，侯世达（Douglas Richard Hofstadter）等认为：“智能的核心在于类比”。（[11]，p.66）要想研究智能的跃迁，需要深入探究归纳、类比、隐喻等。目前，人们对非演绎推理的研究还远谈不上充分，但可以肯定的是，非演绎推理背后也存在一些可抽象化的结构性律则，如斯坦哈特（Eric Charles Steinhart）借助可能世界结构对隐喻进行了系统研究；[12]如张立英认为类比认知和类比推理包含了分类和比较两大要素，[13]在认知结构上还原分类要素则有助于区分类比、隐喻和举例论证[14]等。人工智能领域的开发者一直也非常重视类比智能的刻画，如侯世达、[11]因杜尔希亚（Bipin Indurkhya）、[15]米切尔（Melanie Mitchell）[16]等都深入研究过类比、隐喻与认知相关的问题。

结合对数学推理（演绎推理）的测试，第二节中已经对人类推理的特点做了一些分析，在本节中，我们把对推理的思考扩展到了更广、也更符合日常认知的范围，将演绎推理和非演绎推理综合起来，总结人类推理的特点如下：（1）人类推理中可包含缺省的预设或共识为前提；（2）人类推理遵循演绎逻辑的推理规则，这些规则是具有一般性抽象规则，可通过替换反复应用；（3）人类的推理可以没有（显性）前提，如重言式的得出只依赖于一般性的

公理和规则;(4)人类推理有抽象–替换机制;(5)人类推理涉及语义的解释;(6)非演绎推理包含分类、比较机制,事实上演绎推理中也包含分类;(7)有不同层次的推理可靠性标准(自我评价机制)。其中演绎推理要求保真性,正确的推理能保证如果前提真,则结论一定真;而非演绎推理则无法具有保真性,但同样追求推理的可靠。

3. 大模型推理带来新视角

为了探寻大模型语元关联度预测背后的运行机制[①],陈小平给出了刻画关联度预测的形式公理系统L_C及其形式语义解释,这一工作能够与当下大模型的行为模式相匹配,进一步揭示大模型底层运作的原理。L_C包含三条公理如下,分别是语境关联度、综合单调性和预测选择公理:([8], pp.896–897)

公理1(语境关联度):$0 \le C(a_i, b \mid a_1^n) \le 1$;

公理2(综合单调性):

$\Lambda_{i=1}^{n}(C(a_i, b \mid a_1^n) \le C(a_i, c \mid a_1^n)) \supseteq C(a_1^n, b) \le C(a_1^n, c)$;

公理3(预测选择):$argmax_b C(a_1^n, b)$.

其中,公理1假定了在任何语境下,一语元与另一语元之间都有值在[0, 1]区间内的关联度;公理2给出了n元关联度预测所遵循的规则,预测下一个语元时,要综合考量待选语元与前面出现的n个语元之间的关联度,原则是,如果一个待选语元a与前面n个语元的关联度比待选语元b与前面n个语元的关联度都大,那么语元a的关联度预测值比语元b的关联度预测值大;公理3则规定了以何种原则选择预测结果,这一条根据实际情况可以有不同的选择原则。

陈小平指出:"L_C只含一条推理规则即公理2,同时拥有大量推理前提即公理1,它们在L_C的推理中发挥了巨大的、主要的作用",这与经典逻辑系统和传统AI推理系统中推理规则发挥主要作用形成鲜明对比;"L_C的形式语言和公理都不含变元,所以L_C具有实例性,而以往的公理系统都是用少量含有变元的规则概括大量实例"。([8], p.897)基于这些公理,[8]还证明了系统的内定理,表明类L_C在任何输入下总有输出;而相对于传统推演都具备的一般传递性,类L_C只具有基于语境扩展的传递性。

基于以上分析和上一节中对大模型弱推理表现的分析,总结(关联度预测部分的)大模型推理的特点如下:(1)大模型推理是语元关联度预测模型,推理具有实例型,不预设共识;(2)大模型推理一定是有前提的,需要根据问题和提示(也可没有提示),以及海量的推理前提(语元关联)做答;(3)大模型推理不预设经典逻辑中的一般演绎规则;(4)推理遵循综合单调性规则,具有基于语境的扩展传递性,但不遵循传统的一般传递性规则;(5)大模型推理基于自然语言表层语料,不涉及抽象–替换机制,不涉及语义;(6)大模型推理中涉及比较机制,在公理1和公理2中都有体现;(7)大模型推理没有自我评价机制,无法判断自己的推理是否可靠;甚至还会自发生成看似合理但不真实的内容。

还需注意的是,在日常生活和科学研究中,演绎推理和非演绎推理是混杂在一起的。这也意味着大模型所分析的语料中也同时包含着演绎推理和非演绎推理。

4. 人类推理与大模型推理的比较

将人类推理和大模型推理的特点放在一起的对照(图2),可以看出人类的推理和大模型的推理各有所长。人类的推理在表层语法下另有一个不易察觉的暗线在运作,包括通用的推理规则、抽象–替换机制、隐藏的缺省前提、语义理解等。而大模型推理则是简单直白的外部推理,依靠人类语言规模的语料和强大的算力,基于语料表层的关联频率进行推理。简言之,两者一个主要靠内在的规则运行,一个主要靠外在的海量语料喂养输出。

对于人类推理和大模型推理的比较与人工智能领域的符号主义和联结主义之争有着呼应。符号主义基于形式化和规则来表达知识和推理,可被看成一种理论驱动的"自上而下"的加工,人类的(演绎)推理模式是符号主义的典范;而联结主义则可看作一种"自下而上"的数据驱动,当下基于语义关联度预测的大模型正是联结主义的代表作。在人工智能的发展历史上,曾经

①在大模型训练中,99%的计算时间用于预训练模型(基础模型),其核心功能是关联度预测。

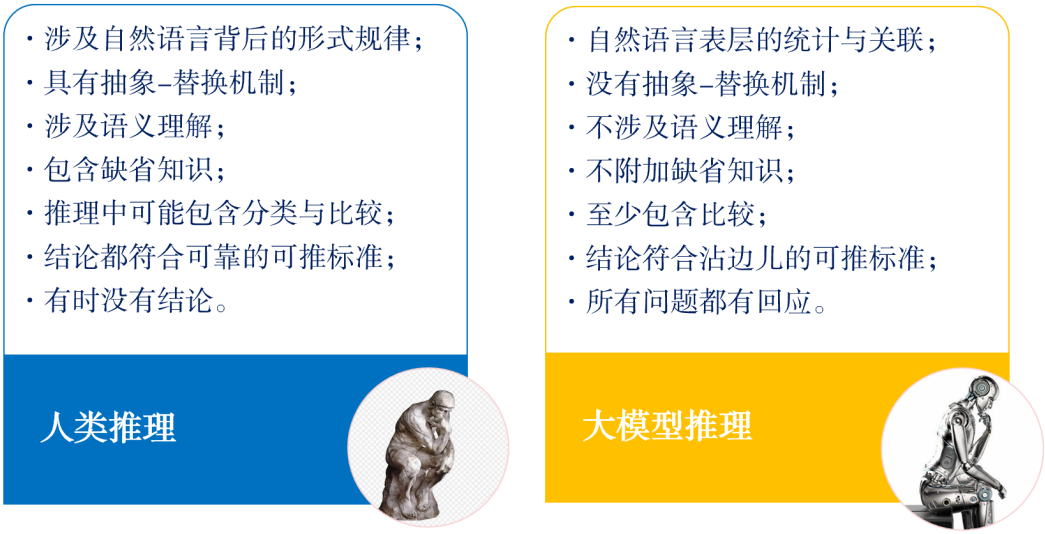

图2 人类推理与大模型推理特点比较

有一度是以符号主义为发展主线，但由于这些尝试的效果并不理想或显明，基于概率运算的联结主义在当下逐渐成为人工智能发展的主线。

但是，符号主义与联结主义真的如此对立吗？不一定。首先，以往人工智能的符号主义进路中所使用的逻辑基底基本是演绎逻辑（如谓词逻辑），但人类推理中还包含了归纳、类比等非演绎推理，尽管逻辑学领域对非演绎推理的研究仍不充分，但这并不是在构造人工智能时仅基于演绎逻辑的理由。这一事实提示一种可能性：也许不是符号主义的路径本身有问题，而是目前符号主义所使用的逻辑基底不够合适或说不够充分。此外，如果把推理考察的范围从数学推理扩展到日常推理，符号主义和联结主义的共性就显现出来了。张立英[13]指出，人类的类比认知包含了分类和比较两大要素，而如L_C的公理1、2、3所体现的，大模型推理中涉及比较机制，语元之间的关联要通过概率数值的比较得出。这意味着，比较作为重要的元素，同时出现在人类推理和大模型推理之中，也意味着同时可以作为符号主义和联结主义刻画的对象。也许可以把符号主义和联结主义之争放在一边，转而考虑是否可以搭建起两者之间的桥梁①。[11]

对于推理的判定标准：人类推理明确要求有判定标准，且可以“推不出”；而L_C的公理3则说明了大模型一定有输出。两者有比较明显的差异。需要注意，判定标准其实是推理之外的元标准，可以从外部施加。

四、大模型推理增强

大模型以人类语言为养料，其任务主要也是与人“交流”，尽管目前其推理呈现不尽人意，但由于大模型的输出很少出现语法错误，这使得人们与大模型打交道时产生了很多幻觉，大模型对人类社会已经产生了巨大的影响。此种情况下，需要尽快考虑大模型推理的提升或说增强问题，以减少大模型对人类社会的负面影响，促进人类与的大模型的深度互动和共生共赢。

目前已有不少对大模型推理增强的尝试，如被广泛热议的DeepSeek-R1主要是通过（前期训练和监督微调之后的）强化学习提升模型的推理能力，提升效果较为显著。DeepSeek-R1的“链式思考”包括步骤分解、自我评估与回顾、多角度尝试、回溯与修正等，展现出了类似人类思考过程的步骤。[9]本文认为其中的重要突破在于自我评估和回顾、回溯与修正，这对于增强大模型推理的自我评价有所助益，尽管其目前自我评估和修正之后的结果仍可能是错误

①早在上世纪80–90年代，侯世达团队开发的Copycat就已经在尝试对联结主义和符号主义加以融合。

的。不过，这样的推理增强是在语元关联度预测前期训练之后，这种强化学习并没有改变语元关联度预测方法的自下而上的根本，大模型推理仍旧未能总结一般性规律，推理具有个例性，不涉及语义。

对于大模型的推理增强，陈小平①提出以形式公理系统L_C为关联度预测和逻辑推理的统一架构，这种方法是从基底上的重新构建。[8]本文讨论的大模型推理增强，也是基底构建意义上的增强。具体有三点建议：在大模型底层增加关于分类认知的架构；如果给大模型增加推理规则，建议增加基于自然语言表层语法的自然逻辑的规则；在大模型内部运作之外，加上推理可靠程度的评判标准。

1. 还原分类认知架构

侯世达团队认为，“认知的本质是概念学习和类比活动”。（[11]，p.388）目前，大模型的语元关联处理并未预设人类的概念，这是其特点，也可能恰恰是其优势所在。但对于类比，无论从重要性还是从技术可行性角度考量，大模型的底层设置中都应尽早予以考虑。第三节中指出，人类的类比认知包含了分类和比较两大要素，其中比较要素是人类推理和大模型推理中体现出的共性。从长远来看，如果想搭建起大模型推理和人类推理共通的桥梁，还需在人类推理的研究和大模型的底层设置分类认知架构。

对于人类推理而言，这种设置是一种还原，即人类认知中本来就包含分类要素。演绎推理方面，第三节中分析了，传统词项逻辑和现代谓词逻辑都包含着分类要素。对于非演绎推理，张立英指出：“举例、类比和比喻可以看作围绕着类展开的三个层次。举例相当于同一物类内部的枚举，类比是在同一类事物间的比较，比喻则是在不同类的事物间比较”。（[14]，p.139）还原认知中的分类要素，可以厘清不同推理之间的关系，而要想表达没有分类或者有某一默认分类，也首先应有分类的概念。

图3展示了分类在区分举例、类比和比喻中立竿见影的作用。需要特别强调的是，以上基于类展开的不同层次，是一种纯粹的结构区分，可以不依赖于人类的概念体系。即举例、类比和比喻的区分是基于分类才产生的，但分类可以不止目前达成强共识的生物学分类一种。这也意味着，基于分类的这种架构或可直接嫁接到语元关联度预测的联结型架构上，而无需额外增加人类的概念化、语义相关的知识，如图4所示。

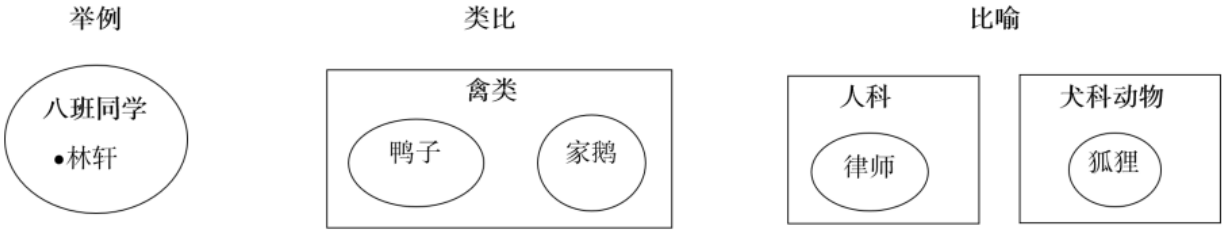

图3 人类推理中举例、类比和比喻的区分例

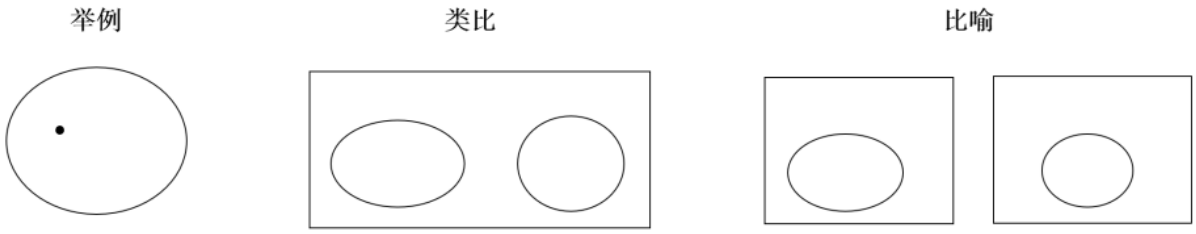

图4 三种基于分类的结构不依赖于现行科学概念体系

①四种大模型推理增强方法的总结来自陈小平2023年在北京师范大学的讲座。

对于大模型推理而言，眼下的关联度预测模式是没有预设分类的模式，可以看作分类架构下的一种特殊情况，这种模式去掉了诸多“条条框框”，直接探索最小语言单位之间的关联，也展示了“简单”的力量。不过，由于大模型所依据的是人类规模的数据，如果完全没有聚类或限定、空泛相比，数据就太多了；而如果完全限制在某一概念框架中，又不能出圈和发现。给人类和大模型共同添加分类底座的优势是大模型的处理内容可以更丰富，更贴合人类的思维方式；而人类也可突破一些分类限制，促进创新发现。

值得一提的是，目前，人工智能领域的很多实际操作其实是人为添加了分类或者聚类的，这也体现了分类的必要性和可行性，但是这种处理是应用层面体现的，本文讨论的底层架构层面的，两种有关联，但属不同层次。

2. 启用自然逻辑

在推理方面，大模型的极简关联模式的工作原理是使其具有强大威力的原因之一，但这种极简关联模式并没有提前预设逻辑基底，也导致了其推理能力的不可靠。目前业界提升大模型推理能力的方式是在大模型运算之前或者运算之后附加一些逻辑规则限制或过滤机制，但实施效果还有待时间验证。经典逻辑，尤其是一阶谓词逻辑，并不是基于自然语言的表层语法建构的，而大模型的语料则主要是来源于自然语言，如果能构建与自然语言相匹配的推理规则库，则可能会对大模型推理能力的提升起到更好的效果，基于自然语言表层语法的自然逻辑是一种备选方案。

自然逻辑尝试重新直接用自然语言来研究人类推理的基本模式。范丙申（Johan van Benthem）指出，像“所有的马是动物，所以，所有的马尾巴是动物尾巴”这样的推理以往被认为是传统词项逻辑无法处理的，需要借助谓词逻辑形式化方法。但自然逻辑认为，如果断定“有一个——尾巴”的某个位置用有更大的外延的“动物”来替换谓词“马”，马尾巴的例子就可以被纳入词项逻辑处理范围之内，这种规则叫做向上单调，[17] 是单调性中的一种，除单调性外，自然逻辑中还有保守性、对称性等推理规则。这些规则无需构建额外的人工语言，基于自然语言的表层语法进行设置即可。目前，莫斯（Lawrence S. Moss）等人对自然逻辑做了较为系统的研究，[18] 并探讨过自然逻辑在人工智能领域可能的应用。[19]

3. 增加推理的评判标准

目前，大模型推理没有自我评价机制，无法判断自己的推理是否可靠。人类推理中对于演绎推理有明确的评价机制；对于非演绎推理，既由于非演绎推理本身具有不确定性，又因为非演绎推理被研究的还不够充分因此并没有特别统一的评价标准，但仍旧有一些推理准则。对于推理是否可靠的判定或评价，属于推理之外的内容，在逻辑学领域属于元逻辑问题，对大模型而言，也应该对于输出结果增加评价机制。这是减少大模型假话、胡话的关键举措。

小　结

当下，大模型的输出呈现弱推理性，尽管各大模型开发公司正在试图对大模型进行推理增强，但结果距离理想还有一定距离。究其原因，大模型的核心是语元关联度预测模型，其输出是某种模式匹配，基底上并未依赖于某种逻辑。如果以“从若干前提出发，依据一定的规则来推出结论的（思维）过程”为推理的界定，那么，大模型推理也可看作一种类型的推理，只是这种推理与人类推理的机制并不相同。通过对大模型推理和人类推理的特点进行比较，可以发现，大模型推理的优势在于以人类规模的海量语料为前提，通过概率运算展开极简关联（自下而上）；而人类推理的优势则在于包含了推理规则、具有抽象–替换机制等（自上而下），这提示我们重新思考自下而上和自上而下两种方式的关系，以及人工智能领域的符号主义和联结主义之争。本文认为，以往的符号主义进路主要以演绎推理（数学推理）的形式规律为基础，但如果把推理的考察范围从演绎推理扩展到还包含非演绎推理的日常推理，就会发现，也许不是符号主义的路径有问题，而

是以往符号主义在具体处理问题时所选取的规律还不够全面恰当；而符号主义和联结主义也不必是对立的。通过对人类推理和大模型推理的比较，还可以发现，两者都以比较为基本要素。

由于大模型已经对人类社会产生了广泛的影响，需要尽快改善大模型的推理，以减少负面影响，促进人类智能和机器智能的共赢。本文认为，[8]所提出的搭建关联度预测与逻辑推理的统一架构能够从根本上解决问题，但在公理系统L_C底层，还应进一步增加关于分类认知的架构，这样能够同时兼顾日常推理的多种类型，且能够真正同时涵盖人类推理和大模型推理；此外，如果尝试给大模型增加一些推理相关的规则，建议增加同样基于自然语言表层语法的自然逻辑的规则；最后，需要在大模型内部运作之外，加上推理可靠性的判定的评价标准。

［参 考 文 献］

[1] Arkoudas, K. 'GPT-4 Can't Reason'[J]. arXiv preprint, arXiv: 2308.03762. 2023.

[2] Qin, C., Zhang, A., Zhang, Z., et al. 'Is ChatGPT a General-Purpose Natural Language Processing Task Solver?'[A], Bouamor, H., Pino, J., Bali, K. (Eds.) *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*[C], Singapore: Association for Computational Linguistics, 2023, 1339–1384.

[3] Zhu, Z. A., Li, Y. 'Physics of Language Models: Part 3.2, Knowledge Manipulation'[J]. arXiv preprint, arXiv: 2309.14402, 2023.

[4] Wei, J., Wang, X., Schuurmans, D., et al. 'Chain of Thought Prompting Elicits Reasoning in Large Language Models'[OL], arXiv preprint, arXiv: 2201.11903, 2023.

[5] Open AI. 'Arithmetic Reasoning on GSM8K'[EB/OL]. https://paperswithcode.com/sota/arithmetic-reasoning-on-gsm8k, 2025–01–26.

[6] Mirzadeh, I., Alizadeh, K., Shahrokhi, H., et al. 'GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models'[J]. arXiv preprint, arXiv: 2410.05229, 2025.

[7] Nezhurina, M., Cipolina-Kun, L., Cherti, M. et al. 'Alice in Wonderland: Simple Tasks Showing Complete Reasoning Breakdown in State-of-the-Art Large Language Models'[J]. arXiv preprint, arXiv: 2406.02061, 2025.

[8] 陈小平. 大模型关联度预测的形式化和语义解释研究[J]. 智能系统学报, 2023, 18（4）: 894–900.

[9] Karpathy, A. 'Deep Dive into LLMs Like ChatGPT'[OL], https://www.deciphr.ai/podcast/deep-dive-into-llms-like-chatgpt, 2025–02–09.

[10] 余俊伟、赵晓玉、裘江杰 等. 数理逻辑[M]. 北京：中国人民大学出版社, 2020, 18.

[11] 侯世达、流动性类比研究小组. 概念与类比——模拟人类思维基本机制的灵动计算架构[M]. 北京：机械工业出版社, 2022.

[12] Steinhart, E. C. *The Logic of Metaphor, Analogous Parts of Possible Worlds*[M]. Dordrecht: Springer Science+Business Media, 2001.

[13] 张立英. 类比的逻辑分析[J]. 科学·经济·社会, 2023,（3）: 13–26.

[14] 张立英. 牟子理惑论之论理辨[J]. 江淮论坛, 2024,（2）: 132–141.

[15] Indurkhya, B. *Metaphor and Cognition: An Interactionist Approach*[M]. Norwell: Kluwer, 1992.

[16] Mitchell, M. *Analogy-Making as Perception*[M]. Cambridge: Bradford Books/MIT Press, 1993.

[17] van Benthem, J. 'A Brief History of Natural Logic'[A], Chakraborty, M. K., Löwe, B., Mitra, M. N., et al. (Eds.) *Navya-Nyāya & Applications: Homage to Bimal Krishna Matilal*[C], Marshalls Creek: College Publications, 2008, 21–42.

[18] Moss, L. 'Natural Logic and Semantics'[A], Aloni, M., Bastiaanse, H., Jageret, T., et al. (Eds.) *Proceedings of the 17th Amsterdam Colloquium Conference on Logic, Language and Meaning, LNAI 6042*[C], Berlin: Springer, 2010, 84–93.

[19] Moss, L., Wollowski, M. 'Natural Logic in AI and Cognitive Science'[A], *Proceedings of Modern Artificial Intelligence and Cognitive Science*[C], Bielefeld: CEUR-WS, 2017, 41–46.

［责任编辑　王巍　谭笑］