认知神经科学

神经语言模型与大脑语义表征的综合文献综述

融合计算模型与语言神经科学的关键议题

Neural language models and brain semantic representations: an integrative review

Huth (2016), Schrimpf (2021), Kumar (2024), Millet (2023), Toneva (2022), Tang (2023) 等

报告人 潘宇轩

学　号 2023K8009991004

院　系 人工智能学院

专　业 人工智能

日　期 2026 年 1 月 18 日

摘　要

近年来，深度学习技术与认知神经科学的交叉研究为理解人类语言处理机制开辟了新途径。本文综述了神经语言模型与大脑语义表征对比研究的最新进展，涵盖预测编码理论及其争议、语义系统的脑映射、自监督学习机制、Transformer 架构的功能专化、组合语义表征以及跨语言和跨模态比较等核心主题。研究表明，语言模型在脑编码任务中的成功可能源于其捕获的丰富语言统计结构，而非单纯的预测机制。语义系统呈现分布式和网络化特征，涉及默认模式网络、注意网络等多个功能网络的协同作用。本综述旨在整合现有研究成果，揭示人工神经网络与大脑语言处理的对应关系，并指出未来研究的关键方向。

目录

1 引言

近年来，随着深度学习技术的突破，人工神经网络在自然语言处理（NLP）领域获得了前所未有的成功。与此同时，认知神经科学的发展使得我们可以通过功能磁共振成像（fMRI）、电生理记录（ECoG）或脑磁图（MEG）等技术测量大脑在理解语言时的动态响应。将这两条研究路径结合起来，比较神经网络模型的内部表征与大脑活动，既有助于解释模型为何有效，也有助于揭示人脑处理语言的规律。特别是语言模型的训练目标往往是预测下一个词，这与某些理论认为大脑通过预测未来输入来减少处理负荷的观点不谋而合。然而，模型解释大脑的成功是否意味着大脑真的在实施同样的预测机制，仍存争议。

另一方面，数据驱动的语义地图绘制为理解语义系统提供了全新的视角。在自然语言刺激下，大脑的语义区域如何分布？语义类别和语义关系如何在皮层中表示？这些问题的解答不仅深化我们对语言认知的理解，也可能反过来指导人工模型的改进。本综述不按单篇文章分别介绍，而是围绕核心主题整合现有研究成果：首先概述预测编码理论及其在语言领域的支持与质疑；随后总结利用神经语言模型建立大脑编码模型的集成建模工作；接着讨论语义系统的脑映射，包括语义类别与语义关系的分布式表征；之后阐述不同学习机制下模型与大脑的对齐，如自监督语音模型和 Transformer 的注意头分析；进一步介绍组合语义（supra-word）表征的发现；最后探讨跨语言、跨模态比较的意义，并在讨论部分提出未来研究方向。通过这种综合视角，我们旨在呈现大脑语言处理与人工神经模型互动的完整景观，并指出当前研究的关键问题和潜在路径。

本领域的交叉研究不仅推动认知科学的发展，也对人工智能的设计产生深远影响。一方面，神经网络模型已从仅追求任务性能的“黑箱”演进为可以解释人类数据的认知模型。通过将脑成像数据纳入模型评测，研究者得以揭示模型所学的语言表示是否具有生物学真实性，从而指导模型架构和训练目标的调整。另一方面，大脑科学家利用模型生成具备特定语言属性的刺激，设计实验以验证大脑的语义组织、加工顺序和层次结构。例如，模型可以产生具有不同预测难度或不同语法依存结构的句子，用于探索大脑在预期违背或结构复杂度下的反应模式。正如后文所述，这种双向互动正在形成新的研究范式，促使人工智能和神经科学共同迈向更高水平的理解。

近年来，随着脑成像技术的进步，研究者不再局限于通过行为数据推断认知过程，而是能够在毫秒乃至更高时间分辨率上观察大脑对语言刺激的响应模式。结合深度学习模

型，这种方法提供了一个跨尺度的桥梁，使我们能同时访问神经元群体的活动和算法级别的信息。与此同时，人工神经网络的内部表征从简单的线性结构发展为复杂的层次化系统，这些系统在训练过程中自发产生语法树状结构、意象表征乃至对世界的统计认识。对比这些表征与脑成像数据，可以揭示哪些结构是普遍存在于人类语言认知中的通用属性，哪些是模型特定的产物。

此外，围绕人工智能的伦理与实际应用问题也提醒我们，理解模型如何与人类认知对齐具有重要意义。语言模型被广泛用于教育、医疗和政策决策等领域，若忽略其与人类理解方式的差异，可能导致误解或偏差，甚至出现严重的社会后果。通过研究神经模型与大脑的对应，我们可以识别模型在语义推理、情感理解或常识推理方面的缺陷，并制定改进策略，例如在训练数据中引入多样化的语境、增加现实世界知识或限制模型的信任范围。总之，人工智能与神经科学的交叉研究不仅拓展科学边界，也对社会和伦理产生深远影响。

值得指出的是，语言不仅是词序列的累积，还包括丰富的形态学和语用学信息。许多语言使用屈折词缀、变化规则和音调来表达语法关系和时态、体、态等语义特征。例如，阿拉伯语的三根辅音根通过元音变换表达不同的词义，因纽特语的多重粘着导致单个词即可表示复杂句子。这些形态学特性为预测提供了额外线索：在富形态语言中，词干和词缀的组合给出句法框架，大脑可能利用形态学规则来缩小下一词的候选范围。未来的神经模型应考虑形态学多样性，不能仅依赖英语语料，否则在解析屈折或黏着语言时可能无法对齐大脑处理。随着研究扩展到更多语言，比较不同形态类型的大脑活动将揭示语言特性对语义系统塑造的影响。

此外，语言的句法结构也对预测机制和语义分布有显著影响。语序固定的语言（如英语）与语序自由的语言（如拉丁语或俄语）在理解过程中的策略不同：在语序固定的语言中，听者和读者可以依赖固定的位置来识别主语、宾语和谓语，而在语序自由的语言中，需要依靠格标记和语义角色进行解析，这可能增加理解负荷并影响预测的层级。此外，汉语等话题优先语言常通过主题—述题结构组织信息，大脑必须在句子前段确定主题并预测后续信息。在某些美洲原住民语言中，主谓宾顺序可以灵活改变以突出新信息或对比，这种突出与焦点结构需要更复杂的语用推理。语言模型和脑编码研究应进一步探索语序变异如何影响预测和语义系统的组织，通过控制句法自由度和引入格标记等操作，比较不同语言的神经反应模式。

2 预测编码理论及其争议

2.1 预测编码的基本思想

预测编码（predictive coding）是一种流行的认知理论，它认为大脑通过不断生成对未来输入的预测来高效加工感知信息。大脑产生内部模型预测即将到来的刺激，并通过比较预测与实际输入的偏差（预测误差）更新内部模型，从而在多层次上实现感知和理解。这一框架最早在视觉系统提出，随后扩展到听觉和语言领域。语言过程中，大脑可能利用上下文信息预测接下来要出现的词语、语法结构或语义内容，减少加工负担。行为实验和脑电研究（如 N400、P600）提供了一些间接证据，表明当语句违背语法或语义预期时会出现特异的神经反应。

人工神经网络语言模型（NNLM）因其训练目标与预测未来词语相关，被视为测试预测编码假说的理想工具。语言模型通常通过最大化下一个词的概率来进行自监督学习，这使模型的隐藏表示内化了大量语法、语义和篇章结构信息。如果人脑确实在理解语言时进行类似预测，那么这些模型的表现应与人脑活动高度一致。一些早期研究发现，基于 RNN 或 Transformer 的语言模型可以解释大量语言相关脑区的活动，且模型的预测准确率与脑解码性能呈正相关。这些结果被视为支持大脑预测编码的证据。

需要强调的是，语言中的预测可能涉及不同层次：在语音层面，人类利用共现概率预测下一个音素或重音；在词法层面，预测词的词形和词性；在句法层面，预测短语结构和句法角色；在语篇层面，预测话题发展或叙事结构。人脑可能在这些层次上并行生成预测，而当前神经语言模型主要专注于词序列层面的预测，忽视了语音和句法层面的单独预测。此外，语言模型中的误差信号仅用于训练过程，推理时不会在层间反馈，而预测编码理论强调误差信号自下而上传递、更新内部模型。早期 ERP 指标如 N400 与 P600 虽常被解释为语义或句法预测误差，但也有研究认为它们反映语义整合困难、冲突监控或重新分析等多种认知过程。因此，尽管神经语言模型的成功为预测编码提供了重要线索，我们仍需谨慎区分模型的预测目标与大脑真实的预测机制。

预测编码理论起源于贝叶斯大脑假说，将大脑视为一个生成模型，在每个时间步预测感官输入的概率分布，并通过最小化预测误差维持内外信息的一致性。在语言领域，这意味着不仅要预测即将出现的词，还要预测其词法形态、语法角色以及在语篇中的功能。例如，英语的限定词后可能接名词，而日语中的助词指示了不同的语法关系，大脑需要根

据语言的语法规则调整其预测。不同语言之间的词序和结构差异也挑战了简单的线性预测框架，提示预测可能是多层次、动态调整的过程。

另一个相关的理论是生成式模型，它认为大脑在理解语言时会构建一个潜在的生成过程，形成关于事件、场景或对话的内在模型。这种生成式推理不仅包含预测，还涉及假设检验和信念更新。例如，在叙事理解中，听者会建立角色间的因果关系和动机结构，对即将发生的情节做出推断。当新信息与内在模型不符时，大脑将产生突出的预测误差信号，从而驱动理解的修正。神经语言模型多采用自回归训练，但真实生成过程可能依赖递归和层次化的预测，这也是模型尚未完全捕捉到的。

预测编码的思想早在上世纪九十年代在视觉系统提出，Rao 和 Ballard 提出了一个分层模型，上层神经元产生低层输入的预测，下层神经元反馈误差信号驱动上层更新。在听觉和语言研究中，类似的框架用于解释为什么语境越丰富，大脑的听觉皮层响应越弱：这种现象被解释为预测误差减少。该理论强调自上而下连接的作用，认为皮层层次之间通过反馈信号调节编码。但大脑皮层的解剖显示出复杂的双向连接，预测编码模型如何映射到这些解剖结构仍具争议。例如，颞上沟的层间连接既有前馈也有反馈，高层视皮层与耳蜗的反馈回路则难以用简单的预测误差解释。此外，语言中的冗余和多义性使得预测必须结合词义和语境，不同语言的后缀、语序自由度也影响预测策略。为了准确评估预测编码理论，研究需要在不同语言和任务上系统测量预测误差信号，并区分语法预测、语义预测和语用推理。

在预测编码的讨论中，还需区分不同层次的预测目标。生成式贝叶斯模型强调，大脑不仅预测即将到来的感官输入，还预测输入的原因。例如，听到特定音节后，大脑可能预测未来出现的语义类别或句法角色，而不仅是具体词形。语言理解涉及对外界事件和他人意图的建模，这种隐式推理远超下一词预测。研究发现，当叙事违反人物动机或事件因果时，大脑的默认模式网络产生强烈反应，但这与词层面的 surprisal 无关。另一方面，预测并非总有利于理解；在幽默或修辞反转中，出乎意料的表达反而提高了记忆和理解。神经语言模型通常通过最大化下一个词的概率学习，但大脑在面对新颖和意料外的句子时可能主动抑制预测，以保持开放性。因此，未来实验应将预测难度、句式复杂度和语用意图作为独立维度操控，通过组合这些因素进一步测试预测编码假说的边界。此外，生成式模型与分布式语义模型并非对立，可通过变分自编码器等框架联合实现预测与解释；这些模型能同时学习生成过程和意义结构，为理解大脑如何整合预测与概念知识提供新的工具。

2.2 对预测编码证据的质疑

尽管语言模型在脑编码任务上的成功常被用作预测编码的证据，但这一推论遭到质疑。Antonello 与 Huth（2023）[1] 对神经语言模型与大脑匹配的机制进行了细致分析。他们发现，同一模型的不同层在解释大脑活动时表现迥异：用于词预测任务的高层隐藏状态并非最佳的神经预测器，相反，中间层的表示更能解释大脑数据。此外，他们提出了一个衡量模型在多种下游任务上迁移能力的“通用迁移性能”指标。研究发现，该指标与模型的脑编码性能同样相关，而与下一词预测性能的关系并不更强。因此，模型在大脑编码任务中的成功可能源自其学习到的丰富语言结构，而不仅仅是预测任务本身。

另一个支持这一观点的证据来自层级分析。对比模型不同层的表示，研究者发现中层能够捕获局部依存和语义结构，而高层则更专注于全局预测。在大脑编码任务中，中层表示的解释方差显著高于高层表示。如果预测编码是唯一关键因素，理论上最擅长预测任务的高层应该最能解释大脑活动，但事实恰恰相反。该研究因此提醒，不能简单把模型的预测目标与大脑的认知目标等同，也不能基于模型的预测准确率直接推出大脑采用预测机制。

在质疑预测编码假说的工作中，Antonello 与 Huth 对神经模型与大脑匹配度展开系统性检验。其数据集由五名健康成人组成，每名受试者在多次扫描中聆听约五小时的英语播客故事。研究者构建了 97 个特征空间，包括词向量、模型不同层的隐藏状态以及手工设计的语言学特征，并通过 Lanczos 插值将这些特征时间序列与 fMRI 采样率对齐。在回归模型中，他们分别比较不同特征对体素级响应的解释能力，结果显示，Transformer 模型的中层表示在预测脑活动时普遍优于后层和早层，且这种优势在不同被试和不同脑区中一致。

除了层级分析，研究者还引入通用迁移性能指标，用以衡量模型在情感分析、命名实体识别、翻译等多任务上的表现。令人惊讶的是，该指标与脑编码性能的相关性与下一词预测相当，甚至在某些情况下更强，这表明模型的迁移能力可能比其预测能力更能反映大脑语言处理的本质。这一发现促使我们重新思考预测编码假说的适用范围：语言理解可能依赖多种学习机制，如统计学习、结构学习和语义推理，而不仅是简单的下一词预测。此外，作者强调，未来研究应通过实验操控模型的训练任务，如比较纯预测任务与多任务学习模型在脑编码中的差异，从而更严格地检验预测编码的贡献。

Antonello 与 Huth 的质疑工作不仅比较了模型层级的性能，还分析了由语言学家手工标注的特征，如词性、句法依存树、形态标签等。他们发现，这些传统特征单独时预测脑活动的能力远低于神经模型表示，但将其与模型嵌入组合可以略微提升性能，表明模型捕获的高级特征包含这些语言学信息。值得注意的是，他们使用约 5 小时的自然播客

故事，总共约 45000 个单词，几乎涵盖多种主题和叙事风格。研究者利用卷积插值将单词特征对齐到 fMRI 时间点，并采用嵌套交叉验证确保统计可靠性。这样的严谨设计减少了过拟合风险，也为其他研究提供了基准。作者还指出，模型在任务外迁移性能上的表现和语法性判断、语义相似度等任务的相关性不高，这表明各任务衡量的语言能力不同，未来应构建多维评估指标综合评估模型与大脑的对应。

2.3 集成建模对预测性的支持

与上述质疑相对，另一系列研究采用大规模集成建模方法，通过比较不同模型、不同任务和不同数据集，系统检验模型性能与大脑匹配度之间的关系。Schrimpf 等（2021）[9] 汇集了 43 种语言模型，包括不同深度的 Transformer、循环神经网络（RNN）和静态词向量，并对比它们在多个神经和行为数据集上的表现。数据集包括人们阅读或听句子时的 fMRI 和 ECoG 信号以及反应时间等行为指标。结果表明，Transformer 模型显著优于 RNN 或静态嵌入，容量越大性能越好。更重要的是，模型的脑拟合度、行为拟合度与其下一词预测准确率之间存在显著正相关，而与其他语言任务（如句法分析、文本分类）无关。此外，未训练的 Transformer 也能解释部分大脑数据，表明模型架构对匹配度有基础性贡献。

综上，集成建模似乎支持预测编码：能够更好预测下一词的模型往往更能解释大脑数据。然而，这种关联并不能排除其他解释。例如，大型模型在训练过程中捕获了更丰富的语言统计规律，其表现优异可能来自于表示的复杂性而非预测任务本身。因此，预测与特征学习之间的贡献仍需通过控制实验加以区分。

集成建模工作汇聚了多个数据集，旨在探索模型性能与脑匹配度之间的普遍规律。Pereira2018 数据集包括 78 名参与者在阅读约 400 个句子时的 fMRI 信号，每个句子呈现多次以提高信噪比。Fedorenko2016 数据集采用 ECoG，记录 12 名癫痫患者在阅读单词或短语时的皮层电活动，具有高时间分辨率，适合分析迅速的语音和语义过程。Blank2014 数据集让参与者聆听约五分钟的自然故事，捕捉更生态的语篇处理。通过在这些数据集上对 43 个模型的各层表示进行线性回归，研究者发现模型的脑拟合度与下一词预测准确率之间呈现强相关，且 Transformer 模型几乎达到了噪声上限。

深入分析表明，模型架构对大脑匹配度的贡献不可忽视。即便未经训练，Transformer 架构也能在一定程度上预测脑活动，这表明多层自注意结构本身具有与语言网络相似的组织方式。此外，模型的容量越大，匹配度越高，提示大脑语言系统可能利用高维表征整合丰富的语境信息。集成建模还强调，模型在其他语言任务（如问答、句法分析）上的性能与脑拟合度几乎不相关，说明预测任务在当前模型框架中仍然是最能反映大脑语言处

理的代理任务。然而，这并不意味着大脑只做预测，而可能是目前的预测任务同时涵盖了语义、语法等多维信息，所以表现出较强的相关性。

Schrimpf 等人的集成建模研究除了下一词预测任务，还评估了模型在机器翻译、问答、句法分析等任务上的性能。他们发现这些任务的准确率与脑拟合度之间相关性较低，这意味着大脑理解语言可能不依赖模型在某些人工任务上的表现。研究还发现，不同受试者、不同刺激材料和不同脑区之间的匹配度差异较小，说明某些规律具有普遍性。他们还使用双任务对比，比较随机初始化模型、预训练模型和经过微调的模型，结果显示预训练是获得高脑拟合度的关键，而微调对特定任务并不能显著提升脑对齐。这一观察支持了模型通过大规模无监督学习捕获人类语言统计结构的重要性，同时也提示未来模型设计应优先考虑预训练阶段的任务和数据多样性。

集成建模之所以能够揭示模型性能与脑拟合度的关系，得益于对多样数据集和多重评测指标的系统整合。Schrimpf 等人在比较 43 个语言模型时，不仅考虑模型的下一词预测准确率，还综合了模型规模、层数、训练语料大小以及是否使用双向或自回归架构。他们发现模型容量与脑拟合度呈非线性关系，层数增加初期可显著提升预测能力，但超过一定深度后收益递减。数据量方面，模型在数十亿词语料上训练可以捕获更丰富的语义和句法统计信息，但超过某一阈值后提升有限。研究还比较了微调模型在特定任务（如翻译、问答）上的表现，结果显示针对下游任务的微调有时会降低脑拟合度，可能因模型过拟合任务数据而破坏其通用表征。这提示我们，在构建大脑对齐的语言模型时，应平衡任务适应与通用语言知识的保持。此外，集成建模涵盖的神经数据主要来自英语，未来应引入其他语言和文化的脑数据，将模型的普适性检验扩展到更广泛的语言生态。

3 语义系统的脑映射与分布式表征

3.1 数据驱动绘制语义地图

在自然语言理解中，大脑对语义信息的表示是分布式的。Huth 等（2016）[4] 通过让受试者聆听长达两个小时的自然故事，利用词共现构建的 985 维词嵌入和体素级正则化回归，绘制了语义系统的高分辨率地图。模型经交叉验证可预测新故事的 fMRI 响应，说明嵌入捕获了稳定的语义特征。通过对模型权重进行主成分分析，他们提取出四个跨受试者共享的主要语义维度，并使用 PrAGMATiC 算法将这些维度投影到皮层，发现左

半球约含 77 个语义区域，右半球约 63 个。这些区域不仅涉及传统语言区，还扩展到顶叶皮层（LPC）、内侧顶叶（MPC）和前额叶的默认模式网络（DMN）区域，其中中心区域偏向社会和人物概念，外周区域偏向数字、视觉或触觉概念。这一发现打破了仅有左侧优势的传统观点，揭示语义系统在左右半球间较为对称。

除绘制地图本身外，Huth 等还详细描述了模型的训练和评估流程。他们构建了 985 维语义特征矩阵，其中每一维度表示英语词与语料中其他词的共现概率。为了消除低级听觉因素，模型在回归时加入词率、音素率和声学特征作为协变量。在训练阶段，研究者利用 10 折交叉验证评估模型对新故事的预测能力，保证了结果不依赖特定刺激。这种严格的验证使绘制出的语义地图具有可重复性和泛化性。通过观察各语义维度在皮层表面的分布，他们发现概念的抽象程度呈后—前梯度：背侧和后侧区域偏向具体感官相关概念，如动作、视觉和听觉；腹侧和前侧区域则偏向抽象、社交和情感概念。该渐变跨越颞叶、顶叶和前额叶，反映大脑可能按抽象度或表象类型组织语义信息。作者建议，这一渐变与从知觉到抽象思维的连续加工路线相一致，为理解概念结构提供了新的神经学证据。

这种数据驱动的语义映射方法具有多重意义。首先，它利用自然故事这一生态刺激，克服了传统实验采用单词或短语的限制，展示出在真实语境下绘制语义地图的可行性。其次，语义地图可作为参照框架，便于不同研究之间比较语义表征位置。作者指出，未来研究需改进区域划分算法，兼顾离散区域与功能渐变。此外，通过跨语言和跨文化采样，可以检验语义系统的普遍性和可变性。

基于自然故事的语义映射提供了更全面的视角。一些后续研究将 Huth 等的方法扩展到其他语言，如西班牙语、法语和中文，发现语义地图的宏观结构具有高度一致性，这表明大脑语义系统的组织具有跨语言普遍性。然而，在某些文化特定概念上，如颜色词、亲属称谓或食物名称，不同语言的激活模式存在细微差异，这可能与语言中相关类别的词汇丰富度或社会文化重要性有关。通过比较不同语言的语义地图，研究者可以揭示概念表示的文化可塑性及其神经基础。

语义地图还揭示了默认模式网络在语义处理中的核心地位。DMN 包含的角回、后扣带皮层和内侧前额叶长期以来被认为参与自发思维、内省和记忆检索。Huth 等的结果显示，这些区域也积极参与语义加工，尤其是涉及社会、情感和思维推理的概念。这一发现使我们重新审视 DMN 的功能定位，提示它可能不是专门处理“与任务无关”的思维，而是在语义推理过程中发挥中枢作用。语义地图的结果还打破了传统认为左半球主导语言的观点，揭示两半球在语义任务上的对称性。不过，由于 fMRI 在前颞叶的信号噪声较大，某些语义区域可能仍未被发现，因此需结合 ECoG 或高场强 fMRI 提高空间分辨率。

在对语义地图的后续分析中，研究者进一步解析四个主要语义维度分别对应哪些概念群。例如，第一个维度从有生命的生物到无生命物体，反映动—静连续体；第二个维度

从社会交往到工具使用；第三个维度从视觉场景到感知属性；第四个维度从数量与空间到情感与心理状态。通过比较这些维度在皮层上的渐变，发现背侧区域偏向具体感官经验，腹侧区域偏向抽象社会知识。实验还比较了不同故事段落中语义向量的变化，结果显示语义维度的激活模式随故事推进而动态演化，表明语义处理具有时间依赖性而非静态。作者提出，可以将语义地图作为生成刺激的指南，选择刺激中激活特定区域的词语或句子，研究这些区域对语义推理的因果作用。

语义地图不仅揭示出概念在皮层表面的分布，还呈现出随着故事语境变化而动态更新的模式。后续研究利用滑动窗口技术，分析语义维度的激活随时间的变化，发现故事高潮时期社会和情感维度的激活显著增加，而叙述背景期则更多激活物体和场景维度。这表明大脑语义表示具有时间敏感性，会根据故事进程调整重点。另一些研究通过比较不同叙事体裁，如对话、新闻报道和诗歌，发现诗歌中的抽象情感维度激活更强，而对话和新闻则更依赖社会互动维度。除此之外，跨语境分析显示，同一概念在不同故事中的激活模式可能不同，反映语义表示与篇章背景的耦合。未来可以结合自然语言生成模型，控制故事内容和风格，系统探索语义表示的动态调节。进一步地，利用联结梯度分析，可以描绘语义网络在皮层内部的连续变化，揭示从感觉相关区域到抽象思维区域的渐变。这种多维度、多时间尺度的语义地图将为理解语言语义的动态生成提供新的视角。

3.2 语义类别与关系的分布式编码

在对语义地图的进一步探索中，Zhang 等（2020）[12] 让受试者听 11 小时故事，建立体素级编码模型，预测数千个单词的脑响应。他们发现，大脑并不以离散模块表示不同语义类别，而是通过广泛重叠的区域同时编码多种类别。例如，工具类概念在左侧顶下小叶、后中颞回和颞上回均有表示；交流与情感等抽象概念则在右前颞区和顶叶更为明显。通过分析词汇的具体性，研究者观察到左半球偏向具体、感官相关概念，而右半球更偏向抽象、内省相关概念。这一左右半球差异揭示语义系统在处理不同类型概念时的功能特化。

不仅语义类别，概念之间的语义关系也可以映射到皮层。Zhang 等利用词向量差表示语义关系，例如整体—部分、类—属、对象—属性等，并构建关系编码模型。他们发现，“整体—部分”关系在默认模式网络区域（如角回、后扣带皮层）呈现明显激活，而前顶叶注意网络呈现抑制；其他关系则在不同网络中表现不同的激活抑制模式。这种共同激活与抑制的模式表明，大脑通过协同的功能网络而非独立区域编码语义推理。作者还发现语义关系的网络模式与语义类别脱钩，例如“手—手指”和“动物—动物园”属于不同类别但具有相似的关系模式。这表明大脑可能存在专门处理抽象关系的网络，与 DMN 中的

思维漫游或内省功能相联系。

值得进一步说明的是，Zhang 等划分的九个语义类别包含工具、人类、植物、动物、地点、交流、情感、变化和数量等，每个类别又由数百个单词组成。对这些类别的分析显示，具体概念（如“锤子”“狗”“苹果”）在多感官和运动相关区域呈现较强激活，而抽象概念（如“自由”“希望”“交流”）则在前额叶和顶叶默认模式网络表现更强。语义关系方面，除了整体—部分、类—属和对象—属性，研究者还考虑了名词与动作之间的关系、时间关联、空间关联和事件因果关系。他们利用 SemEval-2012 评测集中的句子对构建差向量，并利用这些向量预测脑响应。结果表明，不同关系在皮层上呈现高度一致的空间模式，说明大脑可能通过共享的网络处理各类关系推理，而不是为每种关系单独配置区域。这一发现拓展了我们对语义系统功能的理解，表明抽象关系加工与默认模式网络的内在思维密切相关，并可能涉及对情境和情感的整合。

Zhang 等进一步分析了不同类别词汇在皮层中的精细分布。例如，人类和动物概念不仅在后颞与顶叶区域活跃，还在视觉皮层的被动激活区出现，可能与想象或回忆相关；情感和交流类词汇在角回、内侧前额叶等 DMN 区域更强，说明这些抽象概念与自我反省和社会认知密切相关。研究者还比较了不同关系类型，如“物品与使用者”“因果与结果”“部分与整体”，发现关系向量激活模式的相似性与关系的逻辑结构相关。例如，因果关系与时间关系的皮层模式相近，体现故事中事件连贯性在脑中的共通表示；而反义关系激活的网络更分散，可能需要更多注意和工作记忆资源。这些发现提示语义关系不仅是词义差向量，在大脑中也体现为跨网络的协同模式。

3.3　地图与网络的意义

语义地图和语义关系研究表明，语义系统既有局部分区又存在跨区域的连续功能梯度。Huth 等的 PrAGMATiC 算法划分出多个语义区域，但假设每个区域内部同质，这难以捕捉某些功能渐变。未来需要发展既能识别离散区域又能描述功能渐变的模型，如连通性梯度分析。Zhang 等的研究强调语义关系的网络化特点，通过默认模式网络与注意网络的协同活动编码抽象关系。这些发现提示，语义认知不仅依赖单个区域的选择性，也依赖跨网络的动态交互，理解语言中的推理和抽象思维需从网络视角入手。

通过结合皮层连接信息，研究者建议语义系统可划分为若干功能网络：中心的 DMN 支持抽象概念和情景推理，背侧注意网络支持概念的检索和选择，腹侧语义网络支持具体物体和动作信息。语义关系的编码往往跨越这些网络，例如“部分—整体”关系需要同时激活 DMN（处理整体概念）和抑制顶叶注意网络（抑制无关信息）。这表明语义处理可能涉及网络间的抑制与兴奋平衡。未来需要采用动态图模型或有效连接分析，理解在

语义推理过程中网络交互的因果顺序。此外，语义网络与其他认知网络如工作记忆、情绪和奖励系统的交互也值得探索，因为实际语境中的语言理解往往伴随情感体验和行动决策。

4 学习机制与模型：自监督语音与 Transformer

4.1 自监督语音模型 wav2vec 2.0

人工语音学习通常不依赖大规模标注，因此自监督语音模型能更贴近人类语言获得的条件。Millet 等（2023）[6] 考察了自监督语音模型 wav2vec 2.0[2] 与大脑活动的对应性。wav2vec 2.0 由三个主要部分组成：一个由七个卷积块构成的特征编码器，将原始 16 kHz 语音转换为低维潜在表示；一个量化模块，将连续表示映射为有限的离散符号词典；以及一个 12 层的 Transformer 上下文网络，通过自注意机制整合长距离信息。模型的自监督训练目标是预测被掩码帧的离散表示，训练引入对比损失和多样性损失，使模型既能依赖上下文，又能充分利用离散向量。

作者比较了随机初始化、自监督训练、监督训练以及跨语言训练的模型，并以这些模型的层级表示为特征，建立线性回归预测多个受试者聆听有声书时的 fMRI 响应。研究发现，自监督模型在大约 600 小时未标记语音上即可学习出与人脑相似的表示，模型的卷积层、量化层和 Transformer 层分别对应人类语音皮层的不同处理阶段。中层表示在听觉皮层和颞上沟表现最佳，后层 Transformer 表示则更符合语言皮层的反应。这些趋势在法语和普通话受试者中同样显现，表明自监督模型捕获了跨语言的声学与语音规律。行为实验结果也显示，模型的层级专化与人类语音辨别任务表现一致。

Millet 等的分析还指出，自监督模型的成功不仅依赖大规模训练数据，还受数据多样性和音频质量的影响。他们训练了不同尺寸的模型，发现 50 小时的数据即可学习基本的声学表示，但在更高语音层级（如音节、韵律）上需要数百小时的数据才能达到与大脑类似的层级专化。此外，作者比较了在单个语言和混合语言数据上训练的模型，发现跨语言训练的模型能更好地泛化到新语言并保持与大脑的高匹配度。这意味着，模型在学习语音规律时可能捕获了普遍的声学约束而非特定语言的词汇规则。他们还探索了监督学习

模型（如声学—词法一体化模型），发现这类模型在高层表示上过度偏向目标任务（如字符识别），与大脑的匹配度反而下降。因此，自监督学习在模拟人类语言习得方面具有优势，未来可将其推广到更复杂的音系与语调结构，并与基于预测的语言模型整合。

自监督语音模型的优势不仅在于不依赖人工标签，还在于能够捕获语言通用的韵律和声学特征。wav2vec 2.0 的特征编码器由七个卷积模块组成，每个模块通过步长和卷积核大小控制时间分辨率和特征维度。在训练过程中，模型通过对比学习鼓励不同语音片段具有辨别性，同时利用量化模块构建离散的代码本，使连续语音表示能够映射到有限的符号集合。这些符号类似于音素或音节的抽象单位，为后续上下文网络提供清晰的离散输入。研究表明，自监督模型的不同层在处理语音的各个阶段表现出特定功能：早层卷积层对短时间帧的频谱特征敏感，中层表示音节和共振峰结构，后层 Transformer 捕获语调、词汇和发音风格。相比之下，监督训练的声学模型往往过度优化于特定标注任务，如语音识别或声学模型，从而导致其内部表示失去通用性，无法很好地解释大脑数据。通过在多种语言和方言上训练自监督模型，可以构建语言无关的声学基底，然后在此基础上微调特定语言的声学和语音特征。这一策略有望缩小模型与大脑在语音处理上的差距，特别是在处理口音、方言和情绪时。

4.2 Transformer 的计算与功能专化

Transformer 通过多层的自注意机制处理序列信息。此前研究多关注模型的隐向量嵌入，而 Kumar 等（2024）[5] 直接分析了注意头执行的变换，即每个注意头如何更新词表示。他们将模型的变换分解为每个头的线性变换，并使用自然听故事的 fMRI 数据评估其预测能力。结果显示，注意头变换在语言皮层大部分区域能解释大量方差，且在后颞区显著优于传统语言学特征（如词性或句法依存关系）。不同层的头展示出渐进式的功能梯度：早层、短距离回溯的头权重在后侧颞区更高，而高层、长距离回溯的头则在前颞与前额叶区域占优势。这种梯度与语言处理从局部到全局的层次结构一致。

此外，作者发现某些特定头对特定语法依存关系（如补语从句、直接宾语）有选择性，在后颞区表现尤为明显。值得注意的是，在角回等高级语义区域，非上下文嵌入的预测能力反而超过了变换，这表明这些区域可能整合全局语义内容而非依赖局部注意。作者强调，头部变换只是模型的线性近似，不代表大脑的真实计算，但这种方法提供了更加细粒度的功能对齐视角。他们建议未来探索瓶颈 Transformer 等新架构，引入声学特征，并结合行为和语言任务，通过梯度约束和多任务学习进一步靠近大脑机制。

Transformer 的自注意框架不仅提供了线性近似，还可能在不同层捕获句法层次、共指链和话语主题等丰富信息。近期分析发现，注意头不仅沿层次形成回溯距离的梯度，还

沿功能类别发生分化。例如部分头专注于主谓一致、名词短语结构，另一些头捕获语气助词和焦点标记，反映语用提示。更重要的是，头与头之间存在协调机制：在处理长句或嵌套从句时，前层的局部头预先筛选相关修饰语，后层的全局头再根据语篇结构整合信息。实验表明，当人为阻塞关键注意头时，模型在脑编码任务上的预测性能会显著下降，这支持了头之间协同工作的观点。进一步的诊断显示，注意分布与人类阅读的眼动轨迹相关，高层注意与长距离回溯的阅读阶段吻合，说明模型学习到了与人类相似的注意模式。

未来研究可以在模型中纳入生物约束，如稀疏连接和时间常数，使注意机制更贴近突触动力学。例如可以采用动态稀疏化策略限制每层仅激活少数关键头，并让头的权重随时间衰减，模拟大脑对旧信息的遗忘。此外，可以结合符号化模块，将注意头分配给特定的句法功能或语义角色，实现可解释性更强的混合模型。鉴于不同头在不同语言和任务上的作用存在差异，跨语言分析注意模式将有助于识别普遍与特定的注意策略。通过这些改进，我们期望 Transformer 不仅在行为表现上接近人类，还能在内部计算上更符合大脑的层次组织。

此外，注意头的输出可视为一系列线性变换，将当前词向量映射到多个上下文子空间。研究显示，这些变换在处理隐喻、修辞重复和语调变化时表现出特定的模式，表明模型内部对语义和语用的编码比单纯词级向量更丰富。一些工作提出将注意机制与可微栈或记忆模块结合，捕捉跨句段乃至跨文档的长期依存。将这些结构嵌入语言模型，或可更好模拟大脑在保持情节线索和角色身份时的持久性。由于自注意计算的复杂度随序列长度平方增长，与大脑的时间和能量约束不符，近年出现的稀疏或线性注意机制能够在提高效率的同时保留建模能力。这些机制也许更贴近大脑的局部连接和长距离白质束传导特性。对 Transformer 计算的深入理解和重构，将推动下一代高效且生理友好的模型，为解释大脑语言加工提供更多线索。

除了自监督和预测式学习机制外，最近的研究还探索了对比学习、变换器自回归与去噪自编码器结合的框架，以更全面地模拟语言习得。例如，带噪预训练的文本编码器通过随机遮挡句子片段并要求模型复原原句，在学习恢复局部信息的同时，还培养了对全局语义和句法的理解。这种任务结合了预测和重建两个目标，既包含自底向上的模式提取，也包含自顶向下的生成假设。实验证明，这类模型在脑编码任务中表现优异，尤其在预测内侧前额叶和角回等高层语义区域的活动时超越传统自回归模型。另一些研究引入了元学习和持续学习框架，使模型能够在不断变化的语料环境中适应新的语言模式，同时保留旧知识。这与大脑在终生学习中不断更新语义地图的过程类似。将这些先进的学习策略与神经数据对比，不仅可以评估它们的生物合理性，还可以揭示大脑可能采用的学习策略组合。

5 组合语义与 supra-word 表征

语言理解不仅涉及单词的意义，还需要基于上下文组合词汇产生超越字面含义的“超词”（supra-word）意义。Toneva 等（2022）[11] 提出了一种数据驱动方法，利用 ELMo 模型 [7] 的前向 LSTM 隐状态构建上下文嵌入，并通过线性回归消除单独词义的贡献，得到仅包含组合信息的残差嵌入。这种嵌入能捕捉隐含意义，例如“Mary finished the apple”中隐含的“吃完苹果”或“绿色香蕉”表示“未熟香蕉”，也可表示新语义组合。

将 supra-word 嵌入作为预测变量，研究者用其解释 fMRI 和 MEG 数据，发现经典词汇枢纽，如颞上回后部和颞下回，同样维护组合语义。此外，前颞叶也对 supra-word 嵌入敏感，说明词汇与组合语义在大脑中共享神经基础，需要前后颞区域协同维持。然而，他们在 MEG 数据中未检测到 supra-word 表征。这提示 supra-word 意义可能通过持续的、非同步的神经活动体现，而 MEG 对同步活动敏感，难以捕捉这种慢速信号。该结果强调不同脑成像技术对语义组合的敏感性不同，未来需要结合低频功率或相位同步等指标并使用多模态数据共同分析。

值得一提的是，supra-word 嵌入不仅揭示了组合语义的存在，还证明了分布式向量可以通过残差运算表示复杂的语义组合。这种方法与传统基于语法树的组合不同，它不依赖手工定义的规则，而是由模型在大量语料中学习到的统计规律。研究人员发现，将 supra-word 嵌入与词级嵌入结合，可以更准确地预测读者对隐含意义的理解程度。这一发现为改进语言模型提供了启示：未来可尝试不同的向量运算（如加权平均、向量差、张量积）以及递归或注意机制，来模拟人类如何积累和整合组合意义。此外，在神经数据中考察 supra-word 作用的时间动态是一个开放问题，未来可通过在 MEG 或 ECoG 数据中分析低频功率或相位同步来捕捉持续性组合信号，并结合工作记忆和注意任务，探讨语义组合与认知资源的关系。

除了 ELMo 生成的残差嵌入外，后续研究还探索了多种模型和运算来捕捉组合语义。例如，通过比较 GPT-2[8] 或 BERT[3] 的上下文嵌入与各单词嵌入的向量差，可以获取另一种 supra-word 表征，这些表征在脑编码任务中的表现与 ELMo 残差相当。另一个方向是使用张量积或基于张量神经网络的方法，将两个词向量结合为高阶张量，从而显式表示交互项。虽然这些方法在计算上更昂贵，但它们提供了更丰富的组合信息。研究还发现，组合语义的表征不局限于双词短语，复杂句子中的嵌套关系和修饰语也可以通过递归地应用残差运算来分解。实验表明，这种递归嵌套的 supra-word 嵌入能够更好地预测

人类对歧义句子的理解方式，说明模型对多义性消解有所把握。

组合语义不仅依赖词汇的线性组合，还与语法结构和语用知识紧密相关。在一些语言中，形态变化（如词序倒装、语气助词）会改变句子意味和语域，模型需要学习这些语法标记如何影响组合意义。此外，跨语言比较发现汉语、日语等语言的复合词和固定搭配大量依赖语法化的组合规则，这使得 supra-word 表征在这些语言中可能具有不同的统计特征。未来研究应扩展到不同语言的组合语义，检验模型能否捕获这些语言特有的组合规律。为了更精确地测量 supra-word 处理的神经时间动态，可以利用 ECoG 或高时间分辨率 fMRI，结合语音停顿和语调变化，分析大脑如何在听话者缓慢而连续的输入中形成语义组合。总之，supra-word 研究为我们揭示了词汇组合的复杂性，提醒语言理解是一个多层次、多维度的动态过程，需要模型和神经数据同时考虑语法、语义和语用因素。

6 逆向工程：从神经信号重构连续语义

除了“编码”（从刺激预测脑反应），近年来研究也开始系统推进“解码”（从脑反应重构刺激）与“逆向工程”。Tang 等（2023）[10] 展示了利用非侵入式 fMRI 在连续叙事场景下重构语义内容的可行性：模型并非逐字逐句复原原文，而是重构与原叙事在语义上高度相近的表达。这一结果提示，大脑在自然语境下的表征更接近“意义层级”的压缩，而不是对词形的逐点记录。

在方法上，解码任务常被形式化为寻找最可能的文本序列 S：

$$\hat{S} = \arg\max_{S} P(S \mid R) \propto \arg\max_{S} P(R \mid S)\, P(S),$$

其中 R 表示脑反应（如 fMRI 体素时间序列），$P(R \mid S)$ 对应“编码模型”，而 $P(S)$ 则提供语言先验（可由语言模型给出）。在实际求解中，研究者通常在候选空间内采用波束搜索（beam search）等策略，平衡“脑一致性”与“语言可行性”。这类工作也带来重要的科学与伦理问题：它既为检验语义系统的分布式表征提供了新的量化指标，也推动了对脑数据隐私与可解释性的讨论。

7 方法论：编码模型、评价指标与统计检验

7.1 编码模型：岭回归与分带岭回归

当以高维模型表征作为特征时，脑编码模型往往面临 $P \gg N$ 的病态问题，因此线性回归通常配合正则化使用。最常见的形式是岭回归：

$$\hat{\beta} = (X^\top X + \lambda I)^{-1} X^\top y,$$

其中 X 为特征矩阵、y 为神经响应、λ 为正则化系数。在特征由多子空间构成（例如声学特征与语义特征）时，还可使用分带岭回归（banded ridge regression），为不同子空间分配不同的正则化强度，从而进行方差分解并更清晰地区分不同信息源的独立贡献。

7.2 噪声天花板与显著性检验

由于神经测量本身存在噪声，模型预测的相关性上限受到“噪声天花板”（noise ceiling）约束。实践中常以重复试次的一致性或跨被试一致性估计可解释方差，从而避免对模型性能的过度解读。对单体素或单通道的显著性判断，则常结合置换检验（permutation test）构建零分布，以控制多重比较并提升结论的稳健性。

8 跨语言与跨模态比较

跨语言比较有助于检验模型与大脑匹配的普适性。Millet 等纳入英语、法语和普通话受试者，发现 wav2vec 2.0 自监督模型在不同语言中的层次映射非常一致：卷积层对应基本声学特征，中层对应语音特定特征，后期 Transformer 层对应语言特定信息。这一跨语言一致性表明，自监督模型捕捉了普遍的声学和语音规律，并且大脑对不同语言的加工共享相似的功能层级。

跨语言研究还应考虑语言类型学和文化差异。一些初步工作将语义地图方法应用于西班牙语、日语等非印欧语系，发现基本的语义区域位置类似，但某些文化特定概念（如

礼貌等级、敬语、宗教词汇）在皮层中的激活强度存在差异。这可能反映不同语言在语义编码上的策略，以及语言经验对神经表征的塑造。此外，跨语言比较可以揭示不同书写系统对语义加工的影响。例如，表意文字（如汉字）阅读者更依赖视觉形状与语义联想，而表音文字阅读者更依赖声音规则。大脑在这两类文字的处理过程中激活的视觉和语音区域有所不同，这些差异应在模型评估中加以考虑。未来扩展到双语者和方言使用者，将有助于理解语言经验对语义系统的可塑性和适应性。

跨模态比较则揭示不同成像技术对语言处理的敏感性差异。Schrimpf 等的集成建模同时分析了 fMRI 和 ECoG 数据，发现 Transformer 模型对两种模态的预测能力高度一致。然而，Toneva 等发现 MEG 数据无法检测 supra-word 表征。这些对比强调，需要结合多模态数据以捕捉大脑语言处理的不同时间和空间尺度。例如，fMRI 能捕捉慢速血氧反应，适合发现持续性语义信息，而 MEG 则对快速电同步更敏感，适合捕捉即时处理。这些差异需在模型与大脑对齐时仔细考虑。

跨语言研究还强调了语言类型学的多样性对语义系统的塑造作用。黏着语（如土耳其语、芬兰语）通过在一个词上附加多个语素表达语法关系，大脑可能利用这些形态学线索在早期阶段预测词干和附加成分的组合。屈折语（如拉丁语、俄语）则通过词形变化表示时态和格标记，需要跨更长距离的依存关系解析。孤立语（如越南语、泰语）依靠语序和功能词表达语法，使得句子结构的预测依赖于对短语层级的掌握。这些差异影响了语言模型和大脑在预测下一词时使用的策略，也要求我们在模型评估中纳入多样的语言。

文化因素也通过语言使用频率、隐喻习惯和礼貌策略等影响语义系统。例如，某些文化中对亲属关系有精细的词汇区分，而在另一些文化中对颜色或味觉的词汇更为丰富。研究者在跨文化采样时发现，与家庭和情感相关的词在东亚文化中引发的皮层激活更为广泛，而与个人独立性相关的词在西方文化的前额叶激活更强。这些差异提醒我们，语义地图具有可塑性，会随着语言环境和社会规范调整。未来应在不同文化群体中重复语义映射和脑编码实验，构建跨文化的语义参考图谱。

跨模态比较的意义不仅在于技术差异，还在于不同信号反映的神经过程各有侧重。ECoG 捕捉皮层表面电位，能实时反映快速同步活动，适合研究词汇和音素边界的瞬时处理；fMRI 捕捉血氧变化，反映几秒内的总体活动，适合研究持续的语义和情节加工；MEG 介于二者之间，反映大规模神经群体的同步，但对深层结构的敏感性较低。通过在同一受试者身上同时采集或跨实验结合这些数据，可以构建时间分辨率和空间分辨率兼顾的动态语义模型。例如，可先利用 MEG 确定 supra-word 组合发生的时间窗，再利用 fMRI 确定具体位置。跨模态融合将带来更全面的语言加工图景，促进理论与模型的发展。

跨语言语义映射的进一步研究需要涵盖非印欧语言群体以及低资源语言，如非洲的班图语系、美洲的纳瓦荷语或澳大利亚的皮京语。这些语言在语音、形态和句法结构上具

有独特特征，如点击音、序列化动词或双数标记，可能导致不同的语义组织模式。通过将这些语言纳入语义地图项目，可以检验语义系统的普遍性，并为语言多样性保护提供科学依据。此外，随着语言接触和全球化的深入，多语言者的大脑可能展现出更加灵活的语义网络和预测策略。研究显示，精通多种语言的人在切换语言时能够快速调整语义激活模式，这反映出一种在语义空间中的动态抑制与增强机制。将这种多语灵活性纳入神经模型，不仅有助于理解双语脑如何管理多个词汇和语法系统，也能启发开发能够动态切换语境的人工模型。未来的跨语言研究应将样本扩展到不同的社会群体，包含方言和混合语言，如新加坡英语和斯普兰语，从而捕获语言演化和创新对语义系统的影响。

9 讨论：综合视角与未来方向

9.1 预测还是特征学习?

关于神经语言模型在大脑编码任务中成功的解释存在两种观点。一方面，集成建模显示模型的脑拟合度与下一词预测准确率高度相关。这一结果常被解读为大脑优化于预测未来输入。另一方面，层级分析与通用迁移指标则指出，模型的预测能力不一定决定脑编码性能，中间层或迁移能力更能解释大脑数据。因此，我们需要超越简单的二分法，认识到预测和特征学习可能共同作用：大脑在处理语言时既利用预测生成候选，还依赖丰富的统计结构进行解析。未来研究可以通过操纵模型的训练目标（例如仅训练下一词预测 vs. 联合其他任务）来评估这些因素在脑编码中的相对贡献。

9.2 分布式语义系统与网络化视角

语义地图和语义关系研究揭示，大脑语义系统并非由少数专门区域组成，而是由广泛的分布式网络构成。PrAGMATiC 地图揭示 DMN 内有多个语义亚区，且左右半球分布对称。Zhang 等进一步发现语义关系通过 DMN 的激活与前顶叶注意网络的抑制共同编码，说明抽象推理依赖网络的动态交互。理解语义系统需要关注这些网络的连接模式和时空动态，而不仅是个别皮层区域的选择性。未来可以利用功能连接分析或图论方法，探索语义网络在不同任务和状态下的重构与调节。

9.3 模型特性与大脑对应

神经模型的学习机制和架构对与大脑的对应性具有重要影响。自监督语音模型在少量未标记数据上就能学得与大脑相似的层级表示，强调自监督和对比学习的价值。Transformer 的注意头分析提供了更细粒度的功能对齐视角，揭示不同层和头在皮层中的对应关系。这些方法说明，大脑和模型可能共享某些计算原则，如局部与全局整合。但目前模型仍缺乏对生理约束的考虑。加入时间连续性、能量限制、发展过程等约束，可能使模型更符合大脑。

9.4 组合语义的新挑战

supra-word 研究指出，组合语义在大脑中与词汇语义共享神经基础但依赖不同的维持机制。当前的语言模型通常通过固定窗口产生静态嵌入，难以捕捉组合语义的动态特性。未来应探索结构化或递归模型，在模型中显式实现组合运算，并比较不同组合算法对脑数据的预测能力。此外，需要在 MEG 和 ECoG 中寻找捕捉组合语义的合适信号，例如低频功率或相位同步。

9.5 跨文化与多语言的扩展

现有研究主要使用英语和西方语言，受试者背景也相对单一。Huth 等指出不同个体间语义地图的相似性可能源自共同的生活经验；Zhang 等则发现抽象概念与右半球关联的强度可能受文化差异影响。因此，未来需要在不同文化和语言环境中采集数据，比较语义系统的共性与差异。研究应扩展到儿童、双语者或方言使用者，以揭示语义系统的发育和可塑性。这将帮助我们理解语言经验与神经组织如何交互，促进构建具备跨语言普遍性的模型。

讨论未来方向时，还需考虑神经语言模型与其他认知系统的交互。语言理解往往伴随记忆检索、情感评价和动作规划等过程，语义系统必须与海马回、杏仁核以及额顶网络协同工作。现有模型多数仅处理语言输入，缺乏与视觉和情感系统的互动。将视觉和情感通道融入模型，可以模拟故事理解中的场景想象和情绪反应。例如，电影描述不仅包含语言，还伴随画面和音乐，这些多模态刺激触发更复杂的语义及情感网络。开发能够同时处理文本、图像和声音的多模态模型，并用其预测多模态脑数据，将为全面理解自然语境下的语言加工奠定基础。

另一个重要方向是发展因果推断的方法。目前多数研究基于相关分析，难以确定模

型表示对大脑活动的因果作用。结合经颅磁刺激、脑损伤研究或神经反馈，可以测试某些模型表征是否必要或充分。例如，当模型预测出高语义突发点时，相关脑区是否必然会出现因果响应？机器学习中的可解释性工具（如特征重要性和层次可视化）也可与神经干预结合，验证特定特征在语义加工中的作用。通过这些方法，我们可以揭示模型与大脑之间的因果联系，而不仅是相关性。

此外，伦理和公平问题应贯穿研究全过程。语义系统与价值观和社会经验密切相关，训练数据中的偏见可能导致模型学习到有害的语义关联，例如性别刻板印象或种族偏见。在比较不同文化和语言时，也必须尊重多样性并避免文化中心主义。开发公平、透明且可解释的语言模型，并理解其与大脑语义系统的异同，将有助于构建包容、安全的人工智能系统。

另一个值得关注的领域是儿童语言学习与成人语言理解的差异。婴幼儿在缺乏明确监督的情况下，通过与环境互动自然习得语言，其学习过程可能更依赖于语音和语义的统计共生，而非明确的预测。研究表明，儿童的大脑在处理语言时更依赖音节层级和韵律模式，而成人则能利用句法和语义进行更高级的预测。当前神经语言模型主要基于成人语言语料训练，忽视了语言发展阶段的差异。未来可以通过训练自监督模型在儿童对话语料上，模拟语言的发展过程，并与不同年龄段的脑成像数据比较，揭示语言系统的成熟轨迹。此外，还应关注老年人的语言加工变化，研究老化和神经退行性疾病如何影响语义网络和预测机制。通过跨年龄的比较，我们可以建立更加全面的语言认知模型，为教育和康复提供科学依据。

从更广阔的角度看，语言研究与社会科学、哲学和人工智能伦理密切相关。语言不仅是沟通工具，也是思想和文化的载体。神经语言模型的广泛应用有可能影响公共舆论和社会认知结构，因此研究者需要审慎评估模型的社会影响。例如，语言模型生成的内容可能强化现有偏见或创造新的误解；模型的预测机制可能在某些文化中被误解为超越或取代人类思考。通过与哲学家、社会学家和伦理学家的合作，可以更深入理解语义系统与社会结构的相互作用，并制定适合不同文化环境的技术规范。科学家还应通过科普教育，让公众理解神经语言模型的能力与局限，建立合理期待，避免神话化技术。综观全局，语言的神经与计算研究既是科学探索，也是社会实践的一部分，必须兼顾科学价值和社会责任。

10 结论

神经语言模型与大脑语义表征的研究正在快速发展。从预测编码的争论到语义地图的绘制，从自监督学习到注意头的功能专化，这些成果逐渐揭示语言理解的多层次结构。一方面，大型 Transformer 模型的成功强调了预测过程在语言系统中的重要性；另一方面，层级和迁移分析提醒我们，模型与大脑的对应不只是预测，通用语言知识也起决定作用。数据驱动的语义地图和关系研究展示了大脑语义系统的分布式和网络化特点。自监督语音模型及其跨语言一致性为理解人类语言习得提供了新的工具；注意头分析为细粒度的脑—模型对齐提供了方法；组合语义研究则揭示了词汇和组合意义共享但依赖不同神经机制的复杂图景。

展望未来，仍有多重挑战等待解决：获取同时具有高时间和空间分辨率的神经数据，以捕捉语言加工的动态；发展能够整合多模态数据和多任务训练的模型；将发展限制、能耗约束等生物因素引入模型；以及在更广泛的文化与语言背景下验证研究的普适性。通过跨学科合作，结合神经影像学、计算神经科学、语言学和人工智能，我们将进一步接近理解大脑语言机制的真实样貌。

总而言之，将神经语言模型与大脑语义系统进行比较不仅能深化我们对语言机制的理解，还能为人工智能提供新方向。当前证据显示，预测机制、分布式表征、自监督学习和组合语义是构建符合人脑的语言模型不可或缺的元素。未来，我们倡议采用更广泛的语言和文化样本，结合多种神经成像技术和因果干预方法，实现真正的跨学科融合。只有这样，我们才能在知识、技术与伦理层面推动人工智能与人类智能的和谐发展。

综上所述，跨层次、跨模态和跨文化的综合研究将成为未来语义科学的重要方向，为构建真正符合人类思维的智能系统奠定基础。这一目标需要持续努力。

值得一提的是，生成式模型和判别式模型在解释语言理解方面可能各具优势。生成式模型通过显式建模上下文和潜在语义结构，可以模拟听者在理解故事时生成的内部假设和场景想象；判别式模型则侧重于区分候选输出，更适合快速决策和分类任务。大脑可能同时利用两种策略：在构建对叙事的宏观理解时采用生成式推理，而在词汇辨认和句法解析时采用判别式决策。这一观点得到神经证据支持：在默认模式网络和海马区检测到与生成式推理相关的活动，而在颞上回和顶叶注意网络则检测到与判别式加工相关的活动。未来可通过比较生成式预训练模型（如扩散语言模型）和判别式模型在脑编码任务上的表现，探索这两种策略与不同脑区之间的映射关系。同时，在模型训练中引入生成与判别的联合目标，或许能构建更符合大脑处理的混合模型。

另一方面，语言理解与时间意识和未来计划的关系也值得关注。当人们听故事时，他们会构建情节的时间轴，并预测未来事件的走向，这涉及前额叶皮层与海马网络的协同。预测编码模型主要关注短期词序列，而忽略了跨句或跨段落的事件预测。神经语言模型

可以通过引入时间标记和事件链建模任务来更好地模拟这种长程预测能力。例如，可以训练模型根据当前叙事生成可能的下一段情节，或预测角色的后续行为，并用这些预测表示去解释大脑中与情节理解相关的活动。这不仅拓展了模型的功能，也为解释大脑如何处理跨句篇章层次的预测提供了新线索。

另一方面，未来研究还可以考虑将道德推理和价值判断融入语言模型的对齐过程。人类在理解语言时常常自动评估行为的道德性、陈述的可信度和信息的社会意义，这些评估依赖于前额叶皮层和边缘系统的互动。现有语言模型普遍缺乏对道德语义和价值判断的敏感性，通过引入道德推理任务或使用包含道德标注的语料进行微调，可以探索模型在这些方面与大脑的差距。此外，在脑编码实验中加入涉及伦理决策或道德评价的故事，可以观察语义系统与情感和价值网络的协同反应，为发展更负责任的人工智能提供依据。随着深度学习在自然语言处理领域的不断进步，模型与大脑比较的工具也在持续发展。例如，最近的研究使用可微分的哈密顿动力学模拟语言产生过程，将语义和语法编码为物理系统的能量和势能；这种方法在生成语音时显著降低了模型复杂度，并能够捕捉到语调和节奏的微妙变化。与传统神经网络相比，物理启发模型更容易引入生理约束，如能量消耗和处理速度限制，为理解大脑如何高效地编码和生成语言提供新思路。未来可以将这类模型应用于脑解码任务，检验其是否更符合语音与语义的神经表示，同时探索大脑是否利用类似的动力机制来组织语言信息。

此外，针对语义地图和语义关系的研究可以进一步采用图神经网络（GNN）等结构化模型来模拟语义网络的连通性。GNN 能够在节点和边的结构上进行学习，非常适合处理语义图和知识图谱。将 GNN 应用于脑编码任务，可能揭示大脑中概念之间的连接强度和传播路径，进而解释概念激活在时间上的扩散过程。这种结构化建模方式也为跨学科合作打开新途径，将语言学、神经科学与图论结合起来，为理解语义认知的复杂结构提供强大的方法工具。综合这些研究，我们也应认识到当前方法的局限性。首先，大部分神经语言模型只考虑文本或语音输入，忽略了手势、面部表情和语调等非言语信号，这些信号在实际沟通中具有重要作用。多模态语料和模型训练仍处于起步阶段，难以捕捉语言与其他感官输入的交互。其次，脑编码模型主要基于线性回归，无法捕捉大脑和模型表征之间可能存在的非线性映射。尽管线性模型易于解释和训练，但未来应探索深度编码模型或表示对齐技术，以更精准地映射复杂的语义关系。此外，现有研究往往忽视个体差异，如工作记忆容量、阅读习惯、知识背景等，这些因素都会影响语义加工和预测能力。只有在样本量更大、参与者更多样化的情况下，才可能揭示语义系统的共性与个体差异。最后，各研究的数据共享和分析流程尚未完全标准化，重复性和可比性有待提高，倡议建立开放的语料库、共享代码和规范化分析平台，以促进该领域健康发展。

参考文献

[1] Richard Antonello and Alexander G. Huth. Predictive coding or just feature discovery? an alternative account of why language models fit brain data. *Neurobiology of Language*, 2023.

[2] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*, 2020.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019.

[4] Alexander G. Huth, Wendy A. De Heer, Thomas L. Griffiths, Frédéric E. Theunissen, and Jack L. Gallant. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, 532(7600):453–458, 2016.

[5] Sreejan Kumar, Theodore R. Sumers, Tamar Yamakoshi, Ariel Goldstein, Uri Hasson, Kenneth A. Norman, Thomas L. Griffiths, Robert D. Hawkins, and Samuel A. Nastase. Shared functional specialization in transformer-based language models and the human brain. *Nature Communications*, 2024.

[6] Juliette Millet, Charlotte Caucheteux, Yves Boubenec, Alexandre Gramfort, Ewan Dunbar, Christophe Pallier, and Jean-Rémi King. Toward a realistic model of speech processing in the brain with self-supervised learning. *Advances in Neural Information Processing Systems*, 2023.

[7] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2018.

[8] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. OpenAI Technical Report, 2019.

[9] Martin Schrimpf, Idan A. Blank, Greta Tuckute, Carina Kauf, Eghbal A. Hosseini, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45), 2021.

[10] Jerry Tang, Alexandre LeBel, Shailee Jain, and Alexander G. Huth. Semantic reconstruction of continuous language from non-invasive brain recordings. *Nature Neuroscience*, 26:858–866, 2023.

[11] Mariya Toneva and Leila Wehbe. Combining computational controls with natural text reveals aspects of meaning composition. *Nature Computational Science*, 2:745–757, 2022.

[12] Yanchao Zhang, Alec Tetreault, Yaling Xu, John A. Pyles, and Michael J. Tarr. Connecting concepts in the brain by mapping cortical representations of semantic relations. *Nature Communications*, 11:1–13, 2020.