东吴证券 SOOCHOW SECURITIES

汽车行业深度报告

AI 系列深度 16 期：语言智能为何从 RNN 转向 Transformer？

2026 年 08 月 06 日

增持（维持）

证券分析师 黄细里
执业证书：S0600520010001
021-60199793
huangxl@dwzq.com.cn

投资要点

- **语言智能从 RNN 走向 Transformer 的本质，**是序列建模从“递归记忆”转向“全局注意力”，并由此打开并行训练、规模扩展和任务统一的空间。RNN 通过隐藏状态沿时间步传递上下文，适合文本、语音等顺序数据，并在 LSTM/GRU、Seq2Seq 和 Attention-RNN 推动下，于 2014—2016 年达到机器翻译等任务的工程高峰；但其串行计算、长距离依赖衰减和固定向量瓶颈，决定了递归范式很难继续承载大模型时代的规模化路线。

- **Transformer 的关键变化，**不在于设计更复杂的记忆单元，而在于以自注意力让序列中任意位置直接交互，使语言建模从受串行结构约束的算法问题，转变为可随算力和数据扩展的工程问题。其优势集中体现在训练可并行、长距离依赖建模更强、结构更易堆叠扩展、归纳偏置更弱并更能释放大规模数据价值；相应代价是长序列计算成本、位置编码和小数据条件下先验不足等问题，这也是循环类结构在部分效率场景仍会延续的原因。

- **我们以论文与代表成果为锚，将语言 Transformer 演进归纳为七阶段、三大阶段：**2017 年自注意力替代递归，完成架构确立；2018—2020 年，BERT、GPT、T5/BART 推动预训练—微调、理解与生成任务统一以及规模律验证，语言模型从特征提取器转向任务执行器；2021 年后，指令微调、RLHF、CoT、RAG、工具调用、o1 和 DeepSeek-R1 等成果继续叠加，使主线从单纯扩大预训练规模，走向后训练优化与测试时计算。

- **Transformer 出现后，语言领域的研究和产业范式同步改变：**研究主线从为每个任务设计专用架构，转向预训练通用底座加下游适配，再转向以对话式通用模型响应多类任务；产业形态则因“互动性”能力发生质变。相较 AI1.0 时代 CNN 输出标签、检测框等感知性中间结果，语言模型可以直接通过自然语言与用户交互，形成反复使用的入口，ChatGPT 由此成为 AI 第二次爆发的产品拐点。

- **我们认为，数字 AI 底座模型发展路径已逐渐清晰、价值在于确定性，但其难以实现物理世界的建模闭环；我们认为物理 AI 将成为物理世界的 AI 解决方案，从而在 AI 的当前发展阶段成为增量更大的方向，智能驾驶作为最先跑通闭环的代表场景，我们认为应重点关注。**

- **风险提示：**技术迭代不及预期的风险，大模型厂商 ARR 增速放缓的风险，云服务商资本开支不及预期的风险，智能驾驶及机器人商业化落地不及预期的风险。

行业走势

相关研究

《AI 系列深度 15 期：Transformer 如何开启 AI 第二次爆发？》

2026-08-06

《AI 系列深度 14 期：深度学习如何开启 AI 第一次爆发？》

2026-08-05

内容目录

图表目录

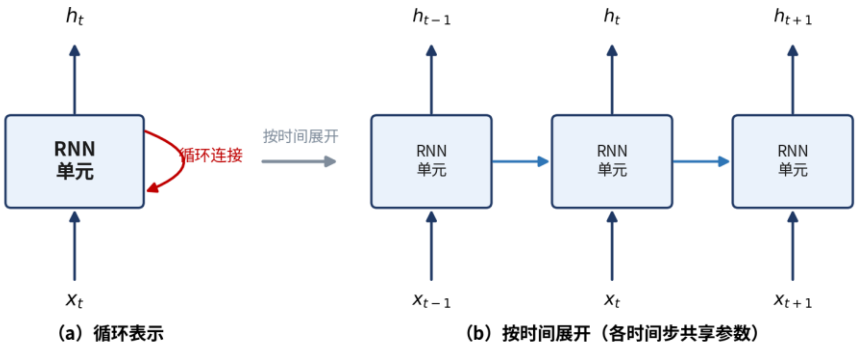
1. 复盘 RNN：递归结构如何定义语言任务上限

1.1. 递归结构：用隐藏状态承载上下文

循环神经网络（Recurrent Neural Network，RNN）是一类专门处理序列数据的神经网络，典型场景包括文本、语音和时间序列。其核心在于引入“循环”结构：模型处理当前位置时，不仅接收当前输入，也会继承上一时刻形成的隐藏状态，并在此基础上更新当前状态。换言之，RNN 通过同一套参数在序列各位置反复运算，把历史信息压缩进隐藏状态中，再沿时间维度逐步传递。该设计使模型能够用固定规模的参数处理不同长度的序列，并在一定程度上保存上下文信息。将这一递归过程沿时间展开，RNN 可理解为一个深度随序列长度增加、但各层共享参数的前馈网络。

与同期卷积神经网络相比，RNN 的优势来自对“顺序”的强假设。它默认数据不是一组彼此独立的点，而是像句子一样前后相连：前一个词会影响后一个词，同一套理解规则也可以在不同位置重复使用。因此，在文本、语音等天然有顺序的数据上，RNN 与任务形态更匹配。

图1：循环神经网络的循环单元与按时间展开示意



数据来源：《Finding Structure in Time》，《Backpropagation Through Time: What It Does and How to Do It》等其他文献，东吴证券研究所绘制

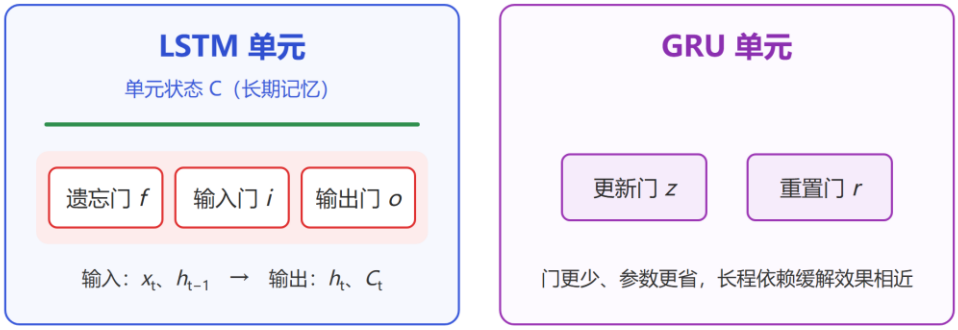
1.2. 从简单循环网络到门控机制与序列到序列框架

RNN 的思想可追溯至 1980 年代。Jordan（1986）与 Elman（1990）先后提出带反馈连接的网络结构，奠定了“用隐藏状态承载上下文”的基本范式；Werbos（1990）系统化了沿时间反向传播训练算法。其后，Bengio 等（1994）从理论上指出，简单 RNN 在信息跨多个时间步传递时容易出现梯度消失或梯度爆炸，导致模型很难记住较早位置的信息。针对这一问题，Hochreiter 与 Schmidhuber（1997）提出长短期记忆网络（LSTM），通过输入门、遗忘门、输出门与单元状态，控制信息“写入、保留、读出”；Cho 等（2014）提出结构更简的门控循环单元（GRU）。门控机制相当于为模型增加一套记忆开关，显

著缓解了长程依赖问题，也使 RNN 在 2010 年代成为序列建模的主流方法。

在门控机制基础上，RNN 进一步向深层化、双向化发展，并在机器翻译中形成序列到序列（Sequence-to-Sequence，Seq2Seq）框架：编码器先把源语言句子压缩成一个向量表示，解码器再据此逐步生成目标语言句子。可以把它理解为“先读懂原文，再一句句写出译文”。我们将 RNN 在语言任务上的演进路径归纳为：反馈网络与隐藏状态记忆思想（1986—1990）→BPTT 训练算法与长程依赖瓶颈的提出（1990—1994）→LSTM/GRU 门控突破（1997—2014）→深层化、双向化→Seq2Seq 与注意力机制推动机器翻译（2014—2016）。

图2：长短期记忆网络（LSTM）与门控循环单元（GRU）的门控结构示意



数据来源：《Long Short-Term Memory》，《Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation》，东吴证券研究所绘制

1.3. 结构性局限

尽管门控机制延长了 RNN 的有效记忆，但其底层递归结构仍带来三项约束，决定了后续架构替换的方向。

其一，串行计算、难以并行。由于当前隐藏状态依赖上一时刻隐藏状态，模型必须沿时间步逐个处理序列，前一时刻计算完成后，后一时刻才能继续。这使 RNN 难以充分利用 GPU 的大规模并行能力，训练与推理耗时随序列长度线性增加，进而制约模型和数据规模的继续扩张。

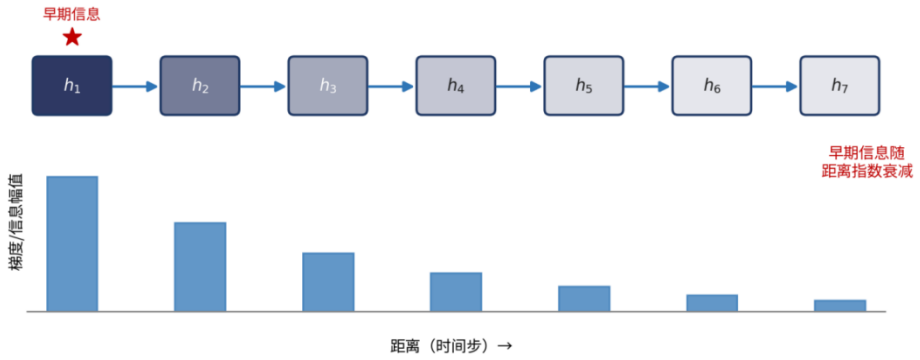
其二，长距离依赖容易衰减。信息在隐藏状态中经过多个时间步传递时，会逐渐被稀释、覆盖或遗忘。LSTM、GRU 通过门控机制缓解了这一问题，但没有完全解决；当依赖跨度很大时，句首或早期信息仍可能在传递过程中丢失。

其三，固定向量带来信息瓶颈。在标准 Seq2Seq 框架中，编码器需要把整段源序列压缩成一个固定长度向量，再交给解码器使用。源序列越长，这个向量越像“过小的行李箱”，越难完整装下全部信息，长句翻译质量也更容易下降。这一瓶颈直接催生了注意

力机制。

我们认为，“递归”让 RNN 具备处理序列的能力，但也通过串行计算、长程衰减与固定向量瓶颈限定了能力上限。理解这三点，是理解 Transformer 为何取代 RNN 成为语言模型主线的前提。

图3：梯度消失与长距离依赖衰减示意



数据来源：《Learning Long-Term Dependencies with Gradient Descent is Difficult》，东吴证券研究所绘制

2. Transformer 爆发之前，循环网络的应用上限

2.1. 循环网络时代语言任务的主要成果

2014—2016 年是 RNN 在自然语言处理（NLP）领域的高峰期，代表性进展集中在机器翻译。Sutskever 等（2014）提出的 Seq2Seq 框架，首次以端到端神经网络完成机器翻译，减少了统计机器翻译对人工特征和短语对齐流程的依赖。针对固定向量瓶颈，Bahdanau 等（2014）提出注意力机制，使解码器在生成每个目标词时，可以对源句不同位置动态分配权重，相当于翻译时有选择地“回看原文”；Luong 等（2015）进一步给出多种注意力计算形式。2016 年，谷歌发布神经机器翻译系统，以深层 LSTM 叠加注意力机制构建工程化系统并上线，标志 RNN 加注意力路线在机器翻译上达到工业级成熟。

机器翻译之外，RNN 在多类语言任务上也曾是主流方法。在序列标注任务中，如命名实体识别和词性标注，双向 LSTM 叠加条件随机场（BiLSTM-CRF）成为标准范式；在语言建模中，LSTM 长期占据困惑度指标的领先位置；在语音识别中，LSTM 配合连接时序分类（CTC）得到广泛应用；此外，文本分类、阅读理解与早期问答系统中也大量使用 RNN。

表1：循环网络在主要语言任务上的代表方法

任务方向	代表方法	主要时期
机器翻译	Seq2Seq、Attention、GNMT	2014—2016
序列标注（NER/词性标注）	BiLSTM-CRF	2015—2016
语言建模	LSTM 语言模型	2014—2016
语音识别	LSTM+CTC	2014—2016
文本分类与情感分析	LSTM/GRU、TextRNN	2015—2016
阅读理解与问答	注意力式阅读模型	2015—2016

数据来源：《Sequence to Sequence Learning with Neural Networks》，《Neural Machine Translation by Jointly Learning to Align and Translate》等其他文献，东吴证券研究所整理

2.2. 应用上限：工程高峰已至，规模化与任务统一受限

我们认为，RNN 把语言任务推进到“递归范式所能达到的工程高峰”，但其上限也由结构性局限决定，主要体现为四点。第一，在有监督、单任务、中短序列条件下，RNN 可达到当时最优水平，但在长文档、长程依赖场景下能力明显受限。第二，不同任务缺乏统一可复用底座，翻译、标注、问答等通常要分别设计网络并从头训练，迁移与复用能力有限。第三，串行计算限制了模型规模与训练数据规模的同步扩张，难以持续依靠“加大模型、加多数据”获得收益。第四，注意力机制已被引入作为 RNN 补充，这一事实本身说明，纯递归结构在长距离关系建模上已显不足，全局关联能力主要来自注意力。

换言之，RNN 加注意力的组合，已把语言任务推到递归范式的工程上限；但规模化与任务统一化两条道路，分别受到串行计算和任务专用性的限制。下一代架构需要解决的问题由此变得清晰：保留并强化注意力，同时摆脱对递归的依赖。

3. Transformer 与循环网络的结构对比及其优点

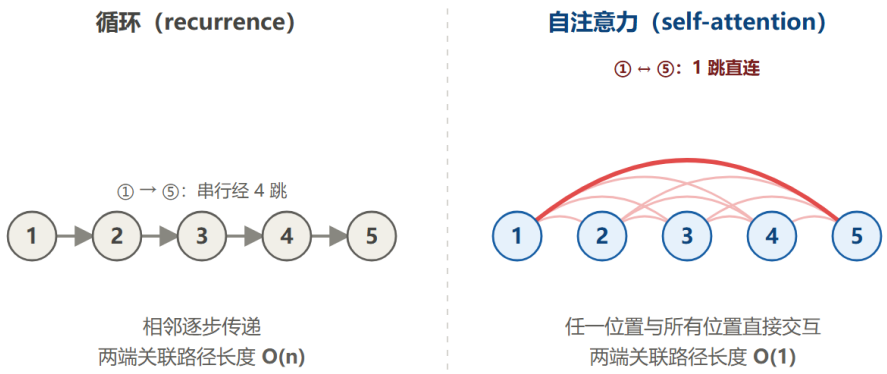
3.1. 两种架构的结构对比

Transformer 最初作为机器翻译架构提出，其与 RNN 最根本的差异，在于序列内信息如何交互。RNN 沿时间步逐次传递信息，任意两个位置之间的关系需要经过多个中间步骤；Transformer 则通过自注意力机制，让序列中每个位置直接与所有其他位置交互。具体而言，模型会为每个位置生成查询（Query）、键（Key）、值（Value）三组向量：查询像“我要找什么”，键像“我能提供什么线索”，值则是“真正要取回的信息”。模型根据查询与各位置键的相似度计算注意力权重，再对值加权求和，从而在单层内建立全局关联。

由于自注意力本身不包含顺序信息，Transformer 还需要通过位置编码显式注入位置信号，并以多头注意力从多个子空间并行捕捉不同关系，再叠加前馈网络、残差连接与

层归一化，构成完整的编码器—解码器结构。此后，该结构进一步分化出仅用编码器的 BERT 路线与仅用解码器的 GPT 路线。

图4：自注意力与循环结构的信息交互方式对比



数据来源：《Attention Is All You Need》，东吴证券研究所绘制

3.2. Transformer 相对循环网络的主要优点

基于结构差异，我们将 Transformer 相对 RNN 的优势归纳为四点，核心在于更适合并行训练、长距离建模和规模化扩展。

第一，训练可并行。RNN 必须沿时间步依次计算，而自注意力对序列各位置的计算相互独立，训练时可以一次性处理整条序列。这使 Transformer 更充分地利用 GPU 算力，缩短训练时间，并支撑更大规模的模型与数据。

第二，长距离依赖建模更强。在自注意力中，任意两个位置之间可以直接建立联系，关联路径不随距离增加而拉长；而 RNN 中两位置之间的关系必须经过多个中间步骤传递，距离越远，信息衰减和传递成本越明显。

第三，可扩展性更强。Transformer 结构均一，便于堆叠、加深和加宽；配合并行训练后，更适合“扩大模型与数据规模”的路线，并成为此后规模律发挥作用的结构载体。规模律本身的机理由本系列第二篇集中讨论，本篇仅讨论其在语言任务中的转折意义。

第四，归纳偏置较弱。RNN 预设了较强的顺序结构假设，而 Transformer 不预设过强结构，主要依靠数据学习关系。在大规模预训练条件下，这更有利于释放数据规模的价值，也为跨任务迁移提供便利。

但客观来看，上述优势也伴随代价：自注意力的计算和显存开销会随序列长度增加而快速上升，处理超长序列时成本较高；由于缺少天然顺序先验，模型需要位置编码补足顺序信息；在数据规模较小时，较弱的归纳偏置也未必优于 RNN。

表2：Transformer 与循环网络的多维对比

对比维度	循环网络（含 LSTM/GRU）	Transformer
信息交互方式	沿时间步递归	自注意力，全局直接交互
训练并行性	串行，难以并行	全序列并行
两位置关联路径	O(n)，随距离增长	O(1)，常数级
计算复杂度	O(n)，随长度线性	$O(n^2)$，随长度平方
顺序信息	结构自带	需位置编码
归纳偏置	强（顺序假设）	弱（依赖数据学习）
可扩展性	受串行计算限制	利于规模扩展
推理显存	常数（隐藏状态）	随上下文增长（KV-cache）

数据来源：《Attention Is All You Need》，《Long Short-Term Memory》，东吴证券研究所整理

我们认为，Transformer 相对 RNN 的胜出，不在于设计了一个“更复杂的记忆单元”，而在于解除递归结构对并行训练和规模扩张的限制，使语言建模从受限于串行计算的算法问题，转变为可随算力与数据扩展的工程问题。

4. Transformer 应用于语言任务的学术成果梳理及发展阶段总结

我们以论文与代表性成果为基础，梳理 2017 年以来 Transformer 在语言任务中的学术演进，并对其发展阶段进行归纳。

4.1. 发展阶段框架

基于相关论文与代表成果复盘，我们将 Transformer 在语言任务中的演化归纳为七个阶段，并进一步归并为三个大阶段（表 3；阶段划分与归并为我们整理，非业界统一标准）。三个大阶段分别为：架构确立，即自注意力替代递归，确立可并行的序列建模架构；范式统一与规模化，即预训练—微调成为主流，理解与生成任务统一为文本到文本，自回归模型经规模扩展获得通用能力；对齐与推理，即指令微调与人类偏好对齐提升模型可用性，复杂推理、工具调用与测试时计算进一步拓展能力边界。

表3：Transformer 语言任务发展七阶段—三大阶段框架

大阶段	阶段	时间	该阶段在语言任务发展中的意义
架构确立	前史：Seq2Seq + Attention	2014—2016	为 Transformer 的编码器—解码器框架与注意力机制奠基
架构确立	1.Transformer 机器翻译架构	2017	最初是翻译架构而非通用大模型，突破点是“序列建模不再必须递归”
范式统一与规模化	2.预训练+微调成主范式	2018—2020	从“每个任务单独设计模型”转向“一个预训练模型迁移多任务”
范式统一与规模化	3.理解与生成任务统一	2019—2020	语言任务接口统一，“提示即任务描述”思想开始成形
范式统一与规模化	4.自回归大模型与规模律	2019—2022	语言模型从“特征提取器”变为“任务执行器”，提示开始替代微调
对齐与推理	5.指令微调与人类偏好对齐	2021—2023	ChatGPT 形态的基础：从“会补全文本”转向“会响应人类任务”
对齐与推理	6.复杂推理、工具与外接	2022—2024	语言模型从“语言生成器”变为“任务规划器与工具接口”
对齐与推理	7.测试时计算与推理模型	2024—2025	能力提升从扩大预训练，转向后训练优化与测试时计算

数据来源：《Attention Is All You Need》，《Language Models are Few-Shot Learners》等其他文献，东吴证券研究所整理

4.2. 里程碑成果与论文分类

在该发展阶段框架下，我们认为以下成果具有里程碑意义：《Attention Is All You Need》（2017，确立 Transformer 架构）、BERT（2018，确立预训练—微调范式的理解路线）、GPT-3（2020，以规模扩展实现少样本与上下文学习）、InstructGPT/RLHF（2022，奠定指令对齐与对话形态）、思维链 CoT（2022，开启显式推理）、o1 与 DeepSeek-R1（2024—2025，确立推理模型与测试时计算路线）。这些成果分别对应架构替换、范式确立、能力跨越、产品可用、推理外化和推理强化等关键转折。

表4：Transformer 语言任务重要成果分类表

大阶段	阶段	代表论文/成果	核心贡献
架构确立	前史	Seq2Seq 、 Bahdanau/Luong Attention	循环网络解决序列到序列任务，注意力解决固定向量瓶颈
架构确立	阶段一	Attention Is All You Need	以自注意力替代循环/卷积，提升并行训练能力
范式统一与规模化	阶段二	GPT-1 、 BERT 、 RoBERTa 、 XLNet、ELECTRA	Transformer 从任务模型变为通用语言表征模型
范式统一与规模化	阶段三	T5、BART	将分类、问答、摘要、翻译统一为文本到文本问题
范式统一与规模化	阶段四	GPT-2、GPT-3、Scaling Laws、Chinchilla、PaLM	仅解码器模型经规模扩展获得零样本、少样本与上下文学习能力
对齐与推理	阶段五	FLAN、InstructGPT/RLHF、Self-Instruct、Constitutional AI、DPO	使模型理解意图、遵循指令、生成有用回答
对齐与推理	阶段六	思维链 CoT 、 RAG 、 ReAct 、 Toolformer、PoT、GPT-4	将推理、检索、工具使用与代码执行结合
对齐与推理	阶段七	o1、DeepSeek-R1、测试时计算扩展	通过更长思考、采样、搜索与强化学习强化推理过程

数据来源：《Attention Is All You Need》，《Language Models are Few-Shot Learners》等其他文献，东吴证券研究所整理

4.3. 逐阶段梳理

前史（2014—2016）：Seq2Seq 与注意力。如第二节所述，Seq2Seq（2014）确立编码器—解码器框架，Bahdanau（2014）、Luong（2015）提出的注意力机制解决固定向量瓶颈。这一阶段为 Transformer 的编码器—解码器结构和注意力机制提供了直接思想来源。

阶段一（2017）：Transformer 作为机器翻译架构。Attention Is All You Need 以自注意力替代循环与卷积，显著提升并行训练能力。Transformer 最初是面向机器翻译的任务架构，并非一开始就是通用大模型；其突破点在于证明“序列建模不再必须依赖递归”。

阶段二（2018—2020）：预训练加微调成为 NLP 主范式。GPT-1（2018）、BERT（2018）及其后的 RoBERTa、XLNet、ELECTRA 等，将 Transformer 从单一任务模型升级为通用语言表征模型。其中，BERT 代表以编码器为主的理解路线，推动自然语言理解基准整体跃升；GPT 代表以解码器为主的生成路线，并最终演化为大语言模型主线。NLP 由此从“每个任务单独设计模型”，转向“一个预训练模型迁移到多个下游任务”。

阶段三（2019—2020）：理解与生成任务统一。T5（2019）与 BART（2019）将分类、问答、摘要、翻译等任务统一表述为“文本到文本”问题，使语言任务接口趋于一致，也为此后“提示即任务描述”的思想奠定形式基础。

阶段四（2019—2022）：自回归大语言模型与规模律。GPT-2（2019）、GPT-3（2020）

等仅解码器模型，通过持续扩大参数和数据规模，获得零样本、少样本与上下文学习能力；Scaling Laws（2020）、Chinchilla（2022）、PaLM（2022）等刻画了模型能力随规模变化的规律。这一阶段的关键意义在于，语言模型从“特征提取器”转变为“任务执行器”：能力不再主要通过参数微调调用，而是通过提示和示例在上下文中直接调用。

阶段五（2021—2023）：指令微调与人类偏好对齐。FLAN（2021）、InstructGPT/RLHF（2022）、Self-Instruct（2022）、Constitutional AI（2022）、DPO（2023）等方法，使模型不仅能续写文本，还能理解用户意图、遵循指令并生成有用回答。GPT-3 已具备较强能力，但不一定按用户意图使用这些能力；指令微调与基于人类反馈的强化学习（RLHF）解决的正是“能力可用性”和“与人类偏好一致性”问题，构成 ChatGPT 形态的基础。

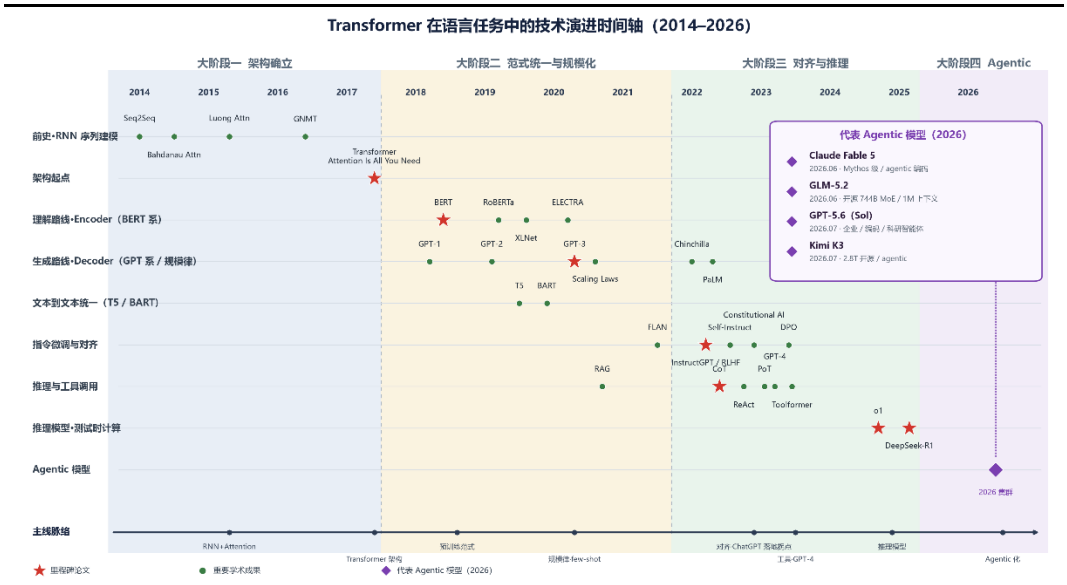
阶段六（2022—2024）：复杂推理、工具调用与知识外接。思维链（CoT）、检索增强生成（RAG）、ReAct、Toolformer、PoT 等方法，使模型把中间推理、外部检索、工具使用与代码执行结合起来；GPT-4（2023）在多任务能力上进一步提升。这一阶段表明，复杂语言任务不再只依赖模型权重一次性生成答案，而是开始外化为推理、检索和工具调用的组合，语言模型由“语言生成器”向“自然语言任务的规划器与工具接口”演变。

阶段七（2024—2025）：测试时计算与推理模型。o1、DeepSeek-R1 等推理模型，通过更长的思考过程、采样、搜索与强化学习强化推理轨迹，并在复杂任务上取得提升。其标志意义在于，能力提升的主线从单纯扩大预训练规模，转向后训练优化与测试时计算。我们提示，该路线仍在快速演化，结论不确定性相对较高。

4.4. 技术演进时间轴

将上述阶段与代表成果按时间和派系展开，可得 Transformer 在语言任务中的技术演进时间轴（图 5）。从主线看，RNN 加注意力（2014—2016）→Transformer 架构确立（2017）→预训练范式形成（2018）→规模律与少样本能力（2020）→指令对齐与 ChatGPT（2022）→工具调用与 GPT-4（2023）→推理模型（2024—2025），构成一条逐层叠加的升级链。从派系看，2017 年后，理解路线（BERT 系）与生成路线（GPT 系）从 Transformer 同源分化，文本到文本（T5/BART）将二者接口统一，对齐与推理两支则在生成路线之上继续叠加。我们认为，这条演进链的突出特征是“叠加”而非“替换”：每一阶段都建立在前一阶段基础之上，而非推倒重来。

图5：Transformer 在语言任务中的技术演进时间轴（2014—2025）



数据来源：《Attention Is All You Need》，《Language Models are Few-Shot Learners》等其他文献，东吴证券研究所整理

5. Transformer 之后：语言模型从专用任务走向通用入口

5.1. 研究与产业范式的变化

Transformer 出现后，语言领域在研究范式与产业形态上均发生显著变化。研究范式上，主线从“为每个任务设计专用架构与特征”，转向“预训练通用底座加下游适配”，并进一步转向“以一个对话式通用模型响应多类任务”；研究重心也从架构设计，转移到预训练数据与规模、指令对齐和推理策略。产业形态上，语言模型以对话助手形式直接面向终端用户，催生 ChatGPT 等新型产品，并带动国内外厂商密集投入大模型研发。

AI 第一次爆发的核心架构（以 CNN 为代表）提供的是“感知性”中间能力，即输出标签、检测框或特征后，还需嵌入既有业务流程才能产生价值，难以独立成为用户反复使用的交互入口，产品形态多表现为对既有功能的增强（1→n）。与之相对，Transformer 语言模型提供的是“互动性”能力：用户可以直接用自然语言与模型交互并获得可用回答，模型本身就构成反复使用的入口。我们认为，这一能力性质差异，是 AI 第二次爆发能够出现消费级现象级产品（0→1）的重要原因。

5.2. Transformer 是作为核心本体还是作为组件

在语言领域，Transformer 多以“核心本体”的方式被直接运用：GPT、BERT 等模型本身就是端到端承载任务的主干，输入直接是离散词元（token）序列，与 Transformer 输入形式高度契合，无需额外特征提取前端。这与计算机视觉领域形成对照：图像需要先切分为图块（patch）并线性映射为序列后才能输入 Transformer（如 ViT，Dosovitskiy

et al.，2020），且视觉 Transformer 常作为骨干网络嵌入检测、分割与多模态等更大流程，或与卷积结构混合使用。也就是说，在视觉中，Transformer 更多扮演“组成部分之一”的角色；在多模态方向，语言侧 Transformer 还常作为统一接口和中枢，将视觉等模态的编码器接入，由语言模型进行统筹。

我们认为，语言是 Transformer“作为本体直接运用”最彻底的领域。这一特征与语言数据的形式和任务形式高度一致；也正是这种“本体直用并兼具接口中枢”的地位，使语言模型成为当前通用人工智能的主要入口。

表5：Transformer 在语言与视觉任务中的角色对比

对比维度	语言任务	视觉任务
输入形式	天然离散词元序列，直接输入	图像须切分图块并序列化（ViT）
典型角色	本体/主干，端到端直接承载（GPT、BERT）	多作骨干网络组件，或与 CNN 混合
与其他结构关系	多为纯 Transformer	卷积与 Transformer 混合常见
多模态中的定位	常作统一接口与中枢，接入其他模态编码器	作为视觉编码器被语言中枢接入
本体化程度	高（最为彻底）	中（以组件化为主）

数据来源：《Attention Is All You Need》，《An Image is Worth 16 × 16 Words: Transformers for Image Recognition at Scale》等其他文献，东吴证券研究所整理

6. 风险提示

技术迭代不及预期的风险：数字 AI、物理 AI 技术进展低于预期，任务或应用场景拓宽速度低于预期。

大模型厂商 ARR 增速放缓的风险：若客户增长、续费率或客单价不及预期，模型厂商 ARR 增速可能放缓。

云服务商资本开支不及预期的风险：若自由现金流承压，或 AI 算力需求及投资回报低于预期，云服务商可能下调资本开支。

智能驾驶及机器人商业化落地不及预期的风险：若产品可靠性、成本下降或用户需求低于预期，相关产品商业化进程可能放缓。

免责声明

东吴证券股份有限公司经中国证券监督管理委员会批准，已具备证券投资咨询业务资格。

本研究报告仅供东吴证券股份有限公司（以下简称“本公司”）的客户使用。本公司不会因接收人收到本报告而视其为客户。在任何情况下，本报告中的信息或所表述的意见并不构成对任何人的投资建议，本公司及作者不对任何人因使用本报告中的内容所导致的任何后果负任何责任。任何形式的分享证券投资收益或者分担证券投资损失的书面或口头承诺均为无效。

在法律许可的情况下，东吴证券及其所属关联机构可能会持有报告中提到的公司所发行的证券并进行交易，还可能为这些公司提供投资银行服务或其他服务。

市场有风险，投资需谨慎。本报告是基于本公司分析师认为可靠且已公开的信息，本公司力求但不保证这些信息的准确性和完整性，也不保证文中观点或陈述不会发生任何变更，在不同时期，本公司可发出与本报告所载资料、意见及推测不一致的报告。

东吴证券投资评级标准

投资评级基于分析师对报告发布日后 6 至 12 个月内行业或公司回报潜力相对基准表现的预期（A 股市场基准为沪深 300 指数，香港市场基准为恒生指数，美国市场基准为标普 500 指数，新三板基准指数为三板成指（针对协议转让标的）或三板做市指数（针对做市转让标的），北交所基准指数为北证 50 指数），具体如下：

公司投资评级：

买入：预期未来 6 个月个股涨跌幅相对基准在 15%以上；

增持：预期未来 6 个月个股涨跌幅相对基准介于 5%与 15%之间；

中性：预期未来 6 个月个股涨跌幅相对基准介于-5%与 5%之间；

减持：预期未来 6 个月个股涨跌幅相对基准介于-15%与-5%之间；

卖出：预期未来 6 个月个股涨跌幅相对基准在-15%以下。

行业投资评级：

增持： 预期未来 6 个月内，行业指数相对强于基准 5%以上；

中性： 预期未来 6 个月内，行业指数相对基准-5%与 5%；

减持： 预期未来 6 个月内，行业指数相对弱于基准 5%以上。

我们在此提醒您，不同证券研究机构采用不同的评级术语及评级标准。我们采用的是相对评级体系，表示投资的相对比重建议。投资者买入或者卖出证券的决定应当充分考虑自身特定状况，如具体投资目的、财务状况以及特定需求等，并完整理解和使用本报告内容，不应视本报告为做出投资决策的唯一因素。

东吴证券研究所
苏州工业园区星阳街 5 号
邮政编码：215021
传真：（0512）62938527
公司网址： http://www.dwzq.com.cn