以下是对《Attention Is All You Need》（Transformer）、《BERT》与《GPT-3》三篇核心文献的横向对比分析，聚焦方法论/架构、训练目标及优势与局限三个维度。

三篇核心文献横向对比分析

对比维度	文献1: Attention Is All You Need Transformer	文献2: BERT	文献3: GPT-3
核心贡献	提出纯自注意力机制架构，替代RNN/CNN，实现高效并行化。	提出双向Transformer预训练模型，通过掩码语言模型（MLM）与下一句预测（NSP）实现深层双向表征。	证明超大规模语言模型（175B参数）在少样本/零样本学习中的强大能力，无需微调即可完成多种任务。
模型架构	编码器-解码器结构，均基于多头自注意力与位置编码。	仅Transformer编码器，双向自注意力，引入 CLS 与 SEP 特殊标记。	Transformer解码器架构，左到右自注意力，支持稀疏注意力以支持长文本。
训练目标	监督任务：机器翻译（编码-解码序列到序列学习）。	自监督预训练： 1. **掩码语言模型（MLM）**：预测随机掩码的词。 2. **下一句预测（NSP）**：判断两句是否连续。	自监督预训练：**自回归语言建模**（预测下一个词），专注于在上下文中学习任务。
训练数据	WMT 2014 英德、英法语料（监督）。	BooksCorpus + 英文维基百科（无标注）。	过滤后的CommonCrawl + WebText + 英文维基百科 + 书籍

对比维度	文献1: Attention Is All You Need Transformer	文献2: BERT	文献3: GPT-3
			（无标注，大规模）。
参数规模	Base: 65M；Big: 213M。	$BERT_{BASE}$: 110M；$BERT_{LARGE}$: 340M。	125M 至 175B（8个不同规模模型，GPT-3为175B）。
主要优势	1. **并行化训练**，效率远高于RNN。 2. **长距离依赖**建模能力强。 3. 为后续所有Transformer模型奠定基础。	1. **双向上下文理解**，优于单向模型。 2. **统一架构**，易迁移至多种NLP任务。 3. 在GLUE、SQuAD等任务上大幅提升SOTA。	1. **卓越的少样本/零样本泛化能力**。 2. 强大的文本生成与开放域问答能力。 3. 展示了模型规模持续提升的潜力。
主要局限	1. 主要针对翻译，未探索大规模预训练范式。 2. 计算复杂度随序列长度平方增长（On^2）。	1. MASK **标记与微调时的不匹配**问题。 2. 在生成任务上弱于自回归模型。 3. 大模型预训练与微调成本高。	1. **资源消耗巨大**，训练与部署成本极高。 2. **偏见与公平性**问题显著。 3. 存在潜在滥用风险（如生成虚假信息）。
最佳应用场景	需要高效处理长序列、并行化训练的序列到序列任务（如机器翻译）。	需要深度语言理解的任务（如文本分类、语义匹配、命名实体识别、阅读理解）。	需要快速适应新任务、强大生成能力或开放域知识问答的场景（尤其是少样本/零样本设定）。
范式	**基础架构革命**：引入自注意力机	**预训练范式确立**：将	**大模型与上下文学习探索**：推动

对比维度	文献1: Attention Is All You Need Transformer	文献2: BERT	文献3: GPT-3
演进定位	制，成为后续所有Transformer模型的基石。	Transformer编码器用于大规模双向预训练，成为NLP下游任务的主流范式。	语言模型向超大规模发展，并探索无需微调的上下文学习潜力。

分析总结

1. **方法论/架构差异**：从Transformer作为**基础模块**，到BERT专注**编码器**以进行双向理解，再到GPT-3采用**解码器**以强化自回归生成能力，三者体现了从通用架构到针对理解与生成任务的特化演进。
2. **训练目标/核心任务**：Transformer最初用于**有监督的序列翻译**。BERT通过**无监督预训练**（MLM+NSP）获取通用语言表征，以解决下游理解任务。GPT-3的核心任务也是**无监督预训练**（自回归LM），但其核心贡献在于评估此任务在**极少监督样本**下完成多种任务的能力。
3. **优势与局限**：
    - **Transformer** 优势在于其**并行化与长依赖建模**，局限在于未探索预训练范式。
    - **BERT** 优势是**统一架构与强大的双向理解能力**，使其成为理解类任务的“瑞士军刀”，但在生成任务上存在弱点。
    - **GPT-3** 的优势在于其**惊人的规模与少样本泛化能力**，展示了通向通用人工智能的一条可能路径，但其**资源成本与伦理风险**是当前的主要局限。

这三项工作共同构成了现代大语言模型发展的关键阶梯：Transformer提供了**核心架构**，BERT确立了**预训练范式**，而GPT-3则引爆了**大模型与上下文学习**的研究浪潮。