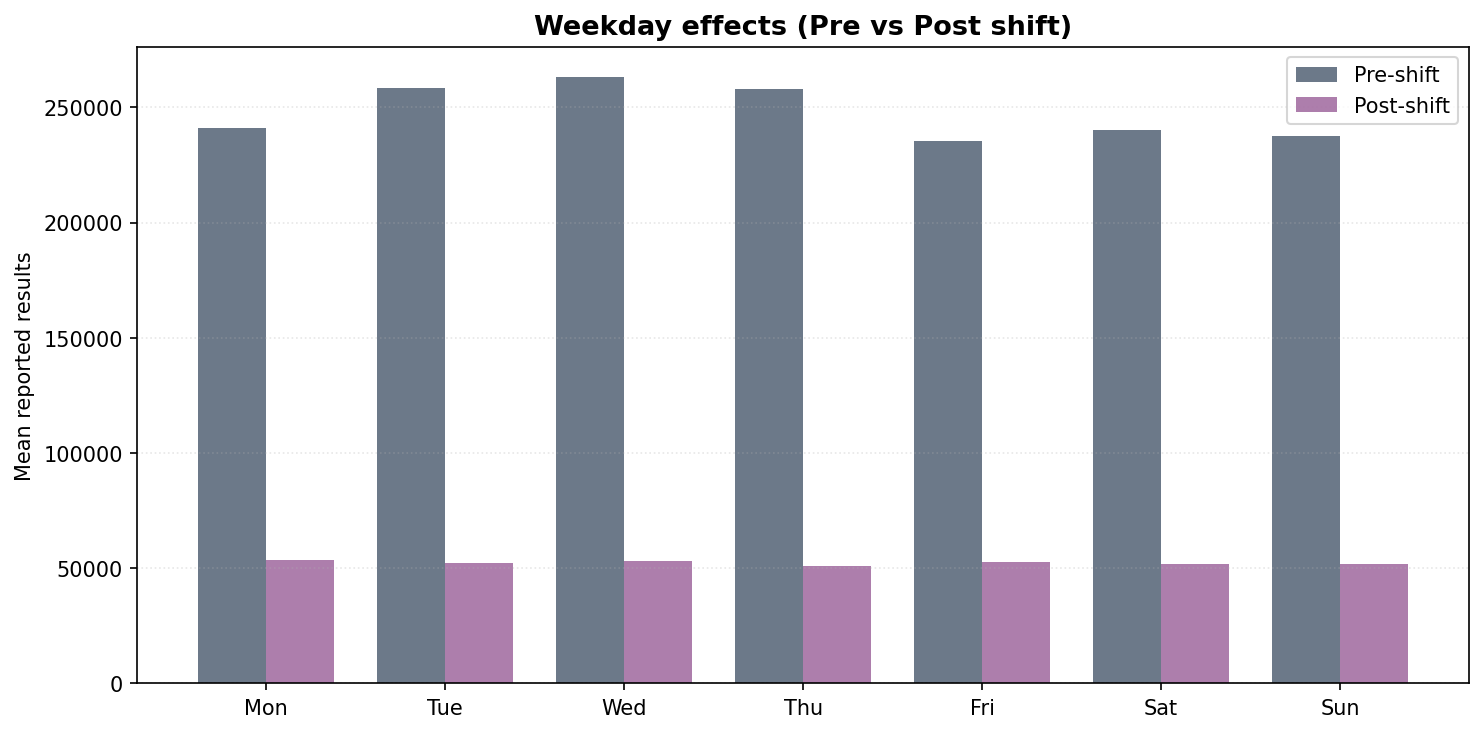
## 1.数据探索性分析

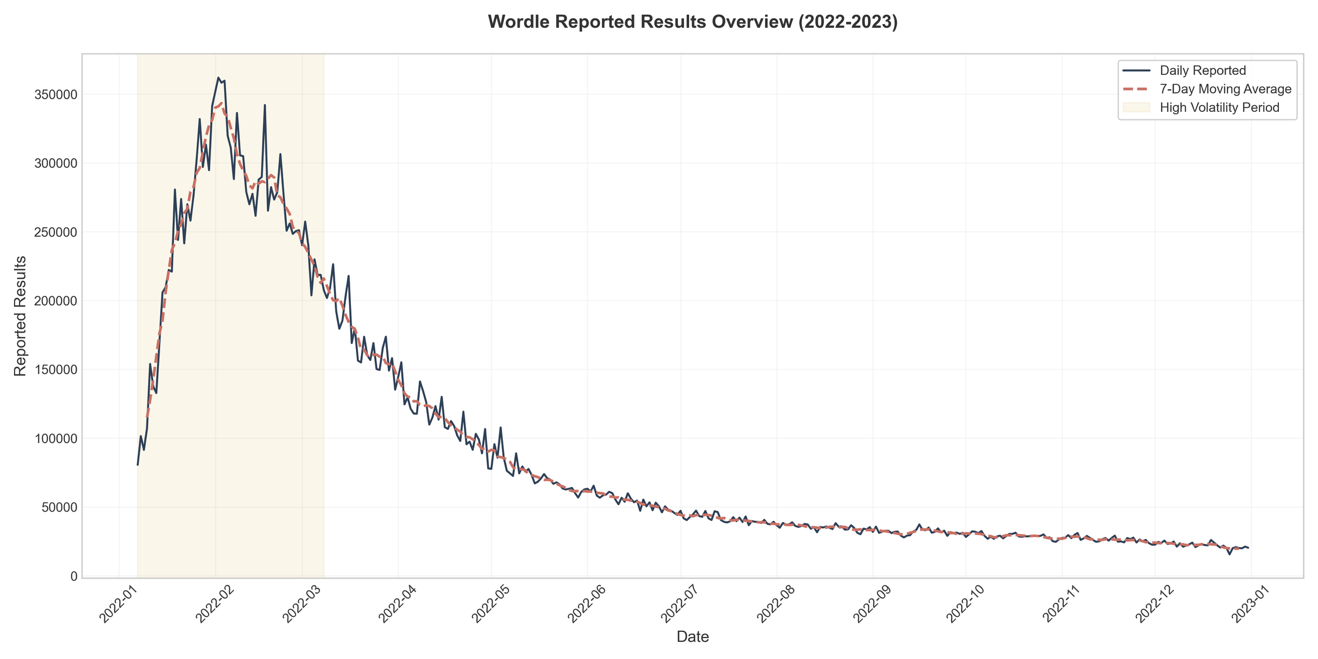
### 1.1描述统计

本研究使用的数据集涵盖了从2022年1月7日至2022年12月31日的Wordle每日报告结果，共计359个观测值。对原始数据的初步统计分析揭示了该时间序列的极端波动性与异质性。总体数据的均值为90,975人/天，但其标准差高达89,225人，变异系数接近98.1%。如此之高的变异系数表明，均值并不能代表该序列的整体水平，数据的分布呈现出极度的偏态与长尾特征。

从时间维度观察，数据明显分为两个截然不同的阶段：

1. 爆发与震荡期：2022年第一季度，数据量级在数十万级别，日间波动幅度巨大，且呈现出随数值增大而扩散的异方差性。
2. 衰减与平稳期：进入2022年第二季度后，数据量级急剧下降至几万级别，随后进入一个缓慢衰减的趋势，波动幅度相对收敛。





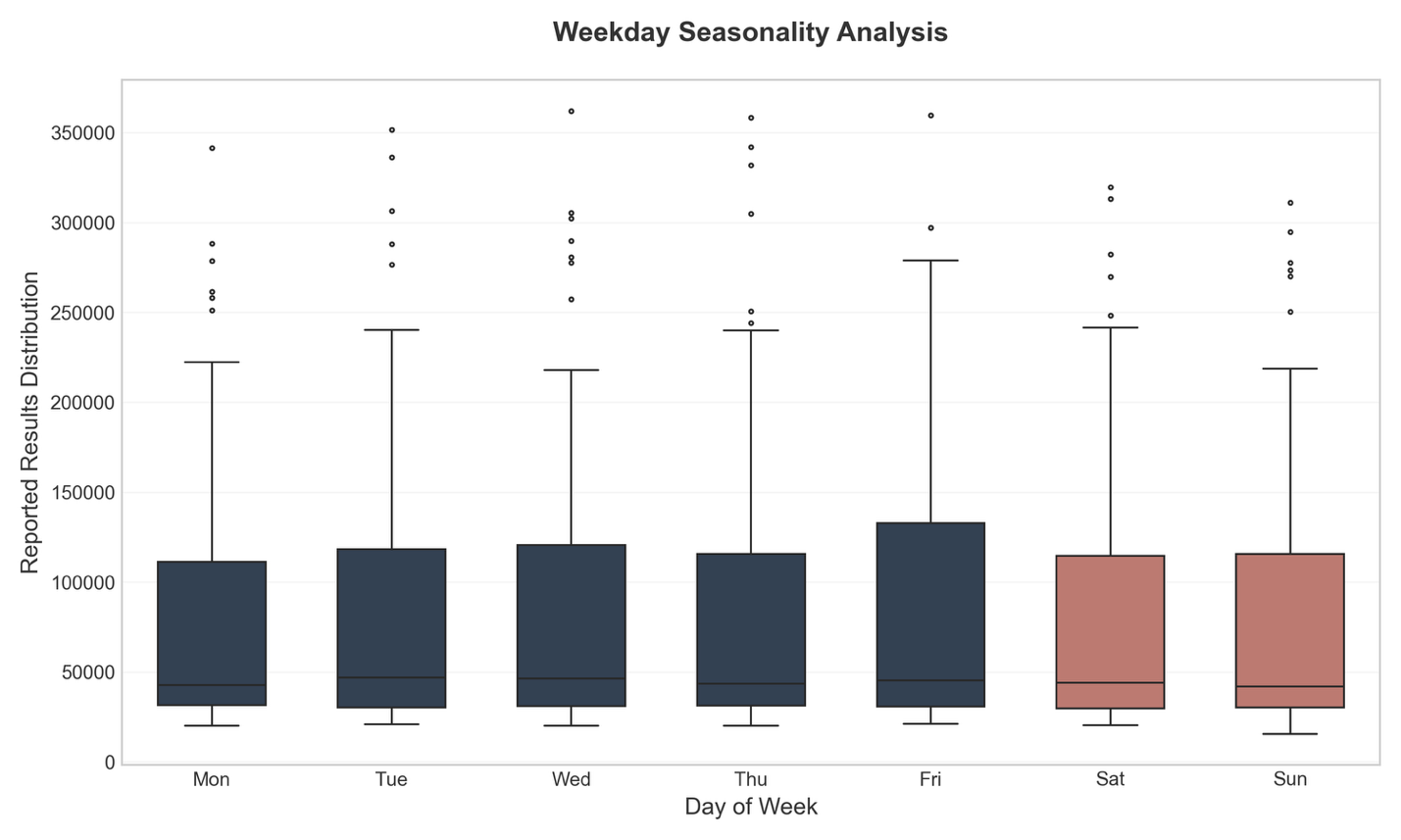
### 1.2数据预处理

原始时间序列呈现出显著的乘性误差结构，即预测误差的大小与预测值本身成正比。如果直接对具有这种特性的序列进行加性模型拟合，高值区域的误差将被赋予过大的权重，从而降低模型对低值区域的敏感度。为了消除这种统计特性的不稳定性，并为后续的线性模型奠定数学基础，我们首先对输入数据实施了变换，即：

其中为原始报告人数，为变换后的对数空间数值。变换不仅能够压缩数值的量级，将异方差转化为近似的同方差，还能将乘性趋势转化为线性趋势，使得模型在几十万量级和几万量级的区间内具有同等的敏感度。并且将原本的指数级衰减趋势转化为近似线性趋势，便于差分算子进行平稳化处理。

### 1.3核心假设

基于上述探索性分析的发现，我们提出三项针对序列特征的建模假设，并分别据此设计相应的模型策略：

1. 结构性断点假设：2022年初数据出现断崖式下跌，这表明时间序列中存在显著的结构性变点。我们假设外部事件导致了序列均值的永久性下降。因此，在模型中引入PELT变点检测机制，动态识别并隔离该变点前后的不同状态，及时切换模型的截距项和趋势参数，防止旧趋势的惯性对后续预测造成干扰。
2. 异方差假设：数据的波动幅度随整体用户规模显著变化，表现出强烈的异方差性。我们假设这反映了用户存量自然流失导致的波动率结构改变——当用户基数下降时，波动率也同步收敛。因此，我们对序列进行了对数变换，并采用逆方差加权的多模型融合来适应这一方差不稳定性，以保证无论高波动期还是低波动期，模型均能有效刻画序列走势。
3. 周期性假设：每日报告人数呈现出周周期波动的迹象，箱线图显示周末与工作日报告人数的中位数差异很小，表明周末效应相对微弱。我们假设Wordle玩家的活跃度存在每周重复的模式（工作日略高于周末），并且在节假日会出现额外的下降。为此，模型中纳入了周度季节性成分（例如SARIMA模型的7天季节项）以及周末和节假日的哑变量，以捕捉常规周期波动并纠正特定日期的异常变化。
4. 

## 2.模型设计

面对 Wordle 日报数据所呈现出的非平稳性与强异方差性，传统的全局拟合模型已难以胜任。本研究摒弃了单一模型的静态假设，构建了一套动态的“检测-隔离-集成”自适应预测框架。该框架以可解释性极强的统计学集成模型为核心，辅以基于Transformer 架构的深度学习模型作为验证，形成了一套“白盒推理 + 黑盒确证”的双重保障体系。

### 2.1 主模型：SARIMA集成模型Ensemble

主模型的设计核心创新点在于提出了一套“检测-隔离-集成”的处理机制，克服传统全局拟合方法在面对训练数据结构性断裂时的滞后性与偏差。

#### 2.1.1 时序隔离的变点检测

传统的变点检测往往是在全量历史数据上进行的事后分析，但在预测场景下，这种做法构成了严重的数据泄露。为了严谨地模拟预测环境，我们采用严格的滚动窗口机制结合PELT算法。这一算法的原理是通过最小化惩罚似然函数来搜索最优的变点位置，在抑制过度切分的同时捕捉真实的结构变化。其目标函数如下：

其中为段内成本函数（通常为负对数似然），为防止过拟合的惩罚项。

对于滚动交叉验证的每一个时间步，系统仅调取时刻之前的历史数据运行PELT算法。这确保了变点识别过程严格遵循时间因果律，杜绝了“利用未来数据预测过去”的数据泄露风险。此外，基于实时检测到的变点，系统将动态生成一个二值化的状态特征。若则表示当前仍处于Wordle游戏的大热期（高均值、高波动）。则表示检测到数据的结构性断裂，序列进入衰退期。这一特征输入使后续模型能够在断裂发生的瞬间切换截距项和趋势参数，阻断旧有趋势的惯性传导。

#### 2.1.2 分段建模与特征工程

为了捕捉细粒度的周期性波动，我们在Regime特征之外，还构建了以下协变量矩阵 ：

1. 周末指标：0代表工作日，1代表周末，用于捕捉周内效应。
2. 节假日指标：基于美国联邦法定节假日生成的0/1变量，用于捕捉特定日期的异常下跌。
3. 趋势项交互：构建时间索引与 的交互项，允许模型在不同阶段下学习不同的衰减斜率（即分段线性趋势）。

#### 2.1.3 异构模型加权集成

单一模型往往难以同时应对数据的长期趋势、短期波动和复杂的季节性。为了提高系统的泛化能力，我们并行部署了三个具有不同参数偏好的子模型，并采用逆方差加权策略进行加权融合。

1. 理论基础

ARIMA (Autoregressive Integrated Moving Average)：这是时间序列预测的基石，适用于捕捉非季节性的短期逻辑。它通过捕捉数据的惯性（自回归部分），通过处理非平稳趋势（集成部分），通过利用过去的预测误差来修正当前值（滑动平均部分）。

SARIMA (Seasonal ARIMA)：这是ARIMA的扩展，增加了季节性组件。对于Wordle数据，显然存在的周度循环。SARIMA不仅考虑“前一天”的影响，还考虑“上一周的同一天”，从而能精准捕捉周期性波动。

1. 候选模型构建

我们没有采取暴力搜索法来确定最优参数组合，而是基于对数据的理解构建了最具代表性的候选模型，分别用于捕捉序列的不同特征，涵盖了从“纯趋势”到“强周期”的多种假设。

SARIMA：意味着仅考虑最近一日的值和误差。配合  的季节性参数，可稳健地捕捉最基础的周度循环和，在数据平稳期表现最佳。

ARIMA：剔除季节性组件，当发生突发事件导致原本的“周周期”模式失效时，强行拟合季节性反而会引入噪声。因此，选取代表模型仅根据近期趋势的二阶导数（加速度）进行预测，能更敏锐地跟踪断崖式下跌或反弹。

SARIMA：引入标准周度季节性的同时赋予历史误差更高的权重，当出现剧烈震荡时该模型能够有效调整，防止预测值过度偏离真实值。

1. 动态权重计算：

权重与模型在验证集上的误差方差成反比（即某模型在最近验证窗口中的误差方差越小，则赋予其的权重越大，反之亦然）：

其中，是模型在最近滚动窗口内的验证误差方差。如果当前数据周期性规律极强，SARIMA模型的方差会很小，权重自动增大；反之，如果周期性规律失效，纯ARIMA模型可能表现更稳健，从而占据主导。这种机制赋予了预测更稳健的模型更高的权重，从而在统计学意义上降低了单一模型的参数风险，形成了模型间的优势互补，提升了整体预测的鲁棒性。

#### 2.1.4 非参数化分布重构

在对数空间完成预测后，通过指数函数 映射回原始空间得到的是预测分布的中位数而非均值。因为在对数正态分布假设下，均值往往大于中位数。为了获得统计无偏的期望值，且不依赖于残差必须服从正态分布的强假设，我们引入了Duan's Smearing Estimator：

其中 为训练集上的残差。这种非参数化的纠偏方法利用经验残差分布来调整预测期望，确保最终输出的点预测和置信区间在原始尺度上具备统计无偏性。

### 2.2 辅助模型：Chronos Time Series Transformer

## 为了检验上述基于人工特征工程和线性假设的SARIMA集成模型是否足以刻画序列规律，我们选择了Amazon推出的Chronos时序Transformer模型作为对照方法。Chronos在海量多样化的时间序列数据上预训练，基于Transformer架构能够自动捕捉复杂的非线性长程依赖，并直接输出未来时间步的概率分布，因而无需人工设定变点或季节性特征即可给出预测区间。作为“黑盒”的深度学习模型，Chronos为我们的结果提供了鲁棒性校验：如果Chronos与我们的Ensemble模型给出了趋同的预测结果，说明线性集成模型所提取的结构断点、周期性等特征是有效的；反之，若两者差异显著，则提示可能存在尚未被捕捉的非线性模式。

## 3.实证结果与对比分析

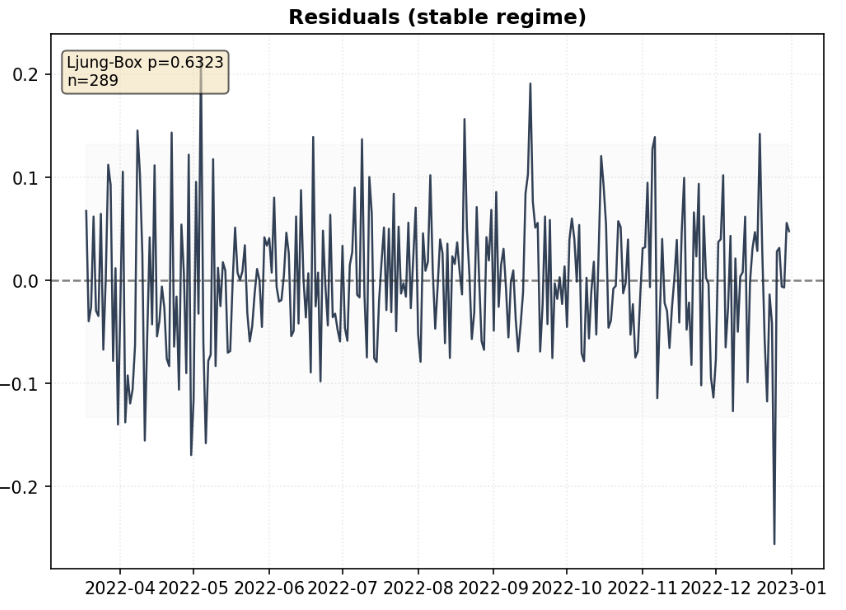
### 3.1 模型诊断与稳健性评估

预测结果显示，2023年3月1日的报告人数点估计为18,932人，90%置信区间为[14762, 24218]。在正式评估预测精度之前，我们首先对主模型的统计性质进行严格诊断，以确保模型的有效性。

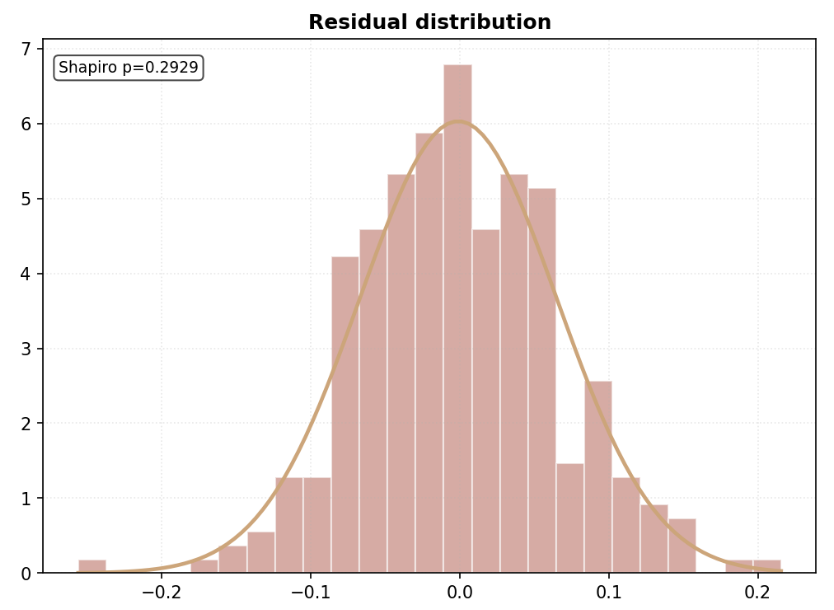
#### 3.1.1 残差白噪声检验

残差是模型未能提取的信息。一个优秀的模型残差应当等同于白噪声，分布随机且无规律。以下是本模型的检验结果：

1. 自相关检验：Ljung-Box检验的p值为 0.6323 (> 0.05)。这意味着残差序列不存显著著的自相关性，历史数据中的有效规律已被模型充分提取。



1. 正态性检验：Shapiro-Wilk 检验的 p值为 0.2929 (> 0.05)。这表明残差服从正态分布。这一结果至关重要，它验证了我们在空间进行建模的合理性，并为置信区间的计算提供了理论前提。



1. 无偏性：残差均值为 -0.0009，极接近于0，说明模型能够较为准确地预测，不存在系统性的高估或低估偏差。

#### 3.1.2 Walk-Forward回测覆盖率

为了评估模型的不确定性量化能力，我们进行了滚动步进式回测。受限于有限的序列长度，我们进行3折交叉验证。结果显示，对于未来60天的预测（预测布长），模型生成的90%置信区间实际覆盖了 95.8% 的真实观测。从统计学的角度看，这一数值高于名义覆盖率（90%），表明模型略显保守。然而，在非平稳的数据输入前提下，这也是模型鲁棒性的体现。

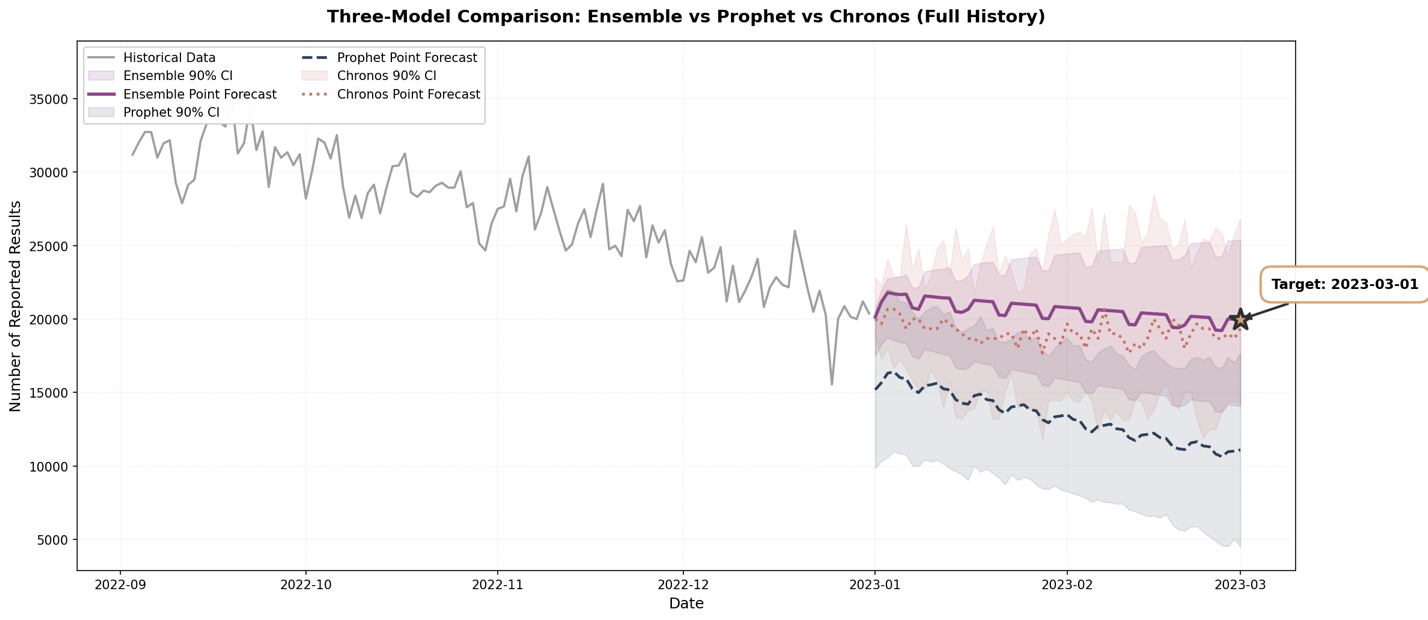
3.2 多模型对比分析

为了验证预测结果的鲁棒性，我们将自适应的Ensemble模型与工业界常用的Prophet模型以及前沿的时间序列基础模型Amazon Chronos进行了严格的横向对比。实验结果不仅揭示了不同建模范式在非平稳环境下的适应能力差异，更为我们的最终预测提供了强有力的交叉验证。

#### 3.2.1 全局拟合

|  |  |  |  |
| --- | --- | --- | --- |
| 模型 (全量数据) | RMSE (均方根误差) | 覆盖率 (Coverage) | MASE |
| Prophet | 10839.56 | 21.1% | 1.04 |
| Chronos Transformer | 2917.53 | 92.2% | 0.23 |
| Ensemble SARIMA | Log RMSE 0.08 | 93.3% | 0.52 |

*注：SARIMA RMSE为对数空间计算所得，其相对误差约为8%，映射回原数值仍优于Prophet。*

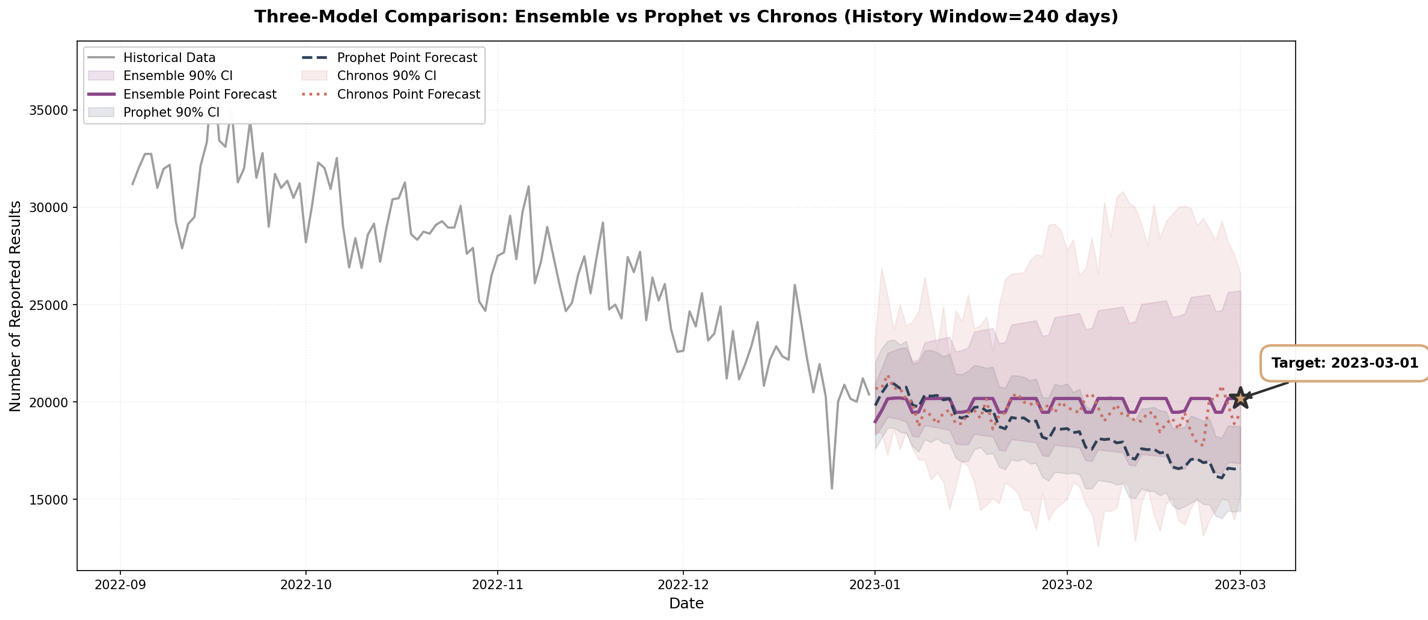


如上表所示，当使用全量历史数据训练Prophet模型时表现不佳。这是因为Prophet模型的核心假设是趋势的连续性。由于Wordle早期数据包含了从数十万到几万的断崖式下跌，在进行全局拟合时，即便使用了分段线性趋势，仍无法完全摆脱前期断崖式下跌的影响。Prophet的全局拟合算法试图用一条连续的曲线去适配这一剧烈变化，导致其误以为这一下降趋势具有巨大的惯性，从而将未来的日活人数预测得过低。21.1%的低覆盖率进一步证实，真实值绝大多数时间都落在Prophet的预测区间之外。这表明了该模型在非平稳数据输入情况下，单纯依赖全局趋势拟合的局限性。这反证了引入 PELT 变点检测机制的必要性：只有主动识别并隔离旧的趋势，模型才能正确理解当前的平稳状态。

相比之下，基于Transformer架构的Chronos模型展现出了对于非平稳数据的强大适应性。尽管作为“黑盒”模型缺乏显式的特征工程，但凭借注意力机制，Chronos成功捕捉了序列中不同阶段的权重关系。在全量历史测试中，其RMSE为2,917.53，覆盖率为92.2%，显著优于Prophet。

#### 3.2.1 部分拟合

|  |  |  |  |
| --- | --- | --- | --- |
| 模型 (近240天数据) | RMSE (均方根误差) | 覆盖率 (Coverage) | MASE |
| Prophet | 2422.73 | 85.0% | 0.85 |
| Chronos Transformer | 2219.61 | 78.3% | 0.77 |
| Ensemble SARIMA | Log RMSE 0.09 | 96.7% | 1.02 |

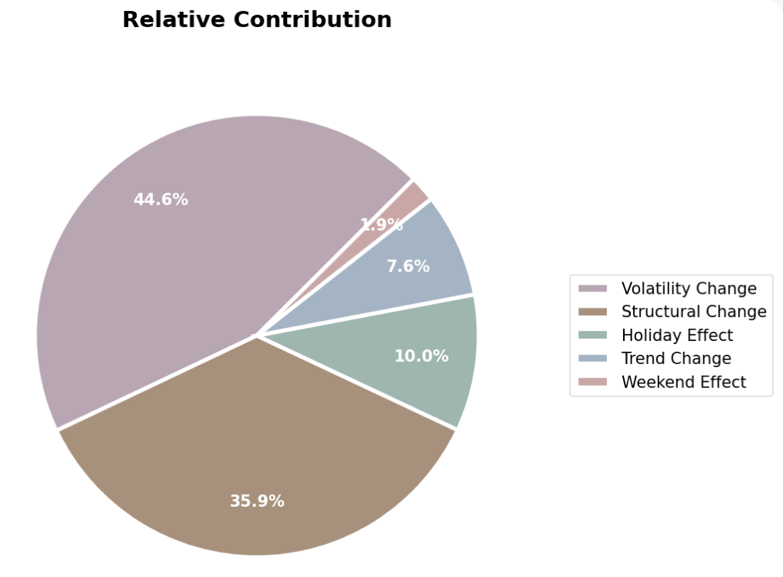


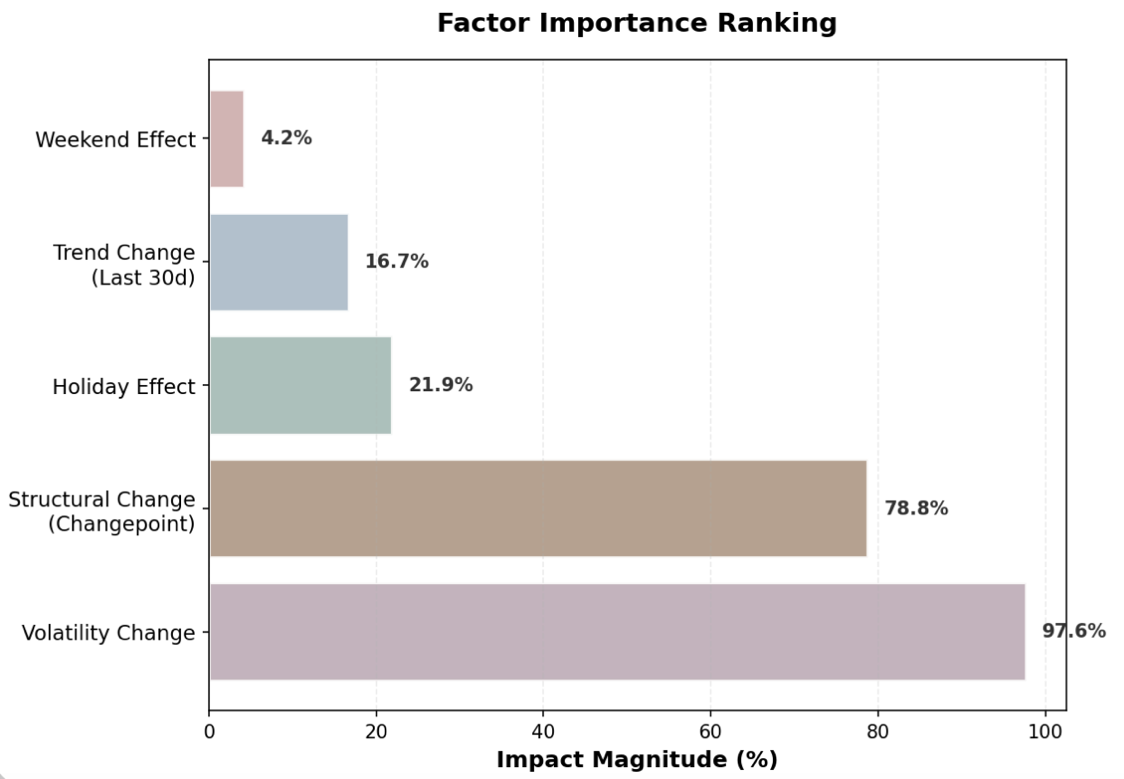
当输入数据为近240天时，观察结果可以发现，手动隔离早期的剧烈变化后Prophet模型性能得到了极大的提升，但整体预测走向依然悲观。Chronos 的预测点与Ensemble模型整体重合度较高，两者均指向20,000人左右的区间。一个是基于海量数据预训练的深度学习模型，一个是基于精细统计推断的线性集成模型，它们在原理截然不同的情况下得出了几乎一致的结论。这种一致性不仅消除了单一模型可能存在的过拟合风险，更极大地提升了我们对于最终预测结果的信息。

同时，Ensemble模型在置信区间的构建上表现得更为稳健。虽然Chronos的MASE更小，但其预测区间在远期显著发散，反映了纯数据驱动模型在长尾预测上的不确定性。而 Ensemble 模型采用了逆方差加权策略，在保持高覆盖率的同时能够给出更收敛的风险边区间界。

### 3.3 解释性分析

既然Ensemble模型已被证明是准确且可靠的，我们利用其透明的参数结构对影响Wordle报告人数的因子进行量化分解。我们从两个互补的视角，即“单次影响强度”与“整体相对贡献”，得出以下关键结论：

****



## 从饼图观察可得，最大的影响因素并非用户流失（即结构性变化），而是波动性变化，其对数据方差的解释度高达44.6%。这意味着 Wordle数据本质上是两个截然不同状态的切换：前70天是“高流量、高噪声”的快速传播期，数据受社交网络情绪驱动，大起大落；而后289天则坍缩为“低流量、低噪声”的稳定期。对于预测而言，这一发现至关重要：如果模型不能通过 变换和分段方差建模来适应这种异方差性，它必然会在后期生成过宽的置信区间，从而失去指导意义。紧随其后的是结构性变化，贡献了35.9%的方差解释度。柱状图显示，断点发生后的日均用户数出现了78.8%的永久性跌落。这两个因子合计解释了超过八成的数据变异，说明 Wordle的演变并非线性的自然衰退，而是整个环境的剧变。我们的模型正是通过捕捉这一变化，避免了像Prophet模型用前期的规律去预测后期的走势。其次，数据显示趋势变化在最近30天内导致了16.7%的量级下滑，其对整体波动的贡献度为 7.6%。这说明变点之后数据持续处于长尾衰减中，是用户新鲜感消退后的自然流失结果。虽然7.6%的贡献度看似不大，但它具有方向性和累积性。这意味着预测模型如果只看均值而不包含负向的趋势项，都会随着时间推移而高估活跃人数。最后，与许多娱乐应用不同，周末效应在其中扮演的角色微乎其微，仅占整体变异的1.9%，且单次影响幅度很小（不到5%）。这一反直觉的数据揭示了Wordle独特的用户粘性特征：它已经融入了核心用户的日常生活，无论是工作日还是周末，打卡习惯都保持了惊人的一致性。相比之下，节假日效应虽然发生频率低，但单次影响较大，是导致数据出现离群值的主要原因。