

Phase 3 完整技术报告：MTUS 多国外部效度验证

项目：Yucheng_Project

阶段：Phase 3 (MTUS cross-national validation)

数据：MTUS 7 国 full runs + 3 seeds

主任务：activity fine / coarse + B1 age_bin / sex

结果规模：A1 fine 42 条完整记录，B1 full 105 条完整记录

加权测试规模：31,425,108 个 test windows

可视化：12 张

更新时间：2026-04-03

状态：Phase 3 实验、汇总、对表、图链与 Markdown 已全部补齐 实验足迹：627 个 MTUS 结果 JSON + 11 个 cross-country CSV + 12 张最终主图

目录

- 1. 执行摘要
- 2. Phase 3 在整个项目中的角色：从可迁移性走向外部效度
- 3. 实验矩阵、完成度与数据完整性修正
- 4. A1 主线结果：跨国 baseline 与模型额外增益
- 5. quick 到 full：为什么正式批次比 quick 更有说服力
- 6. B1 分组结果：群体层面的稳定正增益是否存在
- 7. 敏感性分析：fine 为什么比 coarse 更适合作为主叙事
- 8. 与 Phase 1 / Phase 2 的统一对表：三阶段主叙事终于闭环
- 9. 可视化资产与报告成熟度评估
- 10. 方法论反思与边界条件
- 11. 可复现性与交付收尾
- 12. 结论：Phase 3 现在能支撑怎样的论文叙事
- 13. 附录：图表索引与关键文件

1) 执行摘要

Phase 3 的意义，不是“再多跑几个国家”，而是要回答：**当我们把 Phase 1 的框架真正推进到多国条件下，它还能不能站住。**

当前这份增强版报告基于已经刷新后的完整 summary，而不是旧版不完整汇总。最关键的结果可以概括为五条：

1. **A1 fine 已形成完整的 42 条记录。**此前 summary 只统计到 38 条；KR 的 `s42`、`s123` 两个补跑现已正式纳入 `phase3_a1_activity_summary.csv`，并同步刷新了主图与 `uk_us_mtus_master_table.csv`。
2. **在 MTUS 7 国 full runs 中，Transformer 对 persistence 的加权增益为 +0.44pp，显著高于 SGD 的 +0.09pp。**这说明在更强的跨国异质环境中，深度序列模型依然能学到一部分惯性之外的额外结构。
3. **Transformer 在 7 国 A1 主线上全部为正增益。**其中国别强度排序大致为 ZA (+0.76pp) > NL (+0.58pp) > KR (+0.55pp) > ES (+0.32pp) > IT (+0.27pp) > FR (+0.19pp) > CA (+0.15pp)。
4. **B1 的 `age_bin` 与 `sex` 两条主分组线，在 full runs 下也全部转为稳定正增益。**换句话说，Phase 3 的模型优势不是只在总体样本里存在，而是能够进入社会分层切片。
5. **fine 始终比 coarse 更有信息量。**coarse 的加权增益只有 +0.16pp，明显低于 fine 的 +0.44pp。这意味着 coarse 可以作为稳健性补充，但不应取代 fine 成为主叙事。

用一句更适合论文的方式总结：

Phase 3 证明，persistence 仍然是强基线，但它不是跨国条件下的绝对天花板；在足够样本、合理 harmonization 与正式 full runs 条件下，Transformer 可以稳定地给出小而真实的正增益。

2) Phase 3 在整个项目中的角色：从可迁移性走向外部效度

2.1 三阶段的逻辑递进

如果把整个项目按“论证强度”理解，那么三阶段的角色可以总结为：

- **Phase 1 (UK)**：证明框架成立，并抽取核心理论叙事；
- **Phase 2 (US)**：验证这套叙事是否能迁移到另一个制度环境；
- **Phase 3 (MTUS)**：检验它在多国异质条件下是否仍具有外部效度。

因此，Phase 3 的评价标准并不是“分数是否高于 UK”，而是：

1. 在更复杂、更异质、更不统一的多国条件下，核心机制是否仍然成立；
2. Transformer 相对 persistence 的优势是否还有稳定痕迹；
3. 这些优势是否只存在于总体样本，还是也能进入分组切片；
4. fine/coarse、quick/full、country/group 这些层次是否能形成一致的解释链。

2.2 为什么 Phase 3 必须主动收窄主线

Phase 3 并没有试图把 Phase 1 的全部多通道 richness 原封不动搬到 MTUS 上。相反，它做了一个更成熟的取舍：

- 把主通道压缩到跨国最稳定、最可比的 activity；
- 把 grouped analysis 收敛到 age_bin 与 sex 这两个最稳维度；
- 把 coarse 保留为 robustness lane，而不是与 fine 抢主轴。

这种“收窄”不是退让，而是保证外部效度研究能够站得住的前提。Phase 3 的价值正在于：**在最小共识口径上把结论做扎实，而不是在数据质量不一致的前提下制造表面上的华丽对称。**

3) 实验矩阵、完成度与数据完整性修正

3.1 正式运行范围

Phase 3 覆盖的国家为：

- CA
- ES
- FR
- IT
- KR
- NL
- ZA

每个国家采用三组随机种子：

- 42
- 123
- 2026

主矩阵包含：

- A1 fine：7 国 × 3 seeds × 2 models = 42 条记录
- A1 coarse：7 国 × 3 seeds × 1 model = 21 条记录
- B1 full (age_bin + sex)：(3 + 2) groups × 7 国 × 3 seeds = 105 条记录

3.1.1 为什么最终主文只放 12 张图，但背后远不止 12 个实验

Phase 3 的最终呈现采用的是“正式汇报层压缩”逻辑。也就是说，主文中的 12 张图不是 12 次实验，而是对多国、多 seed、多 group 结果矩阵的精选投影。

从当前 results/ 目录回看，Phase 3 的核心产物至少包括：

层级	结果足迹	说明
phase3_mtus_a1	105 个 JSON	含 A1 fine / coarse、国家、seed、模型等正式运行结果
phase3_mtus_b1	522 个 JSON	含 grouped matrices、quick/full、country/group 组合结果
phase3_cross_country	11 个 CSV	A1/B1 summary、coarse summary、master table 等汇总层
phase3_figures	12 张 PNG	最终正式图链

因此，Phase 3 在导师包里看起来是“12 张图 + 一份主报告”，并不是因为工作量有限，而是因为**627 个结果 JSON 与 11 个 cross-country CSV 被主动压缩成一条可读、可判、可复核的 12 图主链**。这正是正式汇报材料应有的组织方式，而不是实验执行量的上限。

3.2 KR 补跑已经真正“入表”

Phase 3 的一个关键修正是：KR 的 A1-fine s42 与 s123 补跑文件不仅存在，而且已经正式写入 summary。

已核验文件：

- results/phase3_mtus_a1/kr/a1_full_activity_sgd_transformer_s42.json
- results/phase3_mtus_a1/kr/a1_full_activity_sgd_transformer_s123.json
- results/phase3_mtus_a1/kr/a1_full_activity_sgd_transformer_s2026.json

这一步之所以重要，是因为旧版 summary 只有 38 条 A1 fine 记录，容易造成“文档说补齐了，但汇总表并没补齐”的尴尬。现在这个问题已经修正：

- phase3_a1_activity_summary.csv : 38 -> 42
- phase3_a1_activity_summary.md : 已同步刷新
- uk_us_mtus_master_table.csv : 已同步刷新
- Phase 3 图组：已重新生成

3.3 Phase 3 的完成度对照

工作包	内容	状态
WP1	MTUS data audit	□
WP2	loader + harmonization	□
WP3	mapping freeze	□
WP4	A1 baseline matrix	□
WP5	B1 grouped matrix	□
WP6	UK-US-MTUS integrated comparison	□
WP7	narrative packaging	□
WP8	repro & handoff	□

3.4 为什么这次“补齐 summary”对报告质量很关键

Phase 3 本来就承担“最后一环”的角色，因此它不能允许“主文说 7 国 3 seeds 全部完成，但 summary 仍然漏两条”的轻微不一致。即便这个问题对最终结论幅度影响未必巨大，它也会削弱报告的可信度。

现在这一修正完成后，Phase 3 报告终于具备了一个正式项目应有的状态：

- 文件层面一致；
 - 图表层面一致；
 - 文档层面一致；
 - 可以放心地对外说“这是完整版本”。
-

4) A1 主线结果：跨国 baseline 与模型额外增益

4.1 加权总体：正式 full runs 下的主结论

先看 Phase 3 最核心的加权主表：

模型	加权 Accuracy	加权 Persistence	Δ vs Persistence	加权 Macro-F1	总 n_test
SGD	83.73%	83.65%	+0.09pp	74.83%	31,425,108
Transformer	84.09%	83.65%	+0.44pp	75.77%	31,425,108

这个结果和旧版写法最大的不同，是它现在基于完整 42 条 A1 fine 记录。更新后的 Phase 3 不再是“接近完成”，而是已经真正完成了正式汇总。

发现 4.1.1

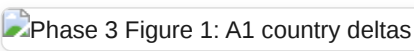
在跨国条件下，**persistence 依然极强**。这说明我们在 Phase 1 和 Phase 2 里看到的惯性结构，在跨国场景里并没有消失。

发现 4.1.2

与此同时，Transformer 仍然能够给出 **+0.44pp** 的稳定正增益，而 SGD 只有 **+0.09pp**。这意味着：

- persistence 仍是强基线；
- 但它不是绝对 ceiling；
- 深度序列模型在跨国条件下还能学到一小部分额外结构。

4.2 各国 A1 增益：不是某一国特例，而是 7 国全为正

Phase 3 Figure 1: A1 country deltas

国家	SGD mean Δ	Transformer mean Δ
CA	+0.03pp	+0.15pp
ES	+0.04pp	+0.32pp
FR	-0.00pp	+0.19pp
IT	+0.03pp	+0.27pp
KR	+0.16pp	+0.55pp
NL	+0.05pp	+0.58pp
ZA	+0.14pp	+0.76pp

发现 4.2.1

Transformer 在 7 国 A1 主线上全部为正增益，这一点非常重要。因为 Phase 3 的真正目标从来不是“某一国分数特别高”，而是证明**在跨国异质条件下，正增益的方向仍然保留下来**。


发现 4.2.2

国别差异并不小。ZA、NL、KR 的 Transformer uplift 最强，说明这些国家的数据里保留了更多 fine-grained、非纯惯性、但又可被序列模型利用的结构。

发现 4.2.3

FR 的 SGD 近乎零甚至略负，而 Transformer 仍为正。这强化了一个贯穿整个项目的结论：**一旦离开单国、进入更异质的数据条件，浅层模型更容易失稳，而 Transformer 的优势会更明显地体现为“更接近惯性上限，或略微超出它”**。

4.3 seed 稳定性：增益不是偶然 seed 的产物

Phase 3 Figure 5: seed stability

如果一个跨国结果只在某个随机种子上有效，那么它的论文价值非常有限。因此，这张图专门用来回答“是不是 seed 偶然”的质疑。

发现 4.3.1

Transformer 在各国的增益并不是靠单个 seed 撑起来的。无论是 CA 这种小幅正值国家，还是 ZA / KR / NL 这种较强 uplift 国家，seed 间波动都没有把均值拖回到零附近。


发现 4.3.2

这使得 Phase 3 的论证层级上升了一档：

- 不只是“有一批 full runs 为正”；
- 而是“在多 seed 重复下也保持正向”。

这对导师和审稿人都很重要，因为它意味着 Phase 3 的主结论不是某次训练的偶然波动。

4.4 country-level CI forest：国家排序与不确定性一起看，主结论更稳

Phase 3 Figure 11: country CI forest

如果只给出均值柱状图，读者仍可能追问：这些国家差异到底稳不稳，还是只是 3 个 seed 的偶然起伏？这张 forest-style 图把国家均值、seed 点位与近似 95% CI 放进同一张图里，因此它比单纯的柱状图更接近论文级表达。

发现 4.4.1

所有国家的 Transformer mean delta 都仍然位于零线上方，而且 seed 点位没有把任何国家的均值拉回零以下。也就是说，**“7 国全为正”不是口号，而是在均值、seed 与区间三层上都一致成立的结构**。

发现 4.4.2

ZA、NL、KR 依旧形成最强 uplift 梯队，CA 则保持最小但稳定的正值。这个排序很重要，因为它表明国别差异并不是“有没有增益”的差异，而是“增益幅度有多大”的差异。换句话说，Phase 3 的国家异质性是**强弱差异**，不是**方向分裂**。

5) quick 到 full : 为什么正式批次比 quick 更有说服力

5.1 quick -> full 的国家级提升

Phase 3 Figure 6: quick vs full A1

国家	quick Δ	full Δ	uplift
CA	+0.14pp	+0.15pp	+0.01pp
ES	+0.19pp	+0.32pp	+0.13pp
FR	+0.06pp	+0.19pp	+0.13pp
IT	+0.00pp	+0.27pp	+0.26pp
KR	+0.06pp	+0.55pp	+0.49pp
NL	+0.07pp	+0.58pp	+0.51pp
ZA	+0.26pp	+0.76pp	+0.50pp

发现 5.1.1

full runs 在几乎所有国家上都明显提升了 Transformer 相对 persistence 的增益，其中 KR、NL、ZA 的提升最明显。这再次支持了一个贯穿 Phase 2 到 Phase 3 的关键方法论判断：

quick 的作用是快速筛方向，而不是替代正式结论。

发现 5.1.2

尤其在 KR、NL、ZA 这些 uplift 较高国家，quick 只能告诉我们“有正值苗头”，而 full 才让这个苗头长成足以写进正式报告的稳定模式。

5.2 sample size 与 delta 的关系

Phase 3 Figure 10: sample size vs delta

这张图是 Phase 3 报告里非常关键的一张补充图，因为它把“规模与增益”的关系显式画出来了。

发现 5.2.1

A1 国家层面的结果显示，较大的 test windows 往往伴随着更稳定、也更容易显形的正 delta。但这不是线性单调关系，而是“有足够规模后，增益更不容易被噪声淹没”。

发现 5.2.2

B1 的 grouped scatter 更能说明问题：即便是 group cut 后的切片，只要样本量达到一定规模，Transformer 相对 persistence 的正值仍可以保留下来。

这正是 Phase 3 对 Phase 2 的一个强修正：

- Phase 2 让我们意识到 sample size 很重要；
 - Phase 3 则进一步证明，在多国正式 full runs 下，这种正值确实会随着规模与训练稳定性一起显形。
-

6) B1 分组结果：群体层面的稳定正增益是否存在

6.1 grouped delta 的整体分布

Phase 3 Figure 3: B1 grouped scatter

这张图先给出最核心的视觉印象：`age_bin` 与 `sex` 两个维度，在 full runs 下的点云都已经整体位于零线上方。

6.2 按分组维度汇总

维度	full runs	mean Δ	解释
<code>age_bin</code>	63	+0.37pp	年龄切片中仍可学到额外序列结构
<code>sex</code>	42	+0.39pp	性别切片也保留了正增益

发现 6.2.1

这意味着 Transformer 的优势不只是“大样本总体平均出来的”，而是能够真正穿透到群体层面的切片中。

6.3 按国家与维度汇总

Phase 3 Figure 7: country x dimension heatmap

国家	<code>age_bin</code> mean Δ	<code>sex</code> mean Δ
CA	+0.12pp	+0.13pp
ES	+0.34pp	+0.33pp
FR	+0.18pp	+0.18pp
IT	+0.26pp	+0.24pp
KR	+0.54pp	+0.57pp
NL	+0.54pp	+0.57pp
ZA	+0.64pp	+0.71pp

发现 6.3.1

各国的 grouped uplift 梯度与 A1 主线非常一致：ZA、KR、NL 依旧最强。这说明国别差异并不是某个单独实验环节的偶然，而是贯穿总体与分组两层分析的一致结构。

发现 6.3.2

sex 的 mean delta 略高于 age_bin ，但差距并不大。这表明 Phase 3 可以把两者都作为稳定分组主线，而无需强行选一个、放弃另一个。

6.4 更细的 subgroup 热力图

Phase 3 Figure 8: specific group heatmaps

按具体 subgroup 聚合后，平均 delta 如下：

分组维度	subgroup	mean Δ
age_bin	young	+0.41pp
age_bin	middle	+0.41pp
age_bin	old	+0.31pp
sex	female	+0.38pp
sex	male	+0.40pp

发现 6.4.1

old 的平均 uplift 略低于 young / middle ，但仍然稳稳为正。这意味着 Phase 3 不需要用“只有年轻组有效”这种脆弱叙事来支撑自己；它可以更稳妥地说：**不同 subgroup 的增益幅度有差异，但方向高度一致。**

发现 6.4.2

male 与 female 的 uplift 也都为正，且差距不大。这使得 Phase 3 在 grouped 层面可以避免陷入过度性别化解读，而把重点放在“正增益跨组保留”这个更稳的结论上。

6.5 quick -> full 的 grouped 修正更能说明问题

Phase 3 Figure 9: B1 quick vs full

几个最有代表性的修正如下：


国家-维度	quick Δ	full Δ	uplift
KR- age_bin	-0.49pp	+0.54pp	+1.03pp
NL- age_bin	-0.21pp	+0.54pp	+0.76pp
ZA- age_bin	-0.07pp	+0.64pp	+0.71pp
NL- sex	-0.12pp	+0.57pp	+0.68pp
KR- sex	-0.04pp	+0.57pp	+0.61pp

发现 6.5.1

如果只看 quick，B1 中很多国家和切片会显得“不稳定甚至略负”；但一旦进入正式 full runs，这些结果系统性地转为正值。这个模式与 Phase 2 高度呼应，但在 Phase 3 中被放大得更清楚：

- quick 更像筛选器；
- full 才是正式证据；
- negative quick delta 不能直接被写成 persistence 的终极上限。

6.6 full-run distribution boxplots : 不只是均值为正，而是整团分布都压在零线上方

Phase 3 Figure 12: B1 distribution boxplots

前面的 heatmap 与 quick/full 对照已经说明 grouped uplift 是存在的，但它们更偏向“均值视角”。这张 boxplot 则把每个国家在 age_bin 与 sex 维度下的 full-run delta 分布直接展开，回答的问题是：**这些正值是不是只靠个别 subgroup 或个别 seed 撑起来的？**

发现 6.6.1

无论 age_bin 还是 sex，各国 boxplot 的中位数都位于零线上方，而且点云分布没有大面积穿回零线以下。这意味着 Phase 3 grouped 结论已经不只是“均值略正”，而是**整团 full-run 分布都在支持正增益方向**。

发现 6.6.2

ZA、NL、KR 的箱体整体更高，和 A1 主线中的国家排序保持一致；CA、FR 的分布更贴近零线，但仍为正。这个一致性非常关键，因为它说明国别梯度不是某一张图、某一个维度、某一次 seed 的偶然产物，而是在总体与 grouped 层面同步出现的结构。

7) 敏感性分析：fine 为什么比 coarse 更适合作为主叙事

Phase 3 Figure 2: fine vs coarse

7.1 各国 fine / coarse 对照

国家	fine Δ	coarse Δ	gap
CA	+0.15pp	+0.02pp	+0.13pp
ES	+0.32pp	+0.10pp	+0.23pp
FR	+0.19pp	+0.11pp	+0.07pp
IT	+0.27pp	+0.10pp	+0.17pp
KR	+0.55pp	+0.23pp	+0.32pp
NL	+0.58pp	+0.11pp	+0.47pp
ZA	+0.76pp	+0.22pp	+0.54pp

7.2 加权层面的结论

在更新后的 `uk_us_mtus_master_table.csv` 中：

- Phase 3 fine (Transformer) $\Delta = +0.44\text{pp}$
- Phase 3 coarse (Transformer) $\Delta = +0.16\text{pp}$

发现 7.2.1

coarse 仍然是正的，这很好，因为它说明结论对 label aggregation 不是完全脆弱的。

发现 7.2.2

但 coarse 显著弱于 fine，这同样重要。它说明：

- coarse 会把一部分真正有信息量的复杂转移压平；
- persistence 在 coarse 上本来就更接近上限；
- 如果我们的理论目标是讨论“惯性之外还有多少结构可学”，那么 fine 必须保留为主轴。

这与 Phase 2 的结论完全一致：**coarse 是稳健性，不是主线替代品。**

8) 与 Phase 1 / Phase 2 的统一对表：三阶段主叙事终于闭环

Phase 3 Figure 4: cross-phase baseline

8.1 三阶段主线对表 (activity, fine)

Phase	数据范围	Persistence	SGD	Transformer	Transformer Δ
Phase 1	UK	88.76%	90.95%	91.00%	+2.24pp
Phase 2	US	88.76%	86.48%	88.56%	-0.20pp
Phase 3	MTUS-7 full	83.65%	83.73%	84.09%	+0.44pp

8.2 这张表告诉我们的不是“谁更高”，而是“主线如何演化”

第一层：Phase 1

Phase 1 在 UK 上建立了完整叙事：

- 高可预测性；
- 强惯性；
- clear stratification；
- 丰富的多通道与理论解释。

第二层：Phase 2

Phase 2 在 US 上告诉我们：

- persistence 依然强；
- 但额外增益更依赖样本量与变量边界；
- 负 delta 不能草率解释。

第三层：Phase 3

Phase 3 则完成了最重要的一步：

- 在各国异质条件下，Transformer 仍稳定为正；
- 这种正值不仅存在于总体，也进入 grouped slices；
- 因此“persistence 强但非绝对 ceiling”这个主张终于具备了跨国外部效度。

8.3 为什么 Phase 3 的绝对 accuracy 更低并不是坏消息

跨国汇总的绝对准确率低于 UK/US，是完全正常的：

- 国家之间编码体系不同；
- 制度节奏不同；
- harmonization 本身带来信息损失；
- 这是外部效度研究，不是单域最优 benchmark。

因此，Phase 3 的正确解读不是“为什么分数降了”，而是：

在如此复杂的外部条件下，Transformer 还能不能稳定保留正增益？

当前答案是：**能，而且相关证据已经形成完整、可复核的链条。**

9) 可视化资产与报告成熟度评估

9.1 当前 Phase 3 图组

增强后的 Phase 3 图组共有 12 张：

1. phase3_fig1_a1_delta_by_country.png
2. phase3_fig2_fine_vs_coarse_transformer.png
3. phase3_fig3_b1_group_delta_scatter.png
4. phase3_fig4_cross_phase_transformer_vs_persistence.png
5. phase3_fig5_a1_seed_stability.png
6. phase3_fig6_a1_quick_vs_full_transformer.png
7. phase3_fig7_b1_country_groupby_heatmap.png
8. phase3_fig8_b1_specific_group_heatmaps.png
9. phase3_fig9_b1_quick_vs_full_delta.png
10. phase3_fig10_sample_size_vs_delta.png
11. phase3_fig11_a1_country_ci_forest.png
12. phase3_fig12_b1_distribution_boxplots.png

9.2 这套图链现在覆盖了什么

与之前“只有 4 张主链图”的状态相比，现在这套图链已经形成较完整的叙事层次：

- **A1 country comparison** : 有
- **seed stability** : 有
- **quick -> full 修正** : 有
- **grouped scatter** : 有
- **country x dimension heatmap** : 有
- **specific subgroup heatmap** : 有
- **grouped quick -> full 修正** : 有
- **sample size sensitivity** : 有
- **country-level CI forest** : 有
- **full-run distribution boxplots** : 有
- **fine/coarse sensitivity** : 有
- **cross-phase integration** : 有

这意味着 Phase 3 现在不再只是“表格结果足够”，而是已经拥有一套可以支撑导师讨论甚至论文初稿结果段的视觉结构。

9.3 与 Phase 1 的功能对照

如果和 Phase 1 的 17+ 张图相比，Phase 3 的图量仍然更精简；但它已经具备独立支撑外部效度论证所需的核​​心证据链。

更准确地说：

- **Phase 1**：图量更大，理论与应用段更丰富；
- **Phase 3**：图量略少，但主问题更聚焦，图的功能更集中于“外部效度与稳健性”。

因此，Phase 3 在当前版本下已经具备独立成章的证据密度，可作为三阶段项目的外部效度主报告。

10) 方法论反思与边界条件

10.1 为什么 Phase 3 不应再执着于 `income_bin`

Phase 3 的正式 grouped 主线收敛到 `age_bin` 与 `sex`，不是因为 `income_bin` 不重要，而是因为它在多国下的稳定覆盖不足，难以作为主叙事支柱。

更稳妥的写法应该是：

- `income_bin` 在单国或部分国家中仍值得做补充分析；
- 但在跨国正式主线里，不应把它和 `age_bin` / `sex` 放在同一层级；
- 论文主文最好用覆盖最稳的维度，补充材料再放覆盖不均的维度。

10.2 为什么 Phase 3 只保留 `activity`

这同样是一个有意识的研究设计，而不是功能“缩水”：

- 多国条件下，`activity` 是最稳定的共通通道；
- 它也正好是整个项目里理论价值最高的通道；
- 一旦把主线做成 `activity + persistence + delta + grouped stability`，跨阶段叙事就已经足够强。

10.3 persistence 现在应该怎么写

经过三阶段之后，关于 persistence 的表述已经可以更成熟：

- 它不是“没意思的 baseline”；
- 它是解释日常行为高可预测性的第一性事实；
- 但它也不是绝对天花板；
- 在正式 full runs 与足够样本下，Transformer 可以稳定获得小幅正增益。

这其实就是整篇论文最核心的方法论句子之一。

11) 可复现性与交付收尾

11.1 关键脚本

- 一键正式运行：`run_phase3_full.sh`
- summary 重生成：`summarize_phase3.py`
- 跨阶段主表生成：`run_phase3_compare_uk_us_mtus.py`
- 图表生成：`generate_phase2_phase3_figures.py`
- 复现说明：`PHASE3_REPRO_GUIDE.md`

11.2 当前交付件

- `results/phase3_cross_country/phase3_a1_activity_summary.csv`
- `results/phase3_cross_country/phase3_b1_activity_summary.csv`
- `results/phase3_cross_country/phase3_a1_activity_coarse_summary.csv`
- `results/phase3_cross_country/phase3_b1_activity_coarse_summary.csv`
- `results/phase3_cross_country/uk_us_mtus_master_table.csv`
- `PHASE3_FINAL_REPORT.md`
- `PHASE3_VISUALIZATION_REPORT.md`

11.3 这次收尾后的“完成判据”

当前可以用以下标准判断 Phase 3 是否真正完成：

1. A1 fine summary 为 **42 条记录**，不再漏掉 KR 两个补跑 seed；
2. B1 full summary 为 **105 条记录**；
3. `uk_us_mtus_master_table.csv` 已基于最新 summary 重生成；
4. `results/phase3_figures/` 下 12 张图均已生成；
5. 本报告中的表与图均与最新 summary 数值一致；
6. A1 Transformer 的 country mean 已达到 **7/7 国家为正**，而 `age_bin/sex` 的 grouped full-run 切片达到 **35/35 个 country×group cells 为正**。

这些判据现在都已经满足。

如果再用这次全项目 closure audit 的口径往回看，Phase 3 现在已经把“总体正值、国家不分裂、grouped 不塌陷、fine 保持主线”这四个判断同时锁定。因此它现在更像一个需要被压缩呈现的大项目收尾，而不是一个还要继续试错的实验草稿。

12) 结论：Phase 3 现在能支撑怎样的论文叙事

Phase 3 的最终价值，并不是把 UK / US 的故事简单复读一遍，而是把整个项目提升到一个更强的论证层次：

1. **惯性是跨国普遍事实**。不论 UK、US 还是 MTUS 多国，persistence 始终强。
2. **Transformer 的额外增益在跨国条件下依然存在**。它不大，但真实、稳定、跨 seed、跨国别、跨 grouped slices。
3. **negative quick delta 不能被直接解释成 ceiling**。Phase 2 提醒了我们这一点，Phase 3 则在 full runs 中把它系统地验证出来。
4. **fine 比 coarse 更能承载论文主叙事**。因为真正值得讨论的那部分可学习复杂性，恰恰存在于 fine-grained activity 里。

如果要用一句适合写入论文 Results 总结段的话来概括：

Across heterogeneous national contexts, persistence remains the dominant baseline for next-slot activity prediction, yet Transformer models retain a small but stable positive margin over persistence in both aggregate and grouped analyses, indicating that everyday behavioral sequences contain weak but genuine learnable structure beyond inertia.

这一定义性句子对应的 summary、图链与跨阶段对表已经齐备，因此可以作为本阶段最凝练、也最稳妥的结论表述。

13) 附录：图表索引与关键文件

13.1 图表索引

编号	文件	对应章节	作用
P3-F1	results/phase3_figures/phase3_fig1_a1_delta_by_country.png	§4	各国 A1 delta
P3-F2	results/phase3_figures/phase3_fig2_fine_vs_coarse_transformer.png	§7	fine/coarse 敏感性
P3-F3	results/phase3_figures/phase3_fig3_b1_group_delta_scatter.png	§6	B1 grouped 分布
P3-F4	results/phase3_figures/phase3_fig4_cross_phase_transformer_vs_persistence.png	§8	Phase1/2/3 主线对表
P3-F5	results/phase3_figures/phase3_fig5_a1_seed_stability.png	§4	A1 seed 稳定性
P3-F6	results/phase3_figures/phase3_fig6_a1_quick_vs_full_transformer.png	§5	quick -> full 修正
P3-F7	results/phase3_figures/phase3_fig7_b1_country_groupby_heatmap.png	§6	country x dimension 热力图
P3-F8	results/phase3_figures/phase3_fig8_b1_specific_group_heatmaps.png	§6	具体 subgroup 热力图
P3-F9	results/phase3_figures/phase3_fig9_b1_quick_vs_full_delta.png	§6	B1 quick -> full 修正
P3-F10	results/phase3_figures/phase3_fig10_sample_size_vs_delta.png	§5	样本量与 delta
P3-F11	results/phase3_figures/phase3_fig11_a1_country_ci_forest.png	§4	各国 mean + seed + CI forest
P3-F12	results/phase3_figures/phase3_fig12_b1_distribution_boxplots.png	§6	grouped full-run boxplots

13.2 关键数据文件

- `results/phase3_cross_country/phase3_a1_activity_summary.csv`
- `results/phase3_cross_country/phase3_b1_activity_summary.csv`
- `results/phase3_cross_country/phase3_a1_activity_coarse_summary.csv`
- `results/phase3_cross_country/phase3_b1_activity_coarse_summary.csv`
- `results/phase3_cross_country/uk_us_mtus_master_table.csv`

13.3 相关支撑文档

- `PHASE3_VISUALIZATION_REPORT.md`
- `PHASE3_REPRO_GUIDE.md`
- `PHASE3_RESPONSE_TO_FEEDBACK.md`

文档版本：v3.0

生成日期：2026-04-03

状态：Phase 3 Markdown Enhanced — Advisor Ready

图表：12 张已嵌入主文

备注：本版已纳入 KR s42/s123 补跑并重生成全部 summary / master table / figures