



Polygenic Risk Score for Alzheimer's Disease in Hispanic Populations Using a Novel Machine Learning Approach

A multi-cohort analysis using **TransPRS** — an attention-based transformer that learns multi-trait polygenic scores end-to-end, with no external GWAS weights.

Sarwan Ali, Giuseppe Tosto

Columbia University Irving Medical Center • ADSP Release 5 • Four Hispanic/Latino cohorts (N = 5,292)



HEADLINE RESULTS

0.704

mean within-cohort AUC

0.681

mean cross-cohort AUC

2.2%

avg. performance decline

100%

cross-cohort configs won by LR

1 Background & Motivation

Alzheimer's disease **disproportionately affects Hispanic/Latino populations**, yet these groups remain severely underrepresented in genomic research.

Standard PRS depend on **European-ancestry GWAS** and are poorly calibrated in admixed populations. **TransPRS** learns multi-trait scores directly from individual genotypes — closing this equity gap.

1.5×

higher AD prevalence

~70%

AD heritability

0

external GWAS weights

2 Cohorts & Study Design

Four independent Hispanic/Latino cohorts from **ADSP Release 5**, spanning distinct geographies.

COHORT	N	SNPS	M / F
ADC	1,353	5.52M	468/885
EFIGA	1,576	4.85M	507/1069
TARCC	607	5.28M	210/397
WHICAP	1,756	4.85M	505/1251
Total	5,292	20.5M	1690/3602

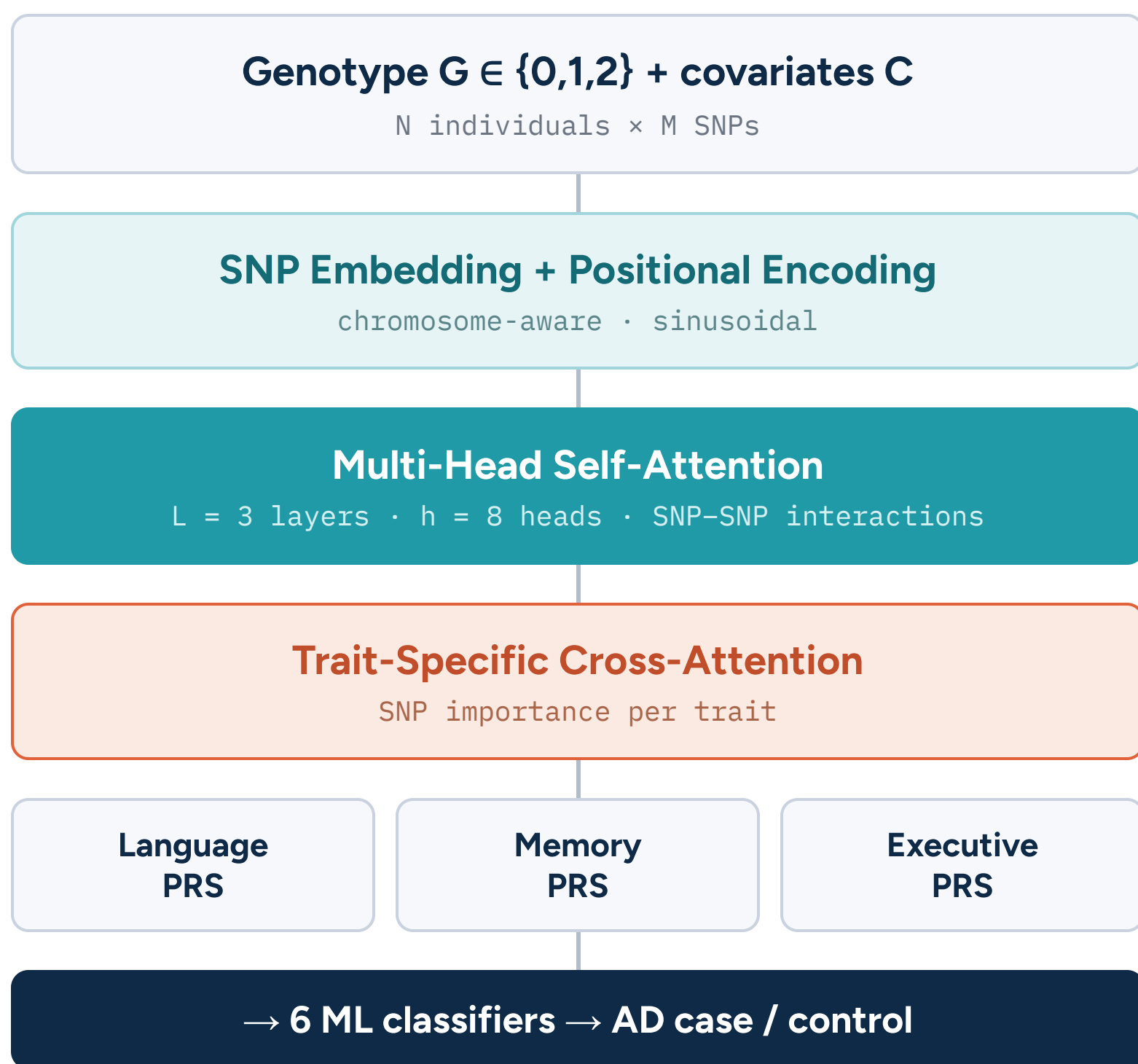
Within-cohort • 5-fold CV

Cross-cohort

16 configurations

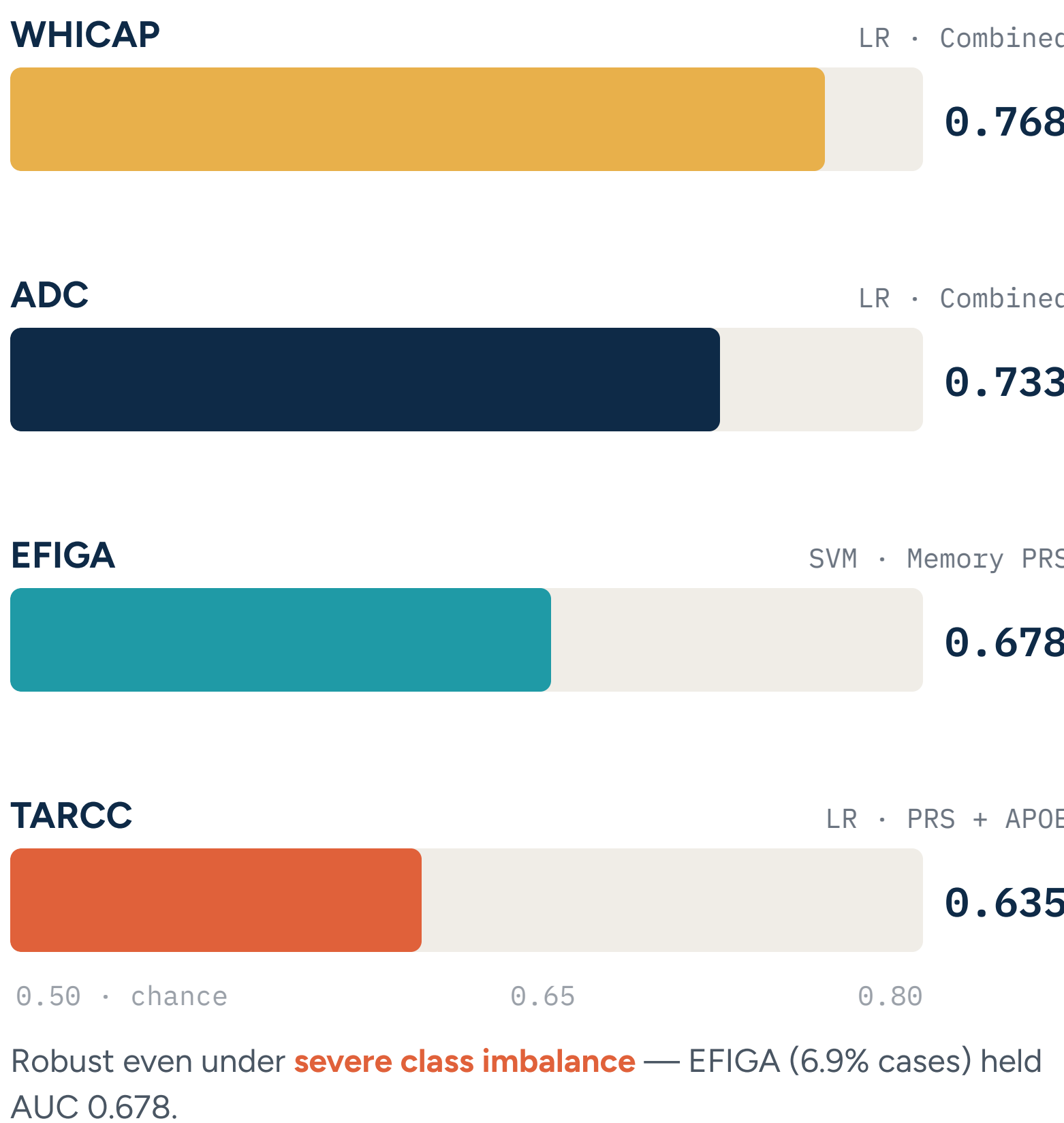
3 The TransPRS Architecture

An end-to-end transformer that jointly predicts **language, memory & executive function**, then extracts trait-specific SNP scores via cross-attention.



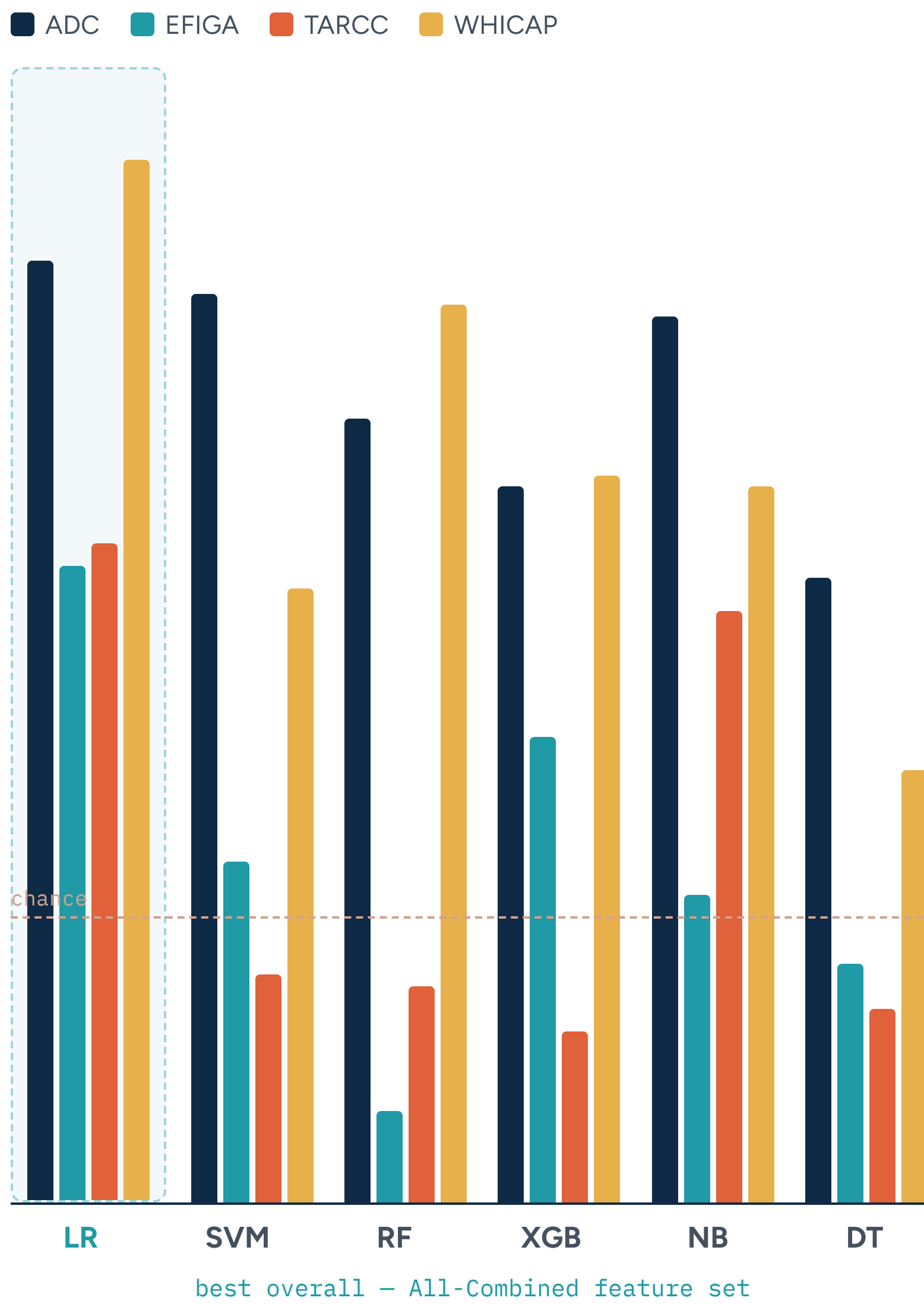
4 Within-Cohort Performance

Best configuration per cohort (5×5-fold CV). **WHICAP** leads with combined PRS + Logistic Regression.



5 Linear Beats Ensembles

Combined feature set across all six classifiers. **Logistic Regression ranked 1st or 2nd in every cohort**; complex ensembles overfit the smaller cohorts.



75%

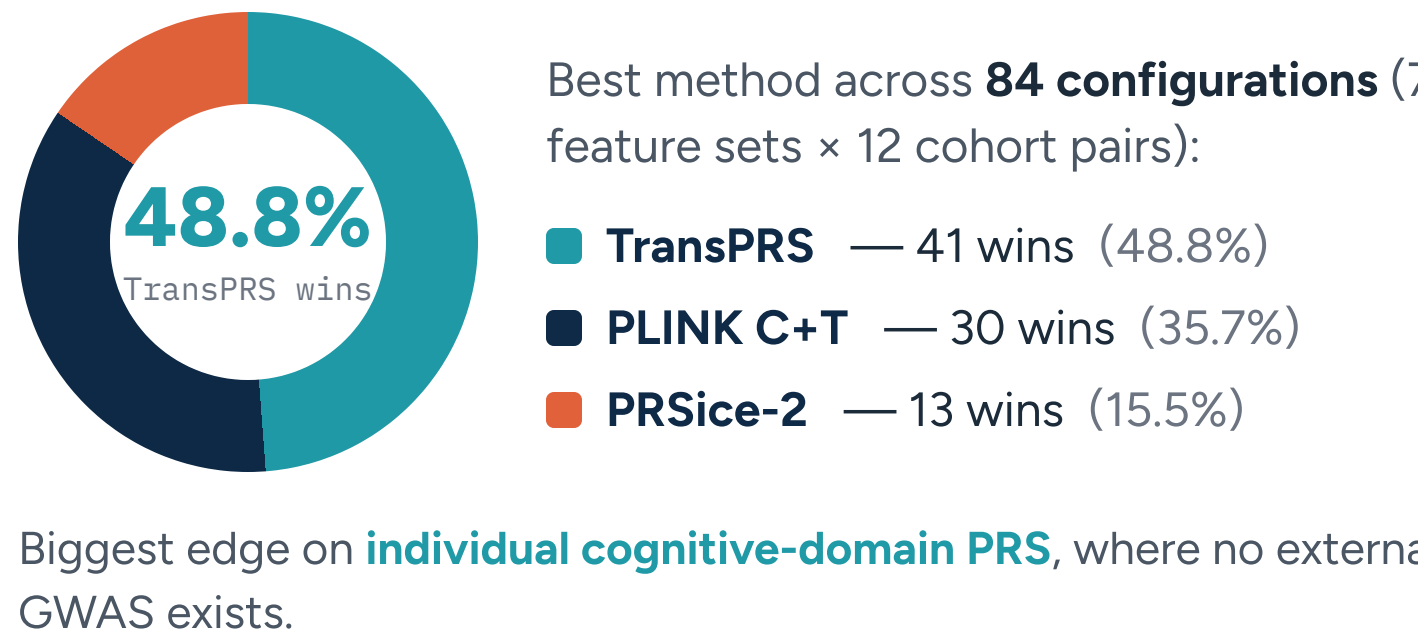
LR ranks best within-cohort configs

100%

LR ranks best cross-cohort configs

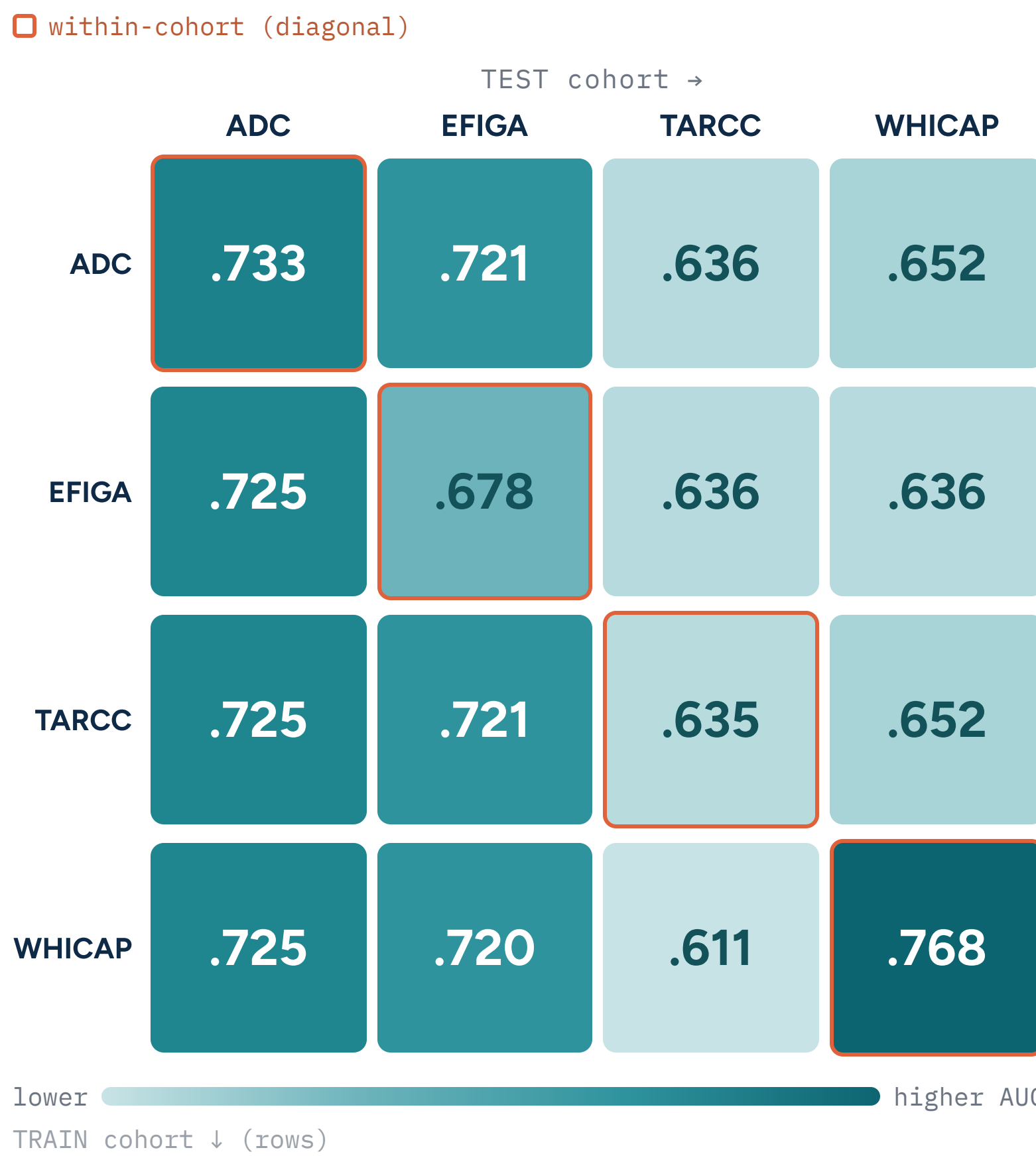
Ensembles (**Random Forest, XGBoost**) collapse on the smallest cohorts — EFIGA & TARCC drop toward chance — classic **overfitting**, while the linear model stays stable.

6 vs. Traditional PRS Tools



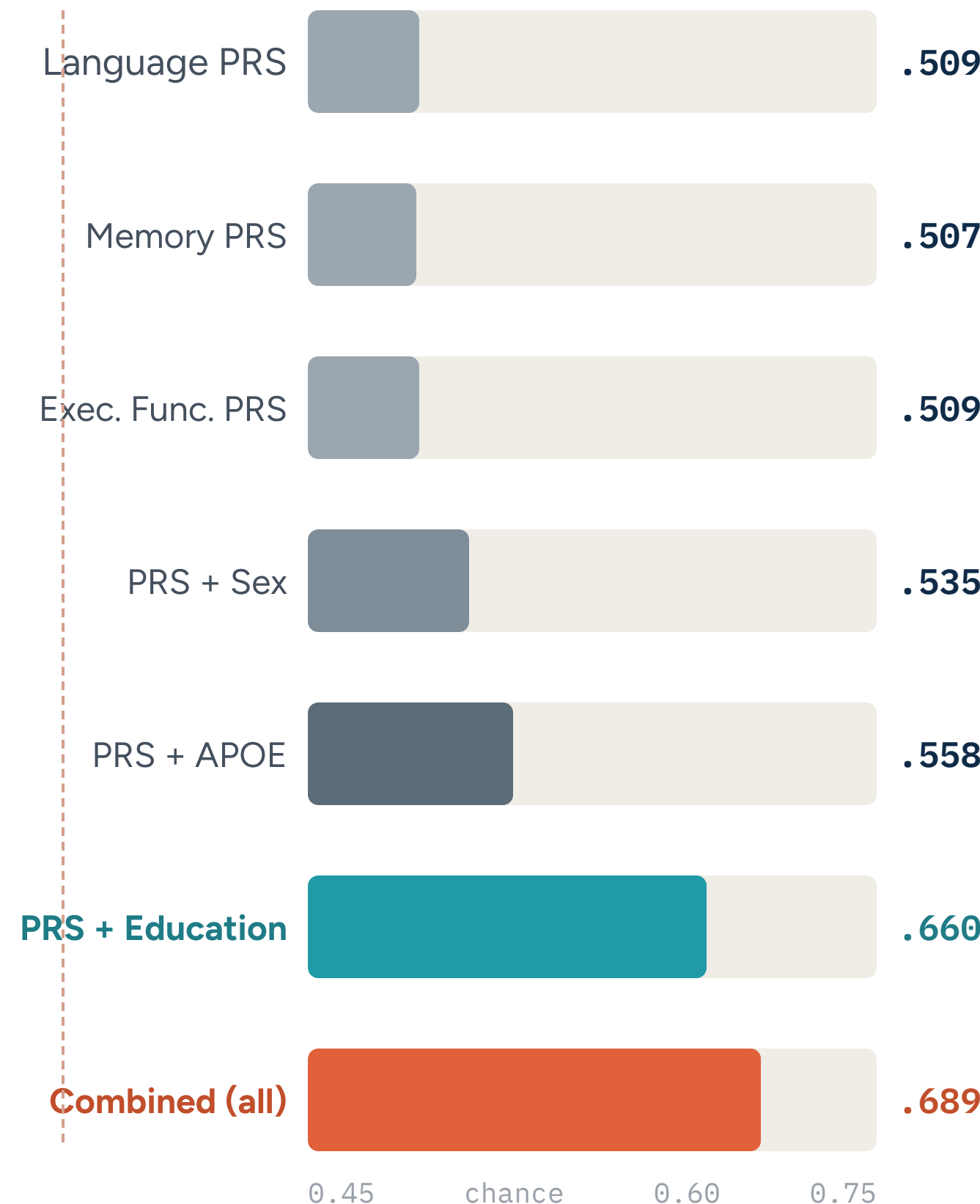
7 Cross-Cohort Generalization

Train on one cohort, test on the other three. Generalization is **strong** — only a 2.2% mean drop.



8 Education Is the Key Covariate

Mean within-cohort AUC by feature set (LR). Single-domain PRS sit near chance; **education delivers the biggest lift, outperforming APOE**.



Conclusions

- TransPRS-based models deliver **reliable AD classification** and strong cross-cohort transfer in Hispanic/Latino populations.
- Education** is an indispensable covariate, surpassing APOE in most cohorts.

- The consistent win of **Logistic Regression** implies aggregated polygenic scores yield approximately linear risk gradients.
- Learning weights from individual-level data **democratizes PRS** for underrepresented populations.

Supporting the clinical utility of PRS-based AD risk stratification in underrepresented populations.