به نام خدا

دانشکده مهندسی کامپیوتر

گزارش جامع پیشرفت و تحلیل علمی پروژه IHARQ BenchGuard

وضعیت تجمیعی، روش‌شناسی و نتایج پس از تکمیل چهار فاز نخست P00–P03

IHARQ BenchGuard: Cumulative Progress and Scientific Analysis through P03

عنوان پروژه:

سنجش و بهبود قابلیت اطمینان رابط مغز و رایانه با کالیبراسیون کم‌هزینه برای توان‌بخشی حرکتی به کمک هوش مصنوعی، راستی‌آزمایی شواهد و تنظیم تطبیقی تصمیم

Assessment and Improvement of Reliability in Brain-Computer Interfaces with Low-Cost Calibration for Motor Rehabilitation using Artificial Intelligence, Evidence Verification, and Adaptive Decision Adjustment

دانشجو: سپهر رجبی

شماره دانشجویی: 4013623008

استاد راهنما: دکتر بهروز شاهقلی

نوع گزارش: گزارش میان‌دوره‌ای پیشرفت، نتایج و تحلیل علمی

پوشش اجرایی: P00–P03 از چرخه P00–P15

تاریخ: مرداد 1405 / August 2026

این گزارش یک جمع‌بندی پژوهش‌محور از وضعیت فعلی پروژه است و جایگزین گزارش نهایی پایان‌نامه یا مجوز ادعای بالینی نیست.

فهرست مطالب

1 خلاصه مدیریتی و وضعیت فعلی 1

1.1 مهم‌ترین دستاوردهای فعلی 1

2 تعریف پروژه، مسئله پژوهشی و اهداف 2

2.1 مسئله اصلی 2

2.2 پرسش پژوهشی مرکزی 2

2.3 اهداف اصلی 2

2.4 نوآوری مورد نظر 3

3 معماری کلان و روش تحقق اهداف 3

3.1 جریان منطقی تصمیم 4

3.2 منطق آزمایش و ابلیشن 5

3.3 واژگان تصمیم در IHARQ 5

3.4 جایگاه Layer 0 و Layer 10 5

3.5 قابلیت‌های مرحله‌ای و توقف‌گاه‌های معتبر 5

4 روش‌شناسی تجربی، حاکمیت شواهد و منطق ارزیابی 6

4.1 زنجیره داده و نقش‌های غیرقابل تعویض 6

4.2 اصل frozen identity و بازتولیدپذیری 6

4.3 واحد تحلیل و جلوگیری از pseudo-replication 7

4.4 خانواده سنجه‌ها: reliability discrimination، و decision 7

4.5 طرح ablation و discipline comparator 7

4.6 منفی‌ها، شکست‌ها و fail-closed به‌عنوان داده علمی 8

4.7 تفکیک canonical، post-hoc و diagnostic 8

5 نقشه راه شانزده‌فازی و سنجش پیشرفت 8

5.1 دو تعریف مکمل از پیشرفت 9

6 فازهای تکمیل‌شده پایه: P00 و P01 10

6.1 P00: زیرساخت حاکمیت، شناسه و بازتولیدپذیری 10

6.2 P01: تبدیل سه دیتاست عمومی به یک substrate مشترک 11

6.2.1 تقسیم‌بندی و جلوگیری از نشت 11

6.2.2 پیش‌پردازش و پنجره رسمی 11

6.2.3 درس مهم از شکست A4 R1 11

7 P02: رمزگشاهای پایه، کم‌برچسب و کنترل‌های A4 12
7.1 بسته‌شدن اجرا 12
7.2 هیچ رمزگشای جهان‌شمولی مشاهده نشد 12
7.3 رفتار کم‌برچسب: بهبود کلی، اما غیرخطی و وابسته به بودجه 13
7.4 حساسیت به policy آموزش 13
7.5 A4: پیچیدگی بیشتر الزاماً بهتر نیست 14
7.6 Extension پسینی Stage18S 14

8 P03: کالیبراسیون، عدم قطعیت و Selective Prediction 14
8.1 A1: کالیبراسیون به‌طور میانگین مؤثر است، اما اجباری نیست 14
8.1.1 بهبود تجمیعی کیفیت احتمال 15
8.2 رابطه با بودجه برچسب 16
8.3 A2: آستانه ثبت‌شده به‌طور قابل اتکا منتقل نمی‌شود 17
8.4 A3: بخشی از uncertainty در task دودویی alias است 18
8.5 خطاهای مطمئن هنوز باقی می‌مانند 18

9 سنتز تجمیعی یافته‌ها و ادعاهای فعلی 19
9.1 یافته‌های تجمیعی با بیشترین ارزش پژوهشی 19
9.2 ادعاهای فعلی که قابل دفاع‌اند 20
9.3 ادعاهایی که هنوز نباید مطرح شوند 20

10 جایگاه پژوهشی فعلی، نوآوری و ارزش علمی برای ادامه کار 20
10.1 تفاوت با یک پروژه صرفاً benchmark محور 21
10.2 بومی‌سازی ایده HARQ بدون قیاس مکانیکی 21
10.3 بومی‌سازی regime risk برای ناپایداری EEG 21
10.4 پیوند نتایج فعلی با ادبیات پایه پروژه 22
10.5 بلوغ پژوهشی حاصل‌شده تا این مرحله 22

11 کار باقی‌مانده و مسیر پیشنهادی از P04 به بعد 23
11.1 گام بعدی: P04 IHARQ-lite evidence verification 23
11.2 معیار موفقیت P04 23
11.3 P05–P07: اعتماد زمانی و تصمیم به تعویق 23
11.4 P08–P13: پیامد، تنش و سازگاری 23
11.5 P14–P15: فازهای سبک‌تر اما ضروری 24

12 محدودیت‌ها، ریسک‌ها و نکات روش‌شناختی 24

12.1 محدودیت‌های فعلی 24
12.2 ریسک‌های ادامه پروژه و کنترل آن‌ها 25
12.3 معیار کیفیت برای گزارش نهایی 25

13 جمع‌بندی نهایی گزارش پیشرفت 25

14 مبنای اسنادی و منابع 27

1 خلاصه مدیریتی و وضعیت فعلی

این پروژه یک چارچوب پژوهشی برای سنجش و بهبود **قابلیت اطمینان تصمیم** در سامانه‌های Brain–Computer Interface (BCI) مبتنی بر EEG است. مسئله محوری صرفاً این نیست که رمزگشا با چه دقتی کلاس حرکتی را پیش‌بینی می‌کند؛ مسئله این است که سامانه بتواند تشخیص دهد **چه زمانی خروجی برای اقدام قابل اتکا است، چه زمانی شواهد ناکافی است، و چه زمانی باید تصمیم را به تعویق انداخت، رد کرد یا مسیر جایگزین را فعال نمود.** این جهت‌گیری، پروژه را از یک مقایسه ساده مدل‌های طبقه‌بندی به یک معماری چندلایه برای reliability-aware decision making تبدیل می‌کند.

وضعیت در یک نگاه

چهار فاز نخست چرخه رسمی شانزده‌فازی، یعنی P00 تا P03، تکمیل شده‌اند. در نتیجه، **پیشرفت شمارشی پروژه برابر 4 از 16، یعنی 25 درصد است.** با این حال این عدد به‌تنهایی بار واقعی کار را نشان نمی‌دهد. دو فاز پایانی، Dashboard and Cards و Final Thesis Integration، عمدتاً فازهای تجمیع، گزارش و بسته‌بندی هستند و P13 نیز یک مقایسه اختیاری OpenSim است. با یک برآورد مدیریتی محافظه‌کارانه که زیرساخت، داده، رمزگشا و لایه اعتماد پایه را سنگین‌تر از فازهای نهایی وزن می‌دهد، **پیشرفت کاری فعلی حدود 30 درصد** برآورد می‌شود. این برآورد یک earned-value رسمی نیست و باید به‌عنوان شاخص برنامه‌ریزی تقریبی خوانده شود.

آنچه تا این نقطه ساخته شده فقط «مقدمه پروژه» نیست. چهار فاز تکمیل‌شده، زنجیره‌ای قابل ردیابی از حاکمیت و قراردادها، داده عمومی و تقسیم‌بندی بدون نشت، رمزگشاهای چندخانواده و آزمایش‌های کم‌برچسب، تا کالیبراسیون احتمال، عدم قطعیت و selective prediction ایجاد کرده‌اند. خروجی فعلی یک **measurement and trust spine** واقعی برای ادامه فازهای IHARQ-lite، RegimeRisk و سیاست تطبیقی است.

1.1 مهم‌ترین دستاوردهای فعلی

- **زیرساخت حاکمیتی و بازتولیدپذیری:** در P00 همه 19 سلول مهندسی ثبت‌شده، 102 آزمون قطعی، 178 دسته ورودی عمداً معیوب، 11 بنیان لایه‌ای و قراردادهای 16 فاز با رفتار fail-closed بسته شدند.
- **زیرلایه داده واقعی:** در P01 سه منبع عمومی EEG با 172 گروه آزمودنی به یک قرارداد مشترک تبدیل شدند؛ 12,910 رویداد پذیرفته‌شده به همان تعداد پنجره رسمی معتبر نگاشت شدند و آزمون‌های نشت و جدایی نقش‌ها بسته شدند.
- **ستون فقرات رمزگشا:** در P02 اندازه‌گیری خام چندخانواده، بودجه‌های کم‌برچسب 1 تا 256 نمونه در هر کلاس، تحلیل A4 و تاریخچه شکست/اصلاح تکمیل شد. نتیجه کلیدی این است که هیچ برنده جهانی در مدل، دیتاست یا بودجه برچسب پشتیبانی نمی‌شود.
- **ستون فقرات اعتماد:** در P03 کالیبراسیون منتخب به‌طور میانگین کیفیت احتمال را بهبود داد، اما انتقال یک آستانه ریسک از validation به TEST به‌شدت ناهمگن بود. این مشاهده عملاً

نشان داد «بهبود کالیبراسیون» و «قابل‌انتقال بودن سیاست پذیرش» دو مسئله جدا هستند.

- **حفظ شواهد منفی:** شکست، نبود آستانه قانونی، پوشش صفر، بدترشدن کالیبراسیون، ناسازگاری ورودی و نتایج ناموفق حذف نشده‌اند. این ویژگی برای یک پروژه پژوهشی جدی مهم است، زیرا مخرج علمی را از selection bias محافظت می‌کند.

نتیجه راهبردی تا این مرحله

پروژه اکنون از مرحله «آیا می‌توان داده و مدل پایه را به شکل معتبر اجرا کرد؟» عبور کرده و به مرحله مهم‌تری رسیده است: **«وقتی رمزگشا یک خروجی می‌دهد، چه سازوکارهایی لازم است تا پیش از تصمیم، کیفیت و کفایت شواهد را بررسی کنیم؟»** پاسخ تجربی P03 نشان می‌دهد یک آستانه اعتماد ساده کافی نیست؛ بنابراین ورود به P04: IHARQ-lite evidence verification از نظر علمی توجیه مستقیم دارد.

2 تعریف پروژه، مسئله پژوهشی و اهداف

2.1 مسئله اصلی

در کاربردهای توان‌بخشی حرکتی، خطای یک رمزگشای قصد حرکتی صرفاً یک خطای طبقه‌بندی نیست. اگر خروجی مدل به فرمان یا اقدام کمکی منتهی شود، کیفیت تصمیم باید پیش از استفاده بررسی گردد. از این رو پروژه به‌جای اتکا به accuracy یا balanced accuracy به‌تنهایی، یک زنجیره تصمیم را دنبال می‌کند که شامل probability calibration، uncertainty، پایداری زمانی، سازگاری شواهد، کیفیت سیگنال، تشخیص تغییر وضعیت و سیاست پذیرش/تعویق/عدم اقدام است.

داده عمومی EEG در این پروژه «محیط آزمون» است، نه مرز مفهومی نهایی. مزیت آن بازتولیدپذیری، دسترسی عمومی و امکان اجرای محاسباتی بدون گردآوری داده انسانی جدید است. بنابراین ادعاهای پروژه در وضعیت فعلی به **آزمایش آفلاین روی بنچمارک عمومی و غیرکلینیکی** محدود می‌شوند.

2.2 پرسش پژوهشی مرکزی

پرسش مرکزی را می‌توان چنین صورت‌بندی کرد:

> آیا ترکیب کالیبراسیون، تحلیل عدم قطعیت، راستی‌آزمایی افزایشی شواهد و تنظیم تطبیقی تصمیم می‌تواند قابلیت اطمینان BCI با کالیبراسیون کم‌هزینه را افزایش دهد؛ به‌گونه‌ای که سامانه علاوه بر پیش‌بینی کلاس، درباره کفایت شواهد برای اقدام نیز تصمیم بگیرد؟

2.3 اهداف اصلی

1. ساخت یک خط لوله بازتولیدپذیر از داده عمومی EEG تا خروجی تصمیم، همراه با شناسه، نسخه، هش، قرارداد نقش داده و کنترل نشت.

2. پیاده‌سازی و مقایسه خانواده‌های کلاسیک، هندسه ریمانی و شبکه‌های عصبی سبک در شرایط FULL_TRAIN و بودجه‌های کم‌برچسب.
3. سنجش کیفیت احتمال با Brier score، NLL، calibration error، نمودارهای قابلیت اطمینان و نرخ خطاهای با اعتماد بالا.
4. ساخت IHARQ به‌عنوان یک لایه بومی‌شده راستی‌آزمایی شواهد؛ نه با «ارسال مجدد» حالت عصبی، بلکه با پنجره بعدی، زیرمجموعه شواهد، مدل/عضو جایگزین و ترکیب شواهد.
5. ساخت RegimeRisk برای پایش ناپایداری زمانی، رانش، افت پیوسته کیفیت و تغییر رژیم اعتماد.
6. ساخت سیاست تطبیقی برای اعمال خروجی‌های چندحالته مانند accept، wait، defer، fallback، no-action یا unsafe.
7. ارزیابی رفتار چارچوب در محیط شبیه‌سازی‌شده حلقه‌بسته، تحت تنش‌های کنترل‌شده و در یک نمایش محدود MyoSuite/OpenSim، بدون تبدیل آن به ادعای درمان یا ایمنی بالینی.

2.4 نوآوری مورد نظر

نوآوری پروژه در یک مدل منفرد خلاصه نمی‌شود. سهم مورد نظر، **معماری ارزیابی و تصمیم‌سازی قابل اعتماد** است: رمزگشا یک منبع شواهد است، نه مرجع نهایی تصمیم. این معماری دو ایده بین‌رشته‌ای را به‌صورت کنترل‌شده بومی می‌کند:

- از HARQ/semantic communications: اصل «پذیرش نکردن پیام فقط به دلیل دریافت شدن» به «پذیرش نکردن خروجی رمزگشا فقط به دلیل تولید شدن» تبدیل می‌شود.
- از مدیریت ریسک سری زمانی ناایستا: مفهوم تغییر رژیم و کاهش اعتماد به مدل به پایش EEG nonstationarity، افت کیفیت، نوسان اعتماد و شکست‌های متوالی تبدیل می‌شود.

این انتقال‌ها تشابه مفهومی هستند و نه ادعای یکسان بودن مسئله‌های مخابرات، بازار مالی و سیگنال عصبی.

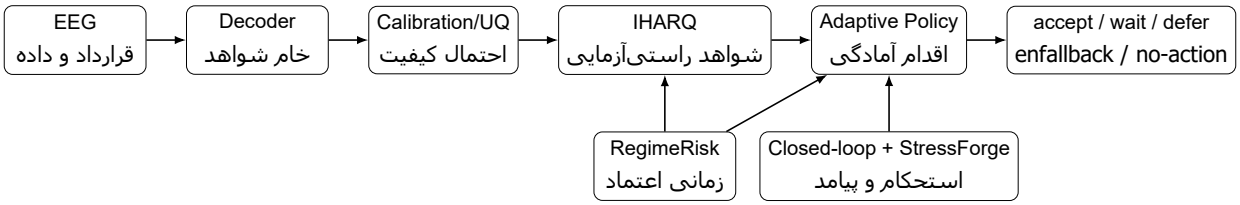
3 معماری کلان و روش تحقق اهداف

معماری نهایی پروژه در یازده لایه Layer 0–10 سازمان‌دهی شده است. هدف این تفکیک آن است که اندازه‌گیری، اعتماد، کنترل، شبیه‌سازی و گزارش‌گری در یکدیگر ادغام مبهم نشوند. هر لایه مالک داده‌ها و ادعاهای خاص خود است و مصرف‌کننده لایه‌های قبل محسوب می‌شود.

لایه	نام رسمی	نقش اصلی
L0	Claim-Safety and Scope Governance	مرزبندی ادعا، محدودیت، شواهد قابل استفاده و جلوگیری از بیش‌ادعایی.
L1	Public-Data and Protocol Anchor	داده، برچسب، تقسیم‌بندی، پیش‌پردازش، پنجره و قرارداد نشت.
L2	Decoder and Baseline Measurement Spine	رمزگشاهای پایه، پیش‌بینی خام، عملکرد و مقایسه‌های کنترل‌شده.
L3	Calibration, Uncertainty, and Selective Prediction	کالیبراسیون، عدم قطعیت، آستانه ثبت‌شده و تحلیل risk–coverage.
L4	IHARQ Evidence Verification	راستی‌آزمایی افزایشی/ترکیبی شواهد و خروجی‌های چندحالته تصمیم.
L5	RegimeRisk Temporal Trust Monitoring	پایش ناپایداری، رانش، نوسان اعتماد و رژیم خطر زمانی.
L6	Adaptive Readiness Policy Layer	سیاست قاعده‌محور، estimator کیفیت، Adaptive-IHARQ، deferral، bandit/RL.
L7	Simulated Closed-Loop Readiness Environment	ارزیابی پیامد تصمیم در حلقه بسته شبیه‌سازی‌شده.
L8	StressForge-Lite Stress Generator	تزریق تنش‌های کنترل‌شده و سنجش استحکام.
L9	MyoSuite/OpenSim Embodiment Demo	نمایش محدود نگاشت تصمیم به اقدام کمکی شبیه‌سازی‌شده.
L10	Dashboard, Cards, and Reproducibility	گزارش خواندنی، کارت‌ها، داشبورد، بسته بازتولید و خروجی پایان‌نامه.

جدول 1: نقشه فشرده لایه‌های پروژه.

3.1 جریان منطقی تصمیم



شکل 1: جریان ساده‌شده از سیگنال تا تصمیم آمادگی اقدام.

3.2 منطق آزمایش و ابلیشن

خط آزمایشی رسمی از A0 تا A13 تعریف شده است: A0 رمزگشای خام، A1 کالیبراسیون، A2 آستانه ساده، A3 عدم قطعیت و selective prediction، A4 کنترل‌های پنجره/ensemble، سپس A5–A13 برای IHARQ، estimator کیفیت، RegimeRisk، learning-to-defer، سیاست تطبیقی، bandit، RL، تنش و embodiment. شناسه A14 عمداً ممنوع نگه داشته شده تا هویت ابلیشن‌ها در نسخه‌های مختلف به‌صورت خاموش جابه‌جا نشود.

در وضعیت فعلی A0–A3 دارای شواهد تجربی مستقیم پروژه‌اند؛ A4 دارای شواهد و زمینه متعلق به لایه 2 است؛ A5–A13 هنوز اهداف فازهای پایین‌دست هستند. این تفکیک از گزارش‌کردن «آمادگی رابط» به‌عنوان «اثبات اثر علمی» جلوگیری می‌کند.

3.3 واژگان تصمیم در IHARQ

برای جلوگیری از ambiguity، لایه IHARQ از های action صریح استفاده می‌کند. vocabulary canonical شامل accept، wait، request_evidence، combine_evidence، fallback، uncertain و unsafe است. این ها action با label حرکتی اشتباه نمی‌شوند: decoder می‌گوید «کدام intent محتمل‌تر است»، اما IHARQ می‌گوید «آیا evidence فعلی برای استفاده از آن کافی است یا نه».

این جدایی از نظر طراحی مهم است. ممکن است یک decoder کلاس را با confidence بالا بدهد ولی policy به دلیل signal-quality issue یا instability آن را uncertain اعلام کند. برعکس، evidence اولیه می‌تواند ضعیف باشد ولی با window بعدی یا source جایگزین به accept برسد. بنابراین لایه decision یک classifier دوم ساده نیست؛ یک controller برای کفایت evidence است.

3.4 جایگاه Layer 0 و Layer 10

Layer 0 و Layer 10 دو مرز کنترلی معماری هستند. Layer 0 اجازه نمی‌دهد statement نهایی از ceiling evidence عبور کند؛ برای مثال یک improvement آفلاین نباید به claim safety تبدیل شود. Layer 10 نیز read-only است: می‌تواند records را validate، summarize، visualize و package کند، اما حق ندارد threshold را metric retune، recompute hidden را failure یا repair را کند.

این تفکیک باعث می‌شود گزارش نهایی خود بخشی از governance باشد. نمودار، card یا dash-board باید از versioned record source تغذیه شود و limitation مرتبط را همراه نتیجه نشان دهد. در نتیجه، layer reporting منبع حقیقت علمی جدید نیست.

3.5 قابلیت‌های مرحله‌ای و توقف‌گاه‌های معتبر

معماری علاوه بر capability phase ها، tier های مرحله‌ای دارد. مفهوم این ها tier آن است که thesis می‌تواند در یک point stopping علمی معتبر جمع‌بندی شود بدون آنکه همه stretch ها component الزاماً موفق شوند. برای مثال اگر IHARQ-lite و RegimeRisk evidence معنادار ایجاد کنند ولی RL benefit اضافه ندهد، نبود superiority در RL نباید foundation قبلی را نامعتبر کند. این طراحی

با ماهیت پژوهش سازگار است: complexity downstream باید ارزش خود را ثابت کند، نه اینکه صرف وجود در roadmap آن را ضروری سازد.

4 روش‌شناسی تجربی، حاکمیت شواهد و منطق ارزیابی

یکی از تفاوت‌های اساسی این پروژه با یک بنچمارک معمول طبقه‌بندی این است که نتیجه علمی فقط از یک جدول performance استخراج نمی‌شود. طراحی از ابتدا به‌گونه‌ای انجام شده که هر ادعا باید به یک زنجیره مشخص از dataset identity، نقش داده، preprocessing، مدل، checkpoint، prediction، metric، مقایسه و limitation متصل باشد. بنابراین روش‌شناسی پروژه را می‌توان در چهار ستون مکمل خلاصه کرد: **ثبات substrate، جدایی نقش‌های داده، مقایسه matched، و نگهداری شواهد منفی**.

4.1 زنجیره داده و نقش‌های غیرقابل تعویض

برای جلوگیری از selection leakage، نقش‌های train، calibration، validation و TEST از هم جدا هستند و در سطح subject تخصیص یافته‌اند. مدل پایه فقط از داده‌ای استفاده می‌کند که نقش آن اجازه training دارد. calibrator روی calibration/validation مجاز fit یا انتخاب می‌شود و آستانه selective نیز روی threshold-validation ثبت می‌شود. TEST فقط برای ارزیابی نهایی همان تصمیم ثبت‌شده مصرف می‌شود؛ اگر rule روی validation وجود نداشته باشد، TEST اجازه ساختن یا نجات rule را ندارد.

این firewall برای سؤال اصلی پروژه حیاتی است. اگر threshold یا calibrator پس از دیدن TEST تنظیم شود، ظاهراً می‌توان risk مطلوب‌تری گزارش کرد، اما دیگر نمی‌توان درباره قابلیت انتقال تصمیم صحبت کرد. در P03 دقیقاً به همین دلیل حالت‌های NO_LEGAL_THRESHOLD و zero TEST coverage به‌عنوان نتیجه معتبر نگه داشته شده‌اند.

4.2 اصل frozen identity و بازتولیدپذیری

هر phase مصرف‌کننده خروجی phase قبل است، اما حق بازتعریف خاموش آن را ندارد. شناسه‌های dataset، label split، map، preprocessing، budget window، model membership، recipe، prediction checkpoint و artifact به‌صورت versioned نگه داشته می‌شوند. این طراحی دو فایده دارد:

- اگر نتیجه پایین‌دست نامطلوب باشد، substrate بالادست برای بهتر کردن نتیجه تغییر نمی‌کند؛
- اگر یک extension جدید اجرا شود، تاریخچه canonical قبلی از بین نمی‌رود و denominator قدیمی با denominator جدید جایگزین مبهم نمی‌شود.

به همین علت، گزارش فعلی هم‌زمان تاریخچه canonical و های‌extension پسینی را نگه می‌دارد. برای نمونه P02 هم سطح historical و هم سطح post-extension 1–256 label/class را مشخص می‌کند و P03 نیز canonical و supplement را جدا گزارش می‌دهد.

4.3 واحد تحلیل و جلوگیری از pseudo-replication

در داده EEG، چندین window از یک participant مستقل از یکدیگر به معنای آماری کامل نیستند. در مقایسه‌های اصلی P02، participant واحد inferential است و windows event-level صرفاً مشاهدات درون participant محسوب می‌شوند. این انتخاب مانع از آن می‌شود که هزاران window از چند فرد محدود به‌طور مصنوعی حجم نمونه را بزرگ نشان دهند.

در P03 تحلیل aggregate روی هایversioned group انجام می‌شود و bootstrap در سطح group اجرا شده است. این سطح با سؤال calibration/threshold سازگار است، زیرا هر group یک ترکیب مشخص از dataset، budget model و lineage تصمیم را نمایندگی می‌کند. در گزارش نهایی نیز سطح واحد تحلیل باید همراه هر نتیجه ذکر شود؛ در غیر این صورت اعداد group participant، و prediction ممکن است به اشتباه با یکدیگر مقایسه شوند.

4.4 خانواده سنجه‌ها: reliability discrimination، و decision

پروژه عمداً یک single score نمی‌سازد. سه خانواده outcome از هم جدا هستند:

1. **Discrimination/performance:** شامل BACC به‌عنوان معیار اصلی، همراه macro-F1، ACC و در هایbranch قانونی ROC-AUC.
2. **reliability: Probability** شامل Brier score، NLL، calibration error و diagnostics کالیبراسیون.
3. **quality: decision Selective** شامل coverage، risk، feasibility، transport rule از vali-dation به TEST، zero-coverage و failure terminals.

این تفکیک مستقیماً با یافته اصلی تا P03 هم‌خوان است: مدل ممکن است discrimination مناسبی داشته باشد ولی score آن calibrated poorly باشد؛ calibration ممکن است بهتر شود ولی threshold به TEST منتقل نشود. بنابراین عبارت کلی «مدل مطمئن‌تر شد» بدون مشخص کردن محور measurement از نظر علمی کافی نیست.

4.5 طرح ablation و discipline comparator

چارچوب A0–A13 برای آن تعریف شده که هر پیچیدگی جدید در برابر baseline مشخص ارزیابی شود. تا این مرحله:

- A0 شواهد decoder خام را در P02 پوشش می‌دهد؛
- A1–A3، threshold selective calibration و uncertainty پایه را در P03 پوشش می‌دهند؛
- A4 در مالکیت P02 نقش کنترل context/window/ensemble دارد؛
- A5–A13 برای هایphase پایین‌دست هستند و هنوز نتیجه تجربی محسوب نمی‌شوند؛
- A14 در stack رسمی مجاز نیست.

منطق مهم این است که هر method جدید باید benefit خود را در یک matched comparison کسب کند. اگر هایkey، dataset، split، window، model، threshold calibration یا cost هم‌تراز نباشند، نتیجه به حالت diagnostic تنزل می‌یابد. این اصل به‌ویژه برای P04 IHARQ-lite ضروری است، زیرا policy پیچیده‌تر نباید صرفاً به دلیل استفاده از evidence بیشتر «بهتر» فرض شود.

4.6 منفی‌ها، شکست‌ها و fail-closed به‌عنوان داده علمی

در پروژه، failure به‌صورت missing silent حذف نمی‌شود. انواع مهم evidence negative شامل موارد زیر هستند:

- input malformed و failure schema در P00؛
- failure feasibility و lineage repair در P01؛
- FAILED، CONDITIONAL_SKIP، INPUT_INCOMPATIBLE و comparison harmful در P02؛
- calibration deterioration، NO_LEGAL_THRESHOLD، zero، ineligible coverage -disagree ment و predictions confident-wrong در P03.

این رویکرد از survivorship bias جلوگیری می‌کند. برای مثال اگر فقط هایی‌group که threshold مناسب دارند گزارش شوند، مسئله feasibility ناپدید می‌شود؛ اگر فقط هایی‌ensemble موفق نگه داشته شوند، هزینه و complexity failure پنهان می‌ماند.

4.7 تفکیک canonical، post-hoc و diagnostic

برای اعتبار علمی، chronology نتیجه بخشی از نتیجه است. هایی‌extension 64/128/256 در P02 و supplement متناظر در P03 resolution علمی را افزایش داده‌اند، اما نباید به‌عنوان evidence از پیش ثبت‌شده بازنویسی شوند. همین اصل برای Stage18S صادق است: این بخش sensitivity و signal جهت اثر را روشن‌تر می‌کند، ولی جایگزین specified prospectively replication نمی‌شود.

قرارداد روش‌شناختی این گزارش

هر عدد در این گزارش باید با سؤال مشخص خوانده شود: **چه جمعیتی؟ چه نقش داده‌ای؟ چه phase و چه status؟ چه comparator؟ و با چه سقف ادعایی؟** هدف از این سخت‌گیری افزایش حجم مستندات نیست؛ هدف این است که در گزارش نهایی بتوان میان «بهبود واقعی»، «نتیجه diagnostic»، «نتیجه منفی» و «آمادگی برای phase بعد» تفاوت روشن ایجاد کرد.

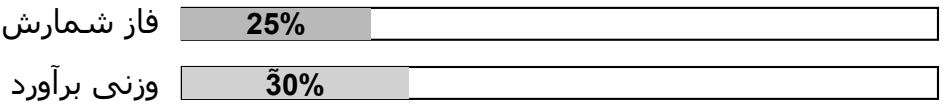
5 نقشه راه شانزده‌فازی و سنجش پیشرفت

چرخه اجرایی رسمی از Phase 0 تا Phase 15 تعریف شده است. در زبان این گزارش، «چهار فاز تکمیل‌شده» یعنی چهار فاز نخست این چرخه: P00, P01, P02, P03. بنابراین فاز بعدی از نظر شماره رسمی Phase 4 / P04 است.

فاز	عنوان	وضعیت	نقش/بار غالب
P00	Repository, configuration, and record schema	**تکمیل**	زیرساخت و حاکمیت
P01	Public data and split protocol	**تکمیل**	داده و قرارداد تجربی
P02	Baseline decoders	**تکمیل**	آموزش/ارزیابی سنگین
P03	Calibration and uncertainty	**تکمیل**	اعتماد پایه و آستانه
P04	IHARQ-lite evidence verification	فاز بعدی	هسته نوآوری راستی‌آزمایی
P05	RegimeRisk temporal trust	برنامه‌ریزی‌شده	تحلیل زمانی/ناایستایی
P06	Evidence-quality estimator and supervised Adaptive-IHARQ	برنامه‌ریزی‌شده	یادگیری کیفیت و سیاست
P07	Learning-to-defer	برنامه‌ریزی‌شده	تصمیم به baseline تعویق
P08	Simulated closed-loop readiness	برنامه‌ریزی‌شده	شبیه‌سازی حلقه‌بسته
P09	StressForge-Lite	برنامه‌ریزی‌شده	استحکام و تنش
P10	Contextual bandit	برنامه‌ریزی‌شده	سازگاری فوری
P11	Reinforcement-learning policy	برنامه‌ریزی‌شده	سیاست ترتیبی
P12	MyoSuite embodiment demo	برنامه‌ریزی‌شده	نمایش embodiment
P13	OpenSim replay or optional comparison	اختیاری/پایین‌دست	مقایسه تکمیلی
P14	Dashboard and cards	پایین‌دست	گزارش/بسته‌بندی، سبک‌تر
P15	Final thesis integration	نهایی	یکپارچه‌سازی پایان‌نامه، سبک‌تر

5.1 دو تعریف مکمل از پیشرفت

1. **پیشرفت شمارشی:** $4/16 = 25\%$. این عدد دقیق و بدون تفسیر است.
2. **پیشرفت وزنی کاری:** حدود **30 درصد**. برای این برآورد، بلوک «زیرساخت/داده» و «رمزگشا/اعتماد پایه» مجموعاً 30 درصد بار کل در نظر گرفته شده‌اند؛ بلوک‌های P04–P07 حدود 28 درصد، P08–P13 حدود 32 درصد و دو فاز نهایی گزارش/یکپارچه‌سازی حدود 10 درصد. این وزن‌دهی فقط برای برنامه‌ریزی است و به‌طور خاص سبک‌تر بودن فازهای نهایی و اختیاری بودن بخشی از P13 را منعکس می‌کند.



شکل 2: پیشرفت شمارشی 25 درصد و برآورد کاری حدود 30 درصد با لحاظ سبک‌تر بودن فازهای نهایی.

نکته برنامه‌ریزی

اگرچه چهار فاز نخست از نظر زیرساختی سنگین بوده‌اند، بخش قابل توجهی از ریسک علمی اصلی هنوز در P04–P11 قرار دارد: آیا راستی‌آزمایی شواهد، اعتماد زمانی و سیاست تطبیقی واقعاً نسبت به هایbaseline ساده سود می‌دهند؟ بنابراین عدد 30 درصد نباید به معنای «یک‌سوم ادعاهای نهایی اثبات شده‌اند» تفسیر شود؛ بلکه بیانگر حجم تقریبی کار بسته‌شده است.

6 فازهای تکمیل‌شده پایه: P00 و P01

6.1 P00: زیرساخت حاکمیت، شناسه و بازتولیدپذیری

P00 عمداً فاز غیرتجربی بود. نقش آن این بود که قبل از ورود به داده و مدل، پروژه دارای قرارداد روشن برای schema، شناسه اجرا، پیکربندی، lineage، lifecycle، اعتبارسنجی، مانفیست، محدودیت و شکست باشد. این تصمیم باعث شده فازهای بعدی نتوانند با تغییر خاموش split، window، budget یا نتیجه ناموفق، نتیجه مطلوب‌تری تولید کنند.

سطح	نتیجه	تفسیر
سلول‌های مهندسی ثبت‌شده	19/19 PASS	بسته‌شدن دامنه ثبت‌شده
آزمون‌های قطعی	102/102 PASS	صحت deterministic در snapshot محلی
ورودی‌های عمداً معیوب	178/178 rejected	رفتار fail-closed بدون پذیرش کاذب
بنیان لایه‌ها	11/11	آمادگی رابطهای Layer 0–10
قرارداد فازها	16/16 dispositioned	آماده‌سازی P00–P15 بدون جعل خروجی آینده
بازتولید محدود	8/8 PASS	بازتولید محلی در snapshot ثبت‌شده

جدول 3: خلاصه بسته‌شدن P00.

نکته مهم P00 این بود که آمادگی A0–A13 به‌صورت READINESS ONLY نگه داشته شد و A14 رد شد. بنابراین در همان ابتدا بین «وجود قرارداد یک آزمایش» و «وجود نتیجه آن آزمایش» مرز صریح ایجاد شد.

6.2 P01: تبدیل سه دیتاست عمومی به یک substrate مشترک

در P01 پروژه وارد داده واقعی شد. سه منبع فعال PhysioNetMI، BNCI2014_001 و Lee2019_MI در یک task دودویی left-hand vs right-hand به قرارداد مشترک تبدیل شدند. ورودی‌های خارج از task به‌عنوان «کلاس منفی» بازتفسیر نشدند؛ بلکه طبق LabelMapRecord خارج از مخرج task باقی ماندند.

شاخص	مقدار	اهمیت
دیتاست فعال / گروه آزمودنی / فایل منبع	3 / 172 / 453	provenance و پوشش بین‌دیتاستی
رویداد پذیرفته‌شده در task	12,910	مخرج رسمی داده
پنجره رسمی معتبر	12,910 / 12,910	حفاظت یک‌به‌یک از مخرج
تقسیم گروه آزمودنی train/cal/val/test	102/35/17/18	جدایی subject-level
پنجره‌های نامعتبر رسمی	0	عبور hard validity
خلاصه کیفیت / soft flags / hard invalid	489 / 20 / 0	سیاست ANNOTATE_NOT_REPAIR
فازهای اجرایی / هاgate / تست‌های regression	27/27, 16/16, 50/50	بسته‌شدن اجرا و package

جدول 4: شاخص‌های اصلی P01.

6.2.1 تقسیم‌بندی و جلوگیری از نشت

تقسیم نقش‌ها در سطح subject انجام شد و گروه‌ها بین train، calibration، validation و test اشتراک نداشتند. علاوه بر disjointness، بررسی duplicate، fit-scope overlap و آلودگی بودجه به TEST انجام شد. نتیجه «نشت قراردادی مشاهده نشد» است؛ این جمله محدود به آزمون‌های پیاده‌سازی‌شده است و ادعای عدم وجود هر نوع نشت قابل تصور نیست.

6.2.2 پیش‌پردازش و پنجره رسمی

هسته شامل EEG-only، ثبت رویداد، demean، average rereference، resampling همزمان event/signal به 160 Hz، فیلتر 8–32 Hz و پنجره +0.5 to +3.5 s با 480 نمونه است. کیفیت به‌جای repair خاموش، annotate شد.

6.2.3 درس مهم از شکست A4 R1

پیشنهاد اولیه پنجره +0.0 to +4.0 s برای A4 روی یک رویداد معتبر نیازمند 80 نمونه غیرموجود بود. clipping padding، یا حذف رویداد مخرج مقایسه را تغییر می‌داد. بنابراین A4 R2 به +0.0 to +3.5 s بازطراحی شد تا همه 12,910 والد حفظ شوند. این یک مثال مهم از رفتار علمی پروژه است:

شکست feasibility به تغییر کنترل‌شده طراحی منجر شد، نه دستکاری داده برای عبور آزمایش.

7 P02: رمزگشاهای پایه، کم‌برچسب و کنترل‌های A4

P02 اولین فاز سنگین عملکرد مدل بود و وظیفه داشت ستون فقرات اندازه‌گیری خام را قبل از افزودن لایه‌های اعتماد بسازد. خانواده‌های کلاسیک/ریمانی و شبکه‌های عصبی/نمایش‌های عمیق در یک substrate ثابت مقایسه شدند. این فاز calibration یا سیاست abstention را ادعا نمی‌کند؛ خروجی آن «رمزگشا و «baseline است.

7.1 بسته‌شدن اجرا

سطح A0 پس از extension دارای **786 سلول terminal** است: 765 SUCCESS، سه FAILED، پانزده CONDITIONAL_SKIP و سه DEPENDENCY_BLOCKED. همچنین 4,590 ردیف metric در سطح participant و 3,057 رکورد metric ایجاد شد. نگه‌داشتن های‌terminal ناموفق بخشی از مخرج است و از تبدیل «موارد سخت» به missing جلوگیری می‌کند.

7.2 هیچ رمزگشای جهان‌شمولی مشاهده نشد

ترتیب مدل‌ها به دیتاست و آزمودنی وابسته است. در FULL_TRAIN، CSP-LDA روی split تک‌آزمودنی TEST در BNCI2014_001 بالاتر است، در حالی که خانواده DNN-SEQ/DBConformer در میانگین participant-first برای Lee2019_MI و PhysioNetMI پیشرو است. روی PhysioNetMI آزمون Friedman ناهمگنی کلی هشت مدل را نشان داد (p=0.004507, Kendall W=0.266805)، اما هیچ مقایسه جایگزین با CSP-LDA پس از Holm به برنده pairwise معنادار تبدیل نشد. برای Lee2019_MI omnibus نیز معنادار نبود (p=0.148455).

این نتیجه از نظر طراحی بعدی مهم است: IHARQ و سیاست تطبیقی نباید بر این فرض ساخته شوند که یک decoder ثابت همیشه بهترین منبع شواهد است.

شکل 3: نمای FULL_TRAIN از میانگین BACC به تفکیک dataset و decoder؛ تغییر ordering بین datasetها یکی از شواهد اصلی ناهمگنی P02 است.

7.3 رفتار کم‌برچسب: بهبود کلی، اما غیرخطی و وابسته به بودجه

سطح کم‌برچسب شامل بودجه‌های 1,2,4,8,16,32,64,128,256 نمونه در هر کلاس است. در 12 مسیر، model×dataset، تمام مسیرهای معتبر BACC در کل بازه 256–1 غیرمونوتون هستند. با وجود این، یک الگوی قوی محلی دیده شد: تمام 12 مسیر از 64 به 128 بهتر شدند و 9 مسیر از 128 به 256 بهتر شدند. با این حال هر 12 نقطه 256 هنوز پایین‌تر از FULL_TRAIN خود هستند؛ میانگین شکاف BACC برابر 0.0525- و میانگین نسبت بازیابی نسبت به FULL حدود 0.9188 است.

این یافته اجازه نمی‌دهد گفته شود «128 برچسب آستانه قطعی» یا «256 معادل FULL» است، زیرا فقط یک مسیر nested membership کنترل‌شده مشاهده شده و هایsubset مستقل تکرار نشده‌اند.

7.4 حساسیت به policy آموزش

یک بخش تشخیصی نشان داد بعضی هاcollapse می‌توانند به training policy مرتبط باشند، نه فقط ظرفیت معماری. CBraMod روی PhysioNetMI از collapse عمومی بازیابی شد و در پنج run تأییدی non-collapsed باقی ماند؛ در مقابل، بازیابی one-shot FBCNet روی BNCI2014_001 در 5 از 5 run تأییدی دوباره به حدود 0.5/collapsed برگشت. بنابراین «یک run خوب» برای ادعای viability کافی نیست.

7.5 A4: پیچیدگی بیشتر الزاماً بهتر نیست

کنترل‌های A4 شامل پنجره طولانی‌تر، multi-window و هایensemble ساده هستند. در 1,080 ردیف C1–C3، 144 p-value جفتی به دست آمد و هیچ مقایسه‌ای پس از Holm معنادار نشد. در C4/C5، یک مقایسه منفی واقعی باقی ماند: hard-vote روی PhysioNetMI FULL_TRAIN در re-peat ۴ نسبت به قوی‌ترین عضو انتخاب‌شده بدتر بود (میانه ΔBACC=-0.04348، Holm p=0.03711).

این نتیجه دو پیام دارد: اول، افزایش تعداد پنجره/عضو به‌خودی‌خود تضمین علمی نیست؛ دوم، baseline ساده باید در فازهای بعدی حفظ شود تا پیچیدگی IHARQ یا policy بدون comparator رشد نکند.

7.6 Extension پسینی Stage18S

تحلیل حساسیت پسینی نشان داد میانگین اثر A4 در بودجه‌های بالاتر از منفی کوچک در 64 به مثبت در 128 و 256 حرکت می‌کند: ΔBACC=-0.00343,+0.00635,+0.02140. در 256، NEURAL C1 میانگین $+0.03291$ داشت و 9 از 9 cell سطح dataset×repeat مثبت بودند؛ با این حال اثر participant-level هنوز ناهمگن است و همه مسیرهای نه‌بودجه‌ای حداقل یک sign flip دارند. بنابراین این بخش یک post-hoc sensitivity است، نه ادعای confirmatory جدید.

جمع‌بندی P02

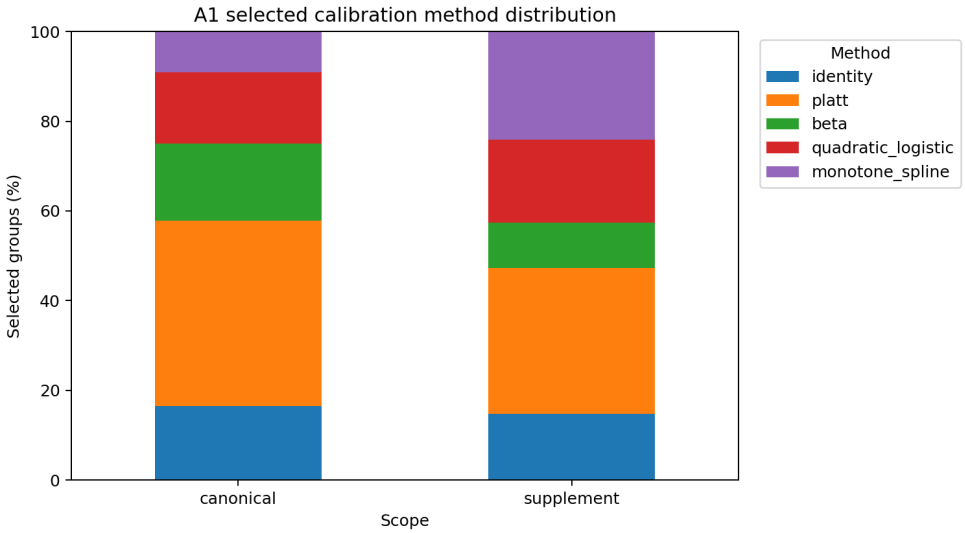
رمزگشاها در این پروژه یک «منبع شواهد ناهمگن» هستند. دیتاست، participant، بودجه برچسب و policy آموزش می‌توانند ترتیب مدل‌ها را تغییر دهند. به همین دلیل، معماری آینده باید کیفیت شواهد را در زمان تصمیم بسنجد و نباید موفقیت FULL_TRAIN یا یک مدل خاص را به شرایط کم‌داده تعمیم دهد.

8 P03: کالیبراسیون، عدم قطعیت و Selective Prediction

P03 مستقیم‌ترین فاز فعلی برای سؤال «آیا اعتماد گزارش‌شده توسط مدل واقعاً قابل استفاده است؟» است. این فاز 25 canonical stage را تکمیل کرده و سه تلاش شکست‌خورده/جایگزین‌شده را نیز حفظ کرده است. جمعیت تحلیل شامل 285 گروه canonical و 108 گروه supplement پس از extension است، بدون exclusion در هسته A1/A2/A3.

8.1 A1: کالیبراسیون به‌طور میانگین مؤثر است، اما اجباری نیست

پورتفوی کالیبراسیون شامل identity، Platt، beta calibration، quadratic logistic و monotone spline است. انتخاب بر اساس threshold-validation NLL و guard بدترشدن انجام می‌شود. در 285 گروه canonical، Platt در 118 گروه برنده است، اما identity نیز در 47 گروه (16.5 درصد) انتخاب می‌شود. این identity branch عمداً وجود دارد تا سیستم مجبور نباشد هر score را recalibrate کند.



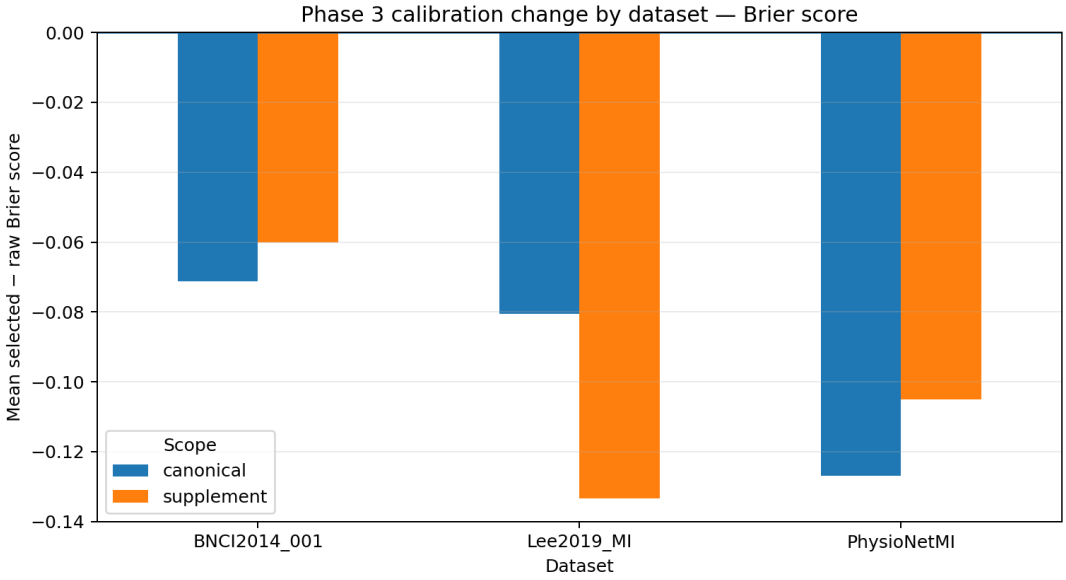
شکل 4: توزیع method منتخب calibration در canonical و supplement؛ هیچ calibrator واحدی winner جهانی نیست.

8.1.1 بهبود تجمیعی کیفیت احتمال

Metric	Raw mean	Selected mean	Δ selected-raw	95% bootstrap CI
Brier	0.5803	0.4875	-0.0929	[-0.1070,-0.0791]
NLL	1.0140	0.6811	-0.3328	[-0.4225,-0.2544]
Calibration error	0.1671	0.0385	-0.1285	[-0.1448,-0.1125]

جدول 5: نتیجه canonical A1 روی TEST؛ CIها با bootstrap گروهی 2000 تکراری محاسبه شده‌اند.

هر سه CI تجمیعی زیر صفر هستند؛ بنابراین در جمعیت frozen آفلاین، کالیبراسیون منتخب به‌طور میانگین Brier، NLL و calibration error را بهبود می‌دهد. اما این میانگین یک تضمین گروهی نیست: 34 گروه در Brier، 29 گروه در NLL و 31 گروه در calibration error بدتر شده‌اند و 47 گروه به دلیل identity/equivalent روی Brier/NLL برابر مانده‌اند. همچنین 44 از 285 گروه canonical status DETERIORATED دارند.

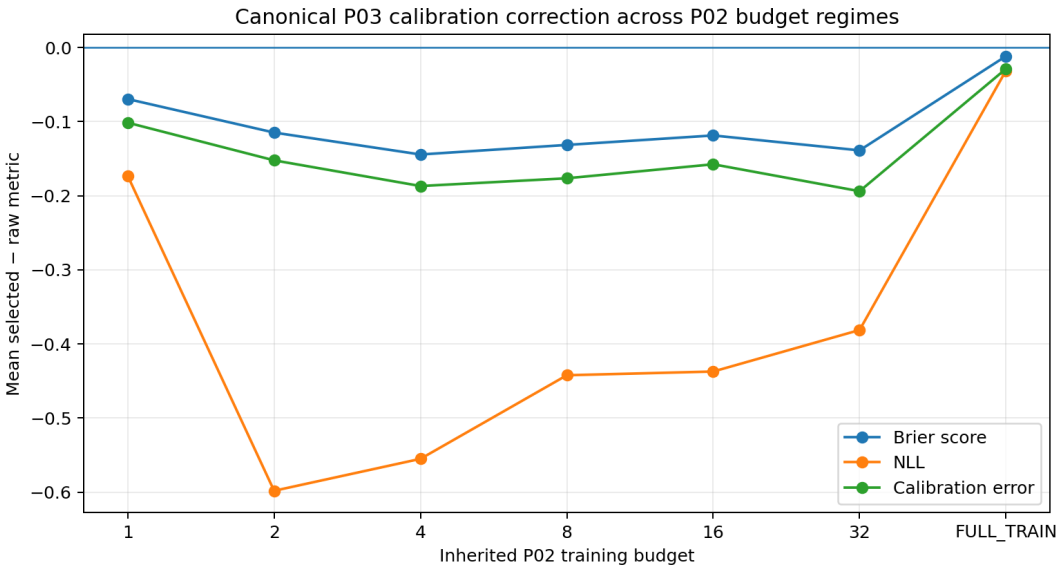


شکل 5: تغییر میانگین Brier به تفکیک دیتاست؛ مقدار منفی به نفع calibrator منتخب است.

در هر سه دیتاست جهت متوسط بهبود یکسان است، اما اندازه اثر متفاوت است: PhysioNetMI بیشترین اصلاح متوسط و BNCI2014_001 کمترین اصلاح را دارد. این ناهمگنی از یک «عدد اعتماد جهانی» حمایت نمی‌کند.

8.2 رابطه با بودجه برچسب

اصلاح کالیبراسیون در FULL_TRAIN به‌طور متوسط بسیار کوچک‌تر از اغلب بودجه‌های low-label است. identity در 23 از 72 گروه FULL (31.9 درصد) ولی در 24 از 213 گروه low-label (11.3 درصد) انتخاب شد. در عین حال deterioration در FULL بیشتر است (25/72، 34.7 درصد) نسبت به low-label مجموع (19/213، 8.9 درصد). این ارتباط توصیفی است و علت آن از این طراحی قابل استخراج نیست.



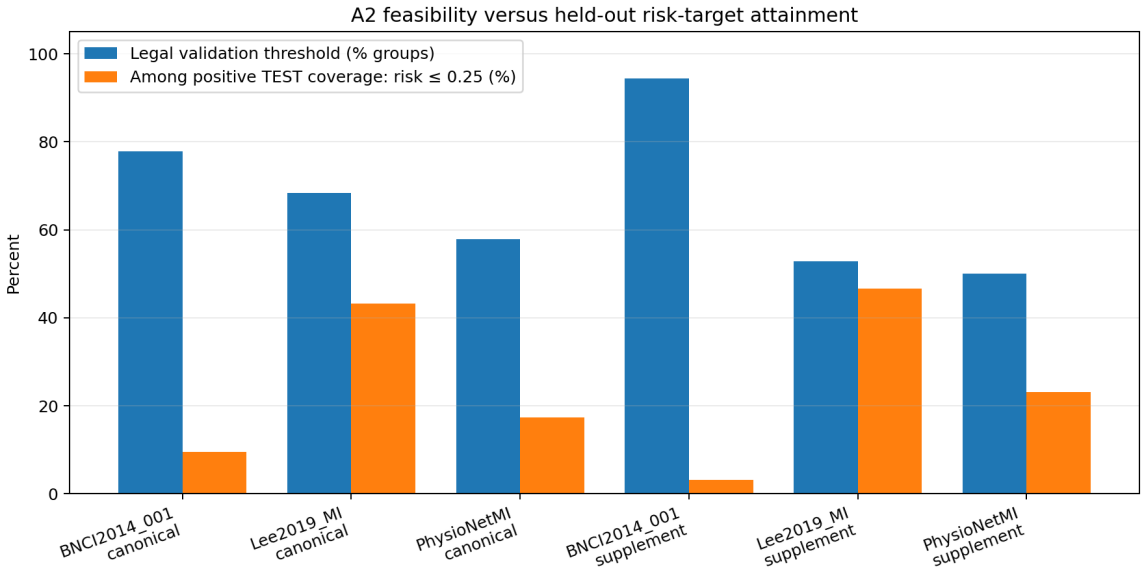
شکل 6: شدت اصلاح کالیبراسیون در بودجه‌های به‌ارث‌رسیده از P02.

8.3 A2: آستانه ثبت‌شده به‌طور قابل اتکا منتقل نمی‌شود

در A2، روی threshold_validation یک confidence floor غیرخالی انتخاب می‌شود که risk حداکثر 0.25 را برآورده کند؛ سپس همان آستانه بدون دست‌کاری روی TEST اعمال می‌شود. اگر هیچ آستانه غیرخالی روی validation وجود نداشته باشد، terminal صحیح NO_LEGAL_THRESHOLD است و TEST حق نجات یا retune ندارد.

Quantity	Canonical result
کل گروه‌ها	285
آستانه قانونی روی validation	194 (68.1%)
NO_LEGAL_THRESHOLD	91 (31.9%)
قانون قانونی با پوشش صفر TEST	30
گروه‌های با پوشش مثبت TEST	164
از میان پوشش مثبت، TEST risk ≤ 0.25	34/164 (20.7%)
میانگین / میانه پوشش TEST در قوانین قانونی	0.301 / 0.058
میانگین / میانه TEST risk در پوشش مثبت	0.412 / 0.400

جدول 6: Feasibility و transport آستانه A2.



شکل 7: تصویر تجمیعی feasibility و transport آستانه‌های A2.

شرطی بر گروه‌های دارای rule قانونی و پوشش مثبت، میانگین تفاوت risk بین A2 و A1 برابر -0.0272 است، اما bootstrap CI برابر [-0.0566,0.0016] صفر را قطع می‌کند. بنابراین حتی پس از حذف no-rule و zero-coverage، ادعای «کاهش متوسط risk به شکل جهانی» پشتیبانی نمی‌شود.

8.4 A3: بخشی از uncertainty در task دودویی alias است

در task دودویی فعلی، confidence، normalized entropy و top-1/top-2 margin از یک score احتمالی واحد مشتق می‌شوند و رتبه‌بندی یکسانی ایجاد می‌کنند. در نتیجه، A3 و A2 در همه ردیف‌های قابل مقایسه برای coverage و finite risk دقیقاً یکسان‌اند. این یک نتیجه مهم منفی است: شمردن این سه feature به‌عنوان سه شاهد مستقل، evidence را مصنوعی بزرگ می‌کند.

شاخه member-disagreement نیز برای همه گروه‌ها ineligible است، زیرا کنترل exact A4 به source مربوط bind نشده است. conformal فقط proxy/diagnostic است و ضمانت پوشش uni-versal ارائه نمی‌کند.

8.5 خطاهای مطمئن هنوز باقی می‌مانند

بهبود میانگین calibration به معنای حذف خطاهای پرخطر نیست. در threshold تشخیصی confi-dence حداقل 0.90، 253 پیش‌بینی اشتباه canonical و 313 پیش‌بینی اشتباه supplement باقی مانده‌اند.

قوی‌ترین نتیجه علمی P03

کالیبراسیون احتمال مفید است، اما مسئله تصمیم را حل نمی‌کند. در همان داده‌ای که کیفیت احتمال به‌طور میانگین بهبود یافته، آستانه ریسک ثبت‌شده در 31.9 درصد گروه‌ها اصلاً feasible نیست و در بخش بزرگی از موارد feasible نیز رفتار مورد انتظار روی TEST منتقل نمی‌شود. این شکاف دقیقاً فضایی است که P04 IHARQ-lite باید با راستی‌آزمایی چندمنبعی و افزایشی شواهد بررسی کند.

9 سنتز تجمیعی یافته‌ها و ادعاهای فعلی

چهار فاز نخست را باید یک زنجیره واحد خواند، نه چهار پروژه مستقل:

P00 foundation → P01 data contract → P02 decoder evidence → P03 trust evidence

این زنجیره بدون تغییر خاموش substrate پیش رفته است. نتیجه فعلی اجازه می‌دهد چند ادعای پژوهشی با سطح اطمینان مناسب مطرح شوند، اما هم‌زمان سقف ادعاها نیز روشن است.

9.1 یافته‌های تجمیعی با بیشترین ارزش پژوهشی

1. **تفکیک لایه‌ها از نظر تجربی ضروری است.** عملکرد طبقه‌بندی، probability calibration و انتقال policy آستانه سه مسئله جدا هستند. یک decoder با BACC خوب لزوماً score قابل اعتماد ندارد؛ یک calibrated score نیز لزوماً threshold قابل انتقال ایجاد نمی‌کند.
2. **دیتاست یک effect modifier جدی است.** هم ترتیب مدل‌های P02 و هم شدت correction و feasibility/transport در P03 به dataset وابسته‌اند. بنابراین رتبه‌بندی یک‌بعدی «بهترین دیتاست/مدل» قابل دفاع نیست.
3. **هیچ برنده universal در decoder یا calibrator وجود ندارد.** این نتیجه به نفع معماری adaptive و evidence-aware است، زیرا policy آینده باید با heterogeneity کنار بیاید.
4. **برچسب بیشتر کمک می‌کند ولی قانون ساده monotonic ندارد.** مسیرهای low-label در کل بازه non-monotonic هستند؛ 64 به 128 در مسیر frozen مفید بود، اما 256 هنوز به FULL نرسیده است. همچنین calibration correction در FULL و low-label رفتار متفاوتی دارد.
5. **پیچیدگی بدون baseline ممنوع است.** ensemble و multi-window در A4 سود تضمین‌شده ندادند و یک ensemble حتی به‌طور corrected بدتر بود. این درس باید در IHARQ، bandit estimator، و RL حفظ شود.
6. **شواهد منفی بخشی از نتیجه هستند.** no-rule، zero-coverage، deterioration-fail، ure-ineligible و incompatible نباید از denominator حذف شوند. این اصل در ادامه برای جلوگیری از winner's curse و report bias حیاتی است.

9.2 ادعاهای فعلی که قابل دفاع‌اند

در نگارش پایان‌نامه یا ارائه میان‌دوره‌ای می‌توان با qualification مناسب گفت:

- پروژه تا P03 یک زنجیره بازتولیدپذیر و fail-closed از داده عمومی تا شواهد اعتماد ایجاد کرده است.
- در جمعیت P03 frozen، calibrator منتخب به‌طور میانگین Brier/NLL/calibration error را بهبود می‌دهد و این جهت در هر سه dataset دیده می‌شود.
- این بهبود universal نیست؛ هایsubgroup بدترشده و identity winners وجود دارند.
- یک threshold ثبت‌شده با target risk روی validation در تعداد قابل توجهی از گروه‌ها وجود ندارد یا روی TEST همان رفتار risk–coverage را حفظ نمی‌کند.
- در task دودویی فعلی، چند feature معمول uncertainty از نظر رتبه‌بندی alias هستند و evidence مستقل محسوب نمی‌شوند.
- افزایش پیچیدگی decoder/ensemble/post-processing به‌صورت خودکار benefit ایجاد نمی‌کند.

9.3 ادعاهایی که هنوز نباید مطرح شوند

- «سامانه برای توان‌بخشی ایمن است» یا «ریسک بالینی را کاهش می‌دهد»؛ داده فعلی آفلاین و غیرکلینیکی است.
- «آستانه 0.25 روی TEST تضمین می‌شود»؛ این target فقط قاعده انتخاب validation است و transport آن محدود بوده است.
- «یک مدل یا calibrator بهترین است»؛ شواهد برعکس، heterogeneity را نشان می‌دهند.
- «256 label/class برابر FULL است» یا «monotonic learning curve است».
- «IHARQ، RegimeRisk یا policy تطبیقی مؤثر است»؛ این‌ها هنوز فازهای پایین‌دست‌اند.
- «نتیجه A4 Stage18S confirmatory است»؛ این بخش post-hoc sensitivity است.

سقف شواهد فعلی

سقف علمی فعلی: **public-benchmark, offline, non-clinical, non-deployment**. این سقف ضعف گزارش نیست؛ نشانه کنترل ادعا است. ارزش علمی فعلی در فهم reliability و failure modeهای تصمیم روی داده عمومی قرار دارد، نه در اثبات درمان یا کنترل واقعی اندام.

10 جایگاه پژوهشی فعلی، نوآوری و ارزش علمی برای ادامه کار

در پایان P03، پروژه هنوز ادعای تکمیل نوآوری اصلی IHARQ/Adaptive Decision را ندارد؛ با این حال، زیرساخت و نتایج فعلی از سطح یک پروژه صرفاً اجرایی فراتر رفته‌اند. ارزش اصلی وضعیت فعلی

در این است که مسئله reliability با چند mode failure واقعی و قابل اندازه‌گیری صورت‌بندی شده است و هایphase بعدی دیگر بر یک فرض انتزاعی بنا نمی‌شوند.

10.1 تفاوت با یک پروژه صرفاً benchmark محور

کارهای benchmark در BCI معمولاً بر مقایسه ،هاdecoder هاdataset و هایperformance metric تمرکز دارند. این پروژه آن پایه را حفظ می‌کند، اما سؤال را یک گام جلوتر می‌برد: **آیا خروجی decoder در یک point operating مشخص برای تصمیم قابل استفاده است؟** این تغییر سؤال باعث می‌شود ،calibration، threshold uncertainty، evidence negative transport و pol-failure icy به اندازه accuracy اهمیت پیدا کنند.

از این منظر، contribution بالقوه پروژه نه «ساخت بزرگ‌ترین neural network»، بلکه ایجاد یک reliability-aware evaluation and decision architecture است که مدل پایه را فقط یکی از منابع ev-idence می‌داند. داده عمومی EEG امکان می‌دهد این ایده با هزینه و ریسک پایین به‌صورت reproducible آزمایش شود، بدون آنکه از این نتایج مستقیماً ادعای safety clinical استخراج شود.

10.2 بومی‌سازی ایده HARQ بدون قیاس مکانیکی

در مخابرات، گیرنده می‌تواند کیفیت پیام را ارزیابی کند، retransmission بخواهد یا چند دریافت را combine کند. در BCI نمی‌توان همان state neural را دقیقاً دوباره ارسال کرد. بنابراین معادل پژوهشی پیشنهادی، درخواست evidence بعدی یا جایگزین است: window زمانی بعدی، chan-nel، family model subset دیگر، disagreement معتبر، indicator quality یا ترکیب چند evidence source.

اهمیت یافته P03 برای این ایده روشن است. وقتی threshold ساده در بخشی از هاgroup وجود ندارد یا روی transport TEST نمی‌شود، یک لایه verification evidence می‌تواند سؤال معناداری داشته باشد: آیا evidence اضافی واقعاً هایgroup شکست‌خورده را نجات می‌دهد یا فقط failure را با complexity بیشتر پنهان می‌کند؟ این سؤال برای P04 یک فرضیه مستقیم و قابل falsify ایجاد می‌کند.

10.3 بومی‌سازی regime risk برای ناپایداری EEG

پیشنهاده پروژه از منطق مدیریت ریسک سری‌های زمانی ناایستا برای پایش تغییر وضعیت استفاده می‌کند. ترجمه مورد نظر، کپی یک مدل مالی نیست؛ بلکه استفاده از مفهوم regime change برای monitor کردن نوسان ،entropy confidence، trend، disagreement کیفیت سیگنال و زنجیره هاfailure است. P05 RegimeRisk باید نشان دهد آیا history زمانی اطلاعاتی فراتر از snapshot فعلی می‌دهد یا خیر.

تا این مرحله، داده P00–P03 هنوز effectiveness online/session-shift را اثبات نکرده است. بنابراین ارزش علمی RegimeRisk در وضعیت فعلی یک **فرضیه طراحی grounded** است، نه یک result. این مرزبندی برای حفظ اعتبار پایان‌نامه ضروری است.

10.4 پیوند نتایج فعلی با ادبیات پایه پروژه

چند خط پژوهشی در اسناد مبنا برای پروژه مهم‌اند: هایbenchmark بازتولیدپذیر مانند MOABB، مدل‌های compact مانند EEGNet، روش‌های EEG convolutional، quan- uncertainty decoding، test-time tification، simulation BCI closed-loop adaptation و کاربردهای توان‌بخشی. پروژه این حوزه‌ها را جایگزین نمی‌کند؛ بلکه تلاش می‌کند شکاف بین benchmark decoder و decision readiness را به‌صورت یک pipeline قابل ارزیابی صورت‌بندی کند.

نتایج فعلی با چند پیام شناخته‌شده این ادبیات سازگارند: heterogeneity dataset/subject جدی است، روش کلاسیک هنوز baseline قابل‌اعتنایی است، complexity neural همیشه برتری تضمین‌شده نمی‌دهد و nonstationarity مانع تعمیم ساده است. سهم خاص پروژه باید از این نقاط فراتر برود و در هایphase بعد نشان دهد آیا verification evidence و ben- policy adaptive efit افزوده واقعی ایجاد می‌کنند.

10.5 بلوغ پژوهشی حاصل‌شده تا این مرحله

از دید گزارش کارشناسی، چهار phase نخست چند مهارت پژوهشی را هم‌زمان پوشش داده‌اند: طراحی، calibra- learning، deep ML، classical reproducibility، governance، data experiment، tion، statistical uncertainty، analysis failure comparison و control. claim این ترکیب برای ادامه پژوهش در مقطع کارشناسی ارشد ارزشمند است، زیرا مسیر بعدی را از «اجرای مدل» به «طراحی و آزمون یک hypothesis درباره «reliability منتقل می‌کند.

اما مهم است که این بلوغ با ادعای کامل شدن پروژه اشتباه نشود. بخش اصلی originality تجربی هنوز در P04–P11 قرار دارد. در نتیجه، وضعیت فعلی از نظر پژوهشی مناسب است: foundation و evidence failure به اندازه کافی قوی شده‌اند که هایphase نوآورانه روی substrate معتبر اجرا شوند، ولی هنوز outcome آن هاphase از پیش فرض نشده است.

صورت‌بندی contribution در وضعیت فعلی

در پایان P03 می‌توان contribution فعلی را چنین بیان کرد: **یک زنجیره reproducible و claim-bounded برای اندازه‌گیری، reliability probability decoder و selective- transport threshold روی سه منبع عمومی EEG ایجاد شده و این زنجیره نشان داده است که بهبود calibration به‌تنهایی برای reliability decision کافی نیست.** contribution نهایی پروژه فقط در صورتی قوی‌تر می‌شود که P04+ نشان دهند evidence verification و policy تطبیقی این شکاف را با مقایسه matched و بدون retuning TEST-side کاهش می‌دهند.

11 کار باقی‌مانده و مسیر پیشنهادی از P04 به بعد

11.1 گام بعدی: P04 IHARQ-lite evidence verification

فاز بعدی باید از نتیجه P03 شروع کند: threshold ساده در بسیاری از گروه‌ها یا feasible نیست یا transport ضعیفی دارد. بنابراین P04 نباید فقط یک threshold پیچیده‌تر بسازد؛ باید بررسی کند آیا ترکیب چند نوع evidence می‌تواند تصمیم را بهتر govern کند.

ورودی‌های مناسب IHARQ-lite شامل confidence/calibrated confidence، entropy/margin با رعایت alias، کیفیت سیگنال، ثبات بین پنجره‌های نزدیک، availability اعضای معتبر، توافق/اختلاف مدل‌ها و reason codeهای ناپایداری است. خروجی باید از ابتدا چندحالته باشد: accept، wait/request more، evidence، defer، fallback، no-action و هایterminal unsafe/blocked.

11.2 معیار موفقیت P04

موفقیت صرفاً افزایش accuracy نیست. مقایسه باید در برابر A2 و هایbaseline ساده انجام شود و حداقل این پرسش‌ها را پاسخ دهد:

- آیا در coverage مشابه، error/risk کاهش می‌یابد؟
- آیا در risk مشابه، coverage قابل استفاده افزایش می‌یابد؟
- آیا no-rule و zero-coverage صرفاً به شکل پنهان منتقل می‌شوند یا واقعاً با شواهد اضافه حل می‌شوند؟
- هزینه محاسباتی و latency اضافه‌شده نسبت به benefit چقدر است؟
- کدام گروه‌ها ضرر می‌کنند و آیا fallback ایمن/شفاف وجود دارد؟

11.3 P05–P07: اعتماد زمانی و تصمیم به تعویق

P05 باید RegimeRisk را از تحلیل لحظه‌ای به هایtrend زمانی گسترش دهد؛ P06 کیفیت شواهد و Adaptive-IHARQ را به‌صورت supervised بسازد؛ و P07 یک learning-to-defer baseline مستقل ایجاد کند تا policy تطبیقی صرفاً با rule-based baseline مقایسه نشود.

11.4 P08–P13: پیامد، تنش و سازگاری

پس از تثبیت evidence policy، P08 رفتار حلقه‌بسته شبیه‌سازی‌شده، P09 تنش‌های StressForge-Lite، P10 contextual bandit و P11 سیاست RL را ارزیابی می‌کنند. P12 نمایش MyoSuite را اضافه می‌کند و P13 OpenSim در صورت زمان و ارزش علمی کافی اختیاری است. ترتیب اهمیت باید حفظ شود: اگر policy پایه benefit معناداری ندارد، پیچیدگی RL/embodiment نباید صرفاً برای تکمیل roadmap اجرا شود.

11.5 P14–P15: فازهای سبک‌تر اما ضروری

دو فاز نهایی عمدتاً evidence presentation، dashboard/cards، traceability، reproducibility package و یکپارچه‌سازی متن پایان‌نامه‌اند. این فازها از نظر محاسبات تجربی سبک‌ترند، اما حذف‌شدنی نیستند، زیرا ادعاهای نهایی باید به هایartifact دقیق و هاlimitation متصل باشند.

اولویت اجرایی پیشنهادی

اولویت فعلی باید روی **P۰۴ و سپس P۰۵** باشد. این دو فاز مستقیماً فرضیه معماری پروژه را آزمایش می‌کنند: آیا به‌جای اعتماد به score واحد، راستی‌آزمایی چندمنبعی و اعتماد زمانی می‌تواند کیفیت تصمیم را در نقاطی که threshold ساده شکست می‌خورد بهتر کند؟

12 محدودیت‌ها، ریسک‌ها و نکات روش‌شناختی

12.1 محدودیت‌های فعلی

- داده‌ها عمومی، آفلاین و عمدتاً مبتنی بر motor imagery EEG هستند؛ نتیجه به بیمار، درمان یا محیط واقعی تعمیم مستقیم ندارد.
- BNCI2014_001 در split فعلی فقط یک participant در TEST دارد؛ استنباط بین‌آزمودنی آن محدود است.
- low-label budgetها بر یک frozen nested membership بنا شده‌اند؛ عدم قطعیت ناشی از sub-های set مستقل هنوز اندازه‌گیری نشده است.
- برخی هاextension مانند Stage18S و هایsupplement 64/128/256 پسینی هستند و chronology آن‌ها نباید به preregistration تعبیر شود.
- در P03 آماره‌های confirmatory که در قرارداد اولیه قابل بازیابی نبوده‌اند نباید اختراع یا بازسازی ظاهری شوند؛ هایbootstrap تحلیلی گزارش‌شده جایگزین آن قرارداد گمشده نیستند.
- uncertainty شاخه‌های confidence/entropy/margin در task دودویی مستقل نیستند؛ برای richer evidence به disagreement واقعی، multiclass یا منبع دیگری نیاز است.
- threshold transport فعلی failure mode را نشان می‌دهد، اما mechanism آن—finite support، nonstationarity، participant shift یا geometry—هنوز به‌صورت causal تفکیک نشده است.

12.2 ریسک‌های ادامه پروژه و کنترل آن‌ها

ریسک	پیامد	کنترل پیشنهادی
گسترش بیش از حد IHARQ/policy	پیچیدگی بدون benefit روشن	ablation baseline-first، و stop rule
overfit estimator/policy	بهبود ظاهری روی validation	frozen TEST، matched keys، repeat و negative terminals
ناپایداری زمانی	افت انتقال policy	RegimeRisk و تحلیل session/sequence
هزینه محاسباتی	latency بالا	گزارش cost/coverage/risk و مقایسه با baseline
کاهش interpretability	دفاع دشوار پایان‌نامه	rule-based anchor و reason code
وسوسه ادعای بالینی	بیش‌ادعایی	حفظ Layer ۰ claim ceiling
کمبود زمان برای شبیه‌سازی سنگین	نیمه‌کاره ماندن مسیر downstream	اولویت P۰۴–P۰۷؛ اختیاری نگه‌داشتن P۱۳ و scope محدود embodiment

جدول 7: ریسک‌های اصلی ادامه کار و کنترل‌های مناسب.

12.3 معیار کیفیت برای گزارش نهایی

گزارش نهایی باید نه فقط بهترین عدد، بلکه denominator کامل، failure state، repeat، limitation، source lineage و comparator را نشان دهد. در این پروژه، نتیجه منفی معتبر از نتیجه مثبت بدون provenance ارزشمندتر است. این رویکرد برای یک خروجی پژوهش‌محور و مناسب ادامه تحصیل در سطح کارشناسی ارشد نیز مهم است، زیرا کیفیت استدلال و کنترل روش‌شناسی را در کنار performance برجسته می‌کند.

13 جمع‌بندی نهایی گزارش پیشرفت

پروژه در پایان P03 از مرحله ایجاد زیرساخت عبور کرده و اکنون یک زنجیره تجربی قابل استفاده برای مطالعه reliability دارد. P00 قراردادها و رفتار fail-closed را بست؛ P01 داده سه‌منبعی، pre- split، processing و denominator را تثبیت کرد؛ P02 نشان داد رفتار decoder به dataset، participant، supervision و policy آموزش وابسته است؛ و P03 نشان داد بهبود probability calibration، هرچند قوی و قابل اندازه‌گیری است، به‌تنهایی یک selective decision policy قابل انتقال ایجاد نمی‌کند.

از نظر پیشرفت، چهار فاز نخست از شانزده فاز رسمی تکمیل شده‌اند: 25 درصد بر اساس شمارش فاز و حدود 30 درصد بر اساس برآورد وزنی کار با لحاظ سبک‌تر بودن فازهای نهایی. مهم‌تر از درصد، این است که هایdependency لازم برای P04 اکنون وجود دارند و پرسش علمی بعدی با داده واقعی روشن شده است.

وضعیت علمی فعلی

آنچه اکنون می‌دانیم: reliability را نمی‌توان با accuracy یا confidence خام خلاصه کرد؛ calibration مفید اما غیرجهانی است؛ threshold ساده transport ضعیف/ناهمگن دارد؛ evidence منفی و failure modeها برای طراحی policy ضروری‌اند.
آنچه هنوز باید نشان دهیم: آیا IHARQ-lite، RegimeRisk و سیاست تطبیقی می‌توانند این شکاف را با benefit قابل اندازه‌گیری، هزینه قابل قبول و بدون overfitting پر کنند؟

بنابراین پروژه از نظر پژوهشی در نقطه مناسبی قرار دارد: خروجی‌های فعلی به اندازه کافی قوی هستند که مسئله reliability را با شواهد واقعی صورت‌بندی کنند، اما هنوز فضای اصلی نوآوری—راستی‌آزمایی شواهد و تصمیم تطبیقی—در فازهای بعدی باقی مانده است. این توازن بین دستاورد واقعی و ادعای محدود، مبنای مناسب برای ادامه پروژه و تبدیل آن به گزارش نهایی پایان‌نامه است.

14 مبنای اسنادی و منابع

این گزارش با تلخیص و هم‌گذاری کنترل‌شده اسناد ارائه‌شده برای پروژه تهیه شده است. محتوای آن جایگزین اسناد authority نیست؛ در صورت اختلاف، نسخه‌های اصلی پروژه و هایartifact machine-readable مرجع باقی می‌مانند. مهم‌ترین اسناد داخلی استفاده‌شده عبارت‌اند از:

1. فرم پیشنهاده پروژه کارشناسی: «سنجش و بهبود قابلیت اطمینان رابط مغز و رایانه با کالیبراسیون کم‌هزینه برای توان‌بخشی حرکتی به کمک هوش مصنوعی، راستی‌آزمایی شواهد و تنظیم تطبیقی تصمیم».
2. Master Architecture Specification — IHARQ BenchGuard, Layers 0–10.
3. Canonical Artifact Record and Interface Registry.
4. Execution and Evidence Plan.
5. Experiment, Ablation, and Evaluation Protocol.
6. Complete Phase Execution Playbook — P00–P15.
7. Integrated Layers 0–10 Method Selection and Design Rationale Register.
8. Integrated Layers 0–10 Detailed Design and Nuts-and-Bolts Specification.
9. Cumulative Phase Evidence, Results and Interpretation through P03, Integrated R4 Completeness Reverified و فایل‌های CSV/JSONL/figures/validation همراه آن.

منابع علمی کلیدی در مبنای پروژه

مراجع

[1] V. Jayaram and A. Barachant, “MOABB: Trustworthy Algorithm Benchmarking for BCIs,” *Journal of Open Source Software*, vol. 3, no. 30, p. 1266, 2018.

[2] MOABB, “Mother of All BCI Benchmarks Documentation.” [Online]. Available: https://moabb.neurotechx.com/docs/index.html

[3] S. Chevallier et al., “The Largest EEG-based BCI Reproducibility Study for Open Science,” arXiv:2404.15319, 2024.

[4] V. J. Lawhern et al., “EEGNet: A Compact Convolutional Network for EEG-based Brain-Computer Interfaces,” arXiv:1611.08024, 2016.

[5] R. T. Schirrmeister et al., “Deep Learning with Convolutional Neural Networks for EEG Decoding and Visualization,” *Human Brain Mapping*, vol. 38, no. 11, pp. 5391–5420, 2017.

[6] “EEG Foundation Models: Progresses, Benchmarking, and Open Problems,” arXiv:2601.17883, 2026.

[7] “Uncertainty Quantification for Motor Imagery BCI,” arXiv:2507.07511, 2025.

[8] “Calibration-free Online Test-time Adaptation for EEG Motor Imagery Decoding,” arXiv:2311.18520, 2023.

[9] J. Shin, D. Suma, and B. He, “Closed-loop Motor Imagery EEG Simulation for Brain-Computer Interfaces,” *Frontiers in Human Neuroscience*, 2022.

[10] “A Closed-loop Adaptive Brain-Computer Interface Framework: Improving the Classifier with Error-Related Potentials,” Microsoft Research.

[11] “Brain-computer Interface-triggered Functional Electrical Stimulation Therapy for Rehabilitation after Stroke.” [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC10605532/

[12] F. Liang, X. Chen, B. Li, and B. Yang, “Brain-Computer Interface Combined with Functional Electrical Stimulation for Post-Stroke Upper Limb Motor Recovery: A Systematic Review and Meta-Analysis,” *Clinical EEG and Neuroscience*.

[13] H. Lorach et al., “Walking Naturally after Spinal Cord Injury Using a Brain-Spine Interface,” *Nature*, vol. 618, pp. 126–133, 2023.

[14] V. Caggiano et al., “MyoSuite: A Contact-rich Simulation Suite for Musculoskeletal Motor Control,” arXiv:2205.13600, 2022.

[15] M. A. Dembia et al., “OpenSim Moco: Musculoskeletal Optimal Control,” *PLOS Computational Biology*, 2020.

[16] “Joint Source-Channel-Check Coding with HARQ for Reliable Semantic Communications,” arXiv:2603.23869, 2026.

[17] “Semantic HARQ for Intelligent Transportation Systems: Joint Source-Channel Coding-Powered Reliable Retransmissions,” arXiv:2504.14615, 2025.

[18] “Online Conformal Model Selection for Nonstationary Time Series,” arXiv:2506.05544, 2025.

[19] A. L. Goldberger et al., “PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals,” *Circulation*, vol. 101, no. 23, pp. e215–e220, 2000.

[20] MOABB, “BNCI2014_001 Dataset Documentation.” [Online]. Available: https://moabb.neurotechx.com/docs/generated/moabb.datasets.BNCI2014_001.html

نحوه استفاده از منابع در این گزارش

منابع علمی بالا برای positioning و تعریف روش‌های پایه استفاده می‌شوند، اما اعداد تجربی گزارش حاضر از هایartifact خود پروژه استخراج شده‌اند. هیچ عدد performance خارجی با اعداد IHARQ به‌عنوان comparison مستقیم هم‌سطح تلقی نشده است، زیرا prepro- split، cessing، metric task و contract evaluation ممکن است متفاوت باشند.