

Mass General Brigham

Division of Pharmacoepidemiology

Harmonizing the Assessment of Real-World Evidence Study Submissions

A Roundtable Discussion

Thursday, October 29, 2026
9:00 AM to 3:00 PM ET

Scan the QR code to join the Q&A on Slido



Shirley Wang



Sebastian Schneeweiss



Laura Pizzi



Wei Hua



Takashi Waki



Wei Liu



Stephen Motsko



Katsiaryna Bykov



Steve Farmer



Marie Bradley



Patrice Verpillat



Páll Jónsson



Stine Hasling Mogensen



Mina Tadrous



Jie Yang



Jiageng Wu



Arup Pramanik



Fatma Shebl



Amy Mulick



Carol Koro



Daniel Prieto-Alhambra

Session 1: Updating the HARPER template

Brief overview and summary of recommendations



Shirley Wang

Associate Professor of Medicine, Harvard Medical School; Brigham and Women's Hospital

Scan the QR code to join the Q&A on Slido

Panelists



Wei Hua

Director, Division of Epidemiology I, Office of Pharmacovigilance and Epidemiology, Office of Surveillance and Epidemiology, Center for Drug Evaluation and Research, FDA, United States



Wei Liu

Director R&D Data, Data Science & AI, Johnson & Johnson Innovative Medicine



Takashi Waki

Epidemiologist, Pharmaceuticals and Medical Devices Agency (PMDA), Japan



Stephen Motsko

Associate Vice President, Center for Observational Research, Amgen

Moderator



Laura Pizzi

Chief Science Officer, International Society for Pharmacoeconomics and Outcomes Research (ISPOR)

TWG1

What we heard — and where we go next

Updating HARPER to improve protocol quality, transparency, usability, and reproducibility

TWG1 makes the science explicit.
TWG2 makes the review explicit.
TWG3 makes both easier to do consistently.

HARPER

HARmonized Protocol to Enhance Reproducibility

ispe ISPOR
Improving healthcare decisions

ORIGINAL ARTICLE | Open Access

HARmonized Protocol Template to Enhance Reproducibility of hypothesis evaluating real-world evidence studies on treatment effects: A good practices report of a joint ISPE/ISPOR task force

Shirley V. Wang, Anton Pottegård, William Crown, Peter Arlett, Darren M. Ashcroft, Eric I. Benchimol, Marc L. Berger, Gracy Crane, Wim Goettsch, Wei Hua, Shaum Kabadi, David M. Kern ... See all authors

What is HARPER?

A protocol template for RWD studies that supports reproducibility, transparent communication, and shared understanding through a common textual, tabular, and visual framework.

EMA GVP
Module VIII PASS

ISPE GPP

NEST-cc

STaRT-RWE

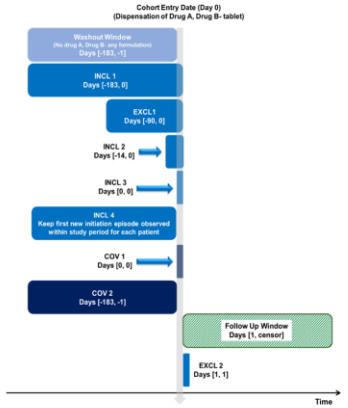
	EMA GVP Mod VIII PASS (Rev 3, Oct 2017)	ISPE GPP (June 2015)	NESTcc (Feb 2020)	STaRT-RWE (2021)	Harmonized Template
High level summary	Largely free text, with guidance on what to include under section headers - details in the *Guidance for the format and content of the protocol of non-interventional post-authorisation safety studies (www.ema.europa.eu)*	Largely free text, with guidance on what to include under section headers - details in report https://www.pharmacoepi.org/resources/policies/guidelines-08027/	Largely free text, with guidance on what to include under section headers - details in NESTcc report	Structured tables that lay out operational parameters to be specified - details in Wang et al, BMJ	Combination of free text with structured tables under section headers
Section Header					
Title page with administrative information (e.g. title, registry ID, drug/device ID, sponsor)	1	A, C, M	2, 10	Table 1	1, table
Table of contents	2			Table of Contents	2, table
Abbreviations	3			Table 9	
Glossary of terminology				Table 8	
Responsible parties	4	B		Table 1	1, table
Abstract	5	D			2, free text
Amendments and updates	6	L		Table 2	3, table
Milestones/timeline	7	E			4, table
Rationale and bac		G	1		5, structured free text
Research questio		F	3	Table 1	6, table
Study design		H1	7	Figure 1, Table 3	7.1, 7.2, free text, table
Setting		H2	4	Table 3A, 3B, 3C, 3D, Table 6	7.3, free text, table
Variables		H4	5, 6	Table 3B, 3E & 3F, 3G, 3H	7.4, free text, table
Device descriptio			2		
Data sources		H3		Table 3A, Appendices	7.5, free text, table
Study size		H5	9	Table 7	7.6, free text, table
Data managemer		H6, H7		Table 3A, Appendices	7.7, free text, table
Data analysis			12	Table 4, Table 5	7.8, free text, table
Quality control			11		
Limitations of the					
Other aspects					
Protection of human subjects			8	Table 1	9, free text
Management and reporting of					10, free text
Plans for disseminating and co					
References					References
Appendices				Appendices	Appendices
ENCePP Checklist for study pro					Appendix

Shaded gray area within bold b



And more…

New initiator, active comparator cohort design evaluating risk of outcome Y for Drug A vs Drug B



INCL/EXCL = Inclusion/Exclusion
Assessment Window
INCL1
Medical and Drug coverage (45 day gaps allowed)
INCL2
Community acquired pneumonia diagnosis and chest radiography
INCL3
Age between 18-65
INCL4
Keep first new initiator episode
EXCL1
Inpatient hospitalization
EXCL2
Censored on day follow up starts

Censoring
183 days
30-Sep-15
Discharged dead
Disenroll medical or drug coverage (45 day gaps allowed)

COV = Covariate Assessment Window
COV1
Age (continuous)
Gender

COV2
Metastatic cancer
Tumor
Arrhythmia
Congestive heart failure
Dementia
Renal failure
Weight loss
Hemiplegia
Alcohol abuse
Number of inpatient hospitalizations
Number of outpatient visits
Number of emergency department visit
Number of unique generics
Prior prescription of penicillins
Prior prescription of cephalosporins
Prior prescription of sulfonamides
Prior prescription of tetracyclines
Prior prescription of aminoglycosides
Pregnancy at time of initiation

Free-text — Context and Rationale

Tables and figure — Operational Definitions

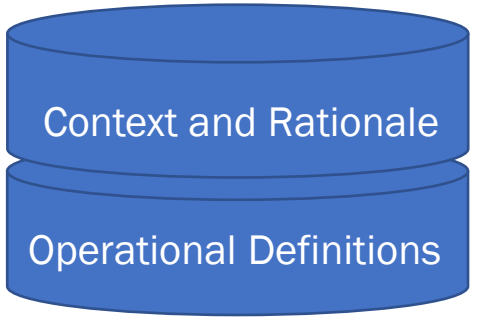
Study population name(s)	Day 0 Description	Number of entries	Type of entry	Washout window	Care Setting[1]	Code Type[2]	Diagnosis position	Incident with respect to…	Measurement characteristics	Source of algorithm
<Text>	<Text>	<drop down>	<drop down>	Number range	<Text>	<Text>	<drop down>	<Text>	<Text>	<Text>
Exposure	*Date of incident dispensation for Drug A (tablets only)*	*Single*	*Incident*	*[-183, 0]*	*n/a*	*NDC*	*n/a*	*Drug A or B (any formulation)*	*Unknown*	*Investigator review of generic names*
Comparator	*Date of incident dispensation for Drug B (tablets only)*	*Single*	*Incident*	*[-183, 0]*	*n/a*	*NDC*	*n/a*	*Drug A or B (any formulation)*	*Unknown*	*Investigator review of generic names*

Exclusion criteria	Code	Code Category	Code Type	Description	Route	Strength
Drug A	00003016050	NDC	11	Awesome Drug	tablet	250mg
Drug B	00003016050	NDC	11	Really Cool Drug	tablet	250mg
Drug A	01003016050	NDC	11	Awesome Drug oral	oral suspension	5 ml
Drug B	00803016050	NDC	11	Really Cool Drug injectable	injectable	5 y injectable
Chest radiography	87.44	PX	09	Routine chest x-ray, so described		
Chest radiography	87.49	PX	09	Other chest x-ray		
Chest radiography	71010	PX	C4	Radiologic examination, chest; single view, frontal		
Chest radiography	71015	PX	C4	Radiologic examination, chest; stereo, frontal		

7.6 Data Sources

Free-text

Table



7.6.1 Context and rationale for data sources

Reason for selection: <Text>

Strengths of data source(s): <Text>

Limitations of data source(s): <Text>

Data source provenance/curation: <Text>

Table 12. Metadata about data sources and software

	Data 1	Data 2	Data 3	Data 4
Data Source(s):	XYZ database version 5.1.2	XYZ database version 5.1.2	n/a	n/a
Study Period:	January 1, 2003 - September 30, 2015	January 1, 2012 – December 31, 2020		
Eligible Cohort Entry Period:	January 1, 2003 - September 30, 2015	January 1, 2012 - September 30, 2015		
Data Extraction Date/Version:	January 1, 2018	January 1, 2021		
Data sampling/extraction criteria:	All enrollees in data source between January, 2003 - September 30, 2015	All enrollees with <1 month Part A, B, D coverage between Jan 2012 and Dec 2020 and <1 diabetes diagnosis		
Type(s) of data:	Commercial claims	Commercial claims		
Data linkage:	None	Linkage to National Death Index. See appendix for details		
Data conversion:	ABC Common Data Model version 7.0	ABC Common Data Model version 7.0		
Software for data management:	Statistical software version 9.4	Statistical software version 9.4		

Since publication in 2022, HARPER has been cited or recommended in numerous guidances

- **Centers for Medicare and Medicaid Services (CMS)**
  Guidance Document: Study Protocols That Use Real-world Data (2025)
- **International Council of Harmonization (ICH)**
  M14 A guide on general principles on plan, design and analysis of pharmacoepidemiological studies that utilize real-world data for safety assessment of medicines. (2024)
- **Council for International Organizations of Medical Sciences (CIOMS)**
  Working Group report: Real-world data and real-world evidence in regulatory decision making (2024)
- **Canadian Agency for Drugs and Technologies in Health (CADTH)**
  Guidance for Reporting RWE (2023)
- **National Institute for Health and Care Excellence (NICE)**
  RWE framework (2023)
- **ENCePP**
  Guide on Methodological Standards in Pharmacoepidemiology (2023)
- **FDA Sentinel Innovation Center**
  A PRocess guide for INferential studies using healthcare data from routine ClinIcal Practice to evaLuate causal Effects of Drugs (PRINCIPLED) BMJ (2023)
- **DARWIN**
  Adopted as template for generic protocols (2023+)
- **More…**

Modified nominal group technique



Kick-off
June

TWG meetings with
14 global stakeholders
July - September

Round-table
October

Five themes

01

Preserve the core structure but refine

Stronger instructions + rationale

02

Make context and rationale more explicit

Decision relevance + transportability

03

Strengthen data fitness assessment

Landscape + feasibility + study-specific fitness

04

Build a resource library

Examples + reusable implementation support

05

Use AI to improve consistency/ implementation

AI + future proofing

01 **Preserve the core structure but refine**

02 Make context and rationale more explicit

03 Strengthen data fitness assessment

04 Build a resource library

05 Use AI to improve consistency/ implementation

The recurring problem is less often a missing section than incomplete or superficial use of existing sections.

01 **Preserve the core structure but refine**

02 Make context and rationale more explicit

03 Strengthen data fitness assessment

04 Build a resource library

05 Use AI to improve consistency/ implementation

More guidance

≠

more boxes

- ✓ Improve the quality and transparency of scientific reasoning.
- ? For each new topic: revised instructions, worked examples?

01 Preserve the core structure but refine

02 **Make context and rationale more explicit**

03 Strengthen data fitness assessment

04 Build a resource library

05 Use AI to improve consistency/ implementation

CONTEXT OF USE

What decision will this evidence inform?

Proposed sections to add:

+ “Context of use / intended decision”
+ “Transportability/relevance to target pop”

Evidentiary bar depends on context of use

- Strengthen structure for discussion of how limitations or anticipated biases affect interpretation for context of use

01 Preserve the core structure but refine

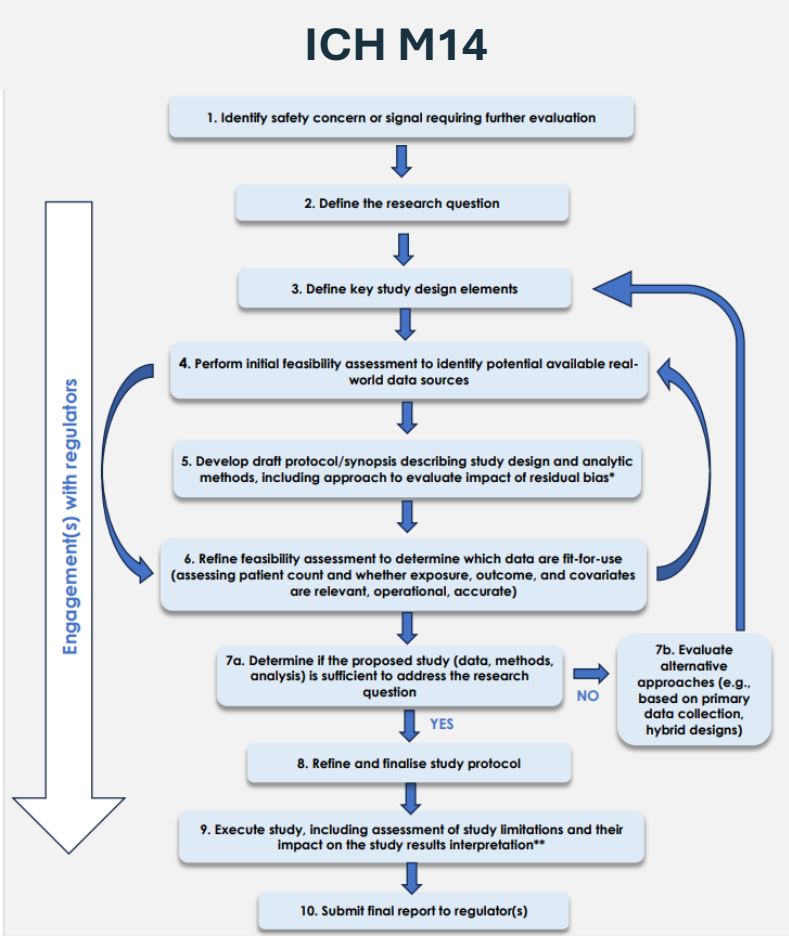
02 Make context and rationale more explicit

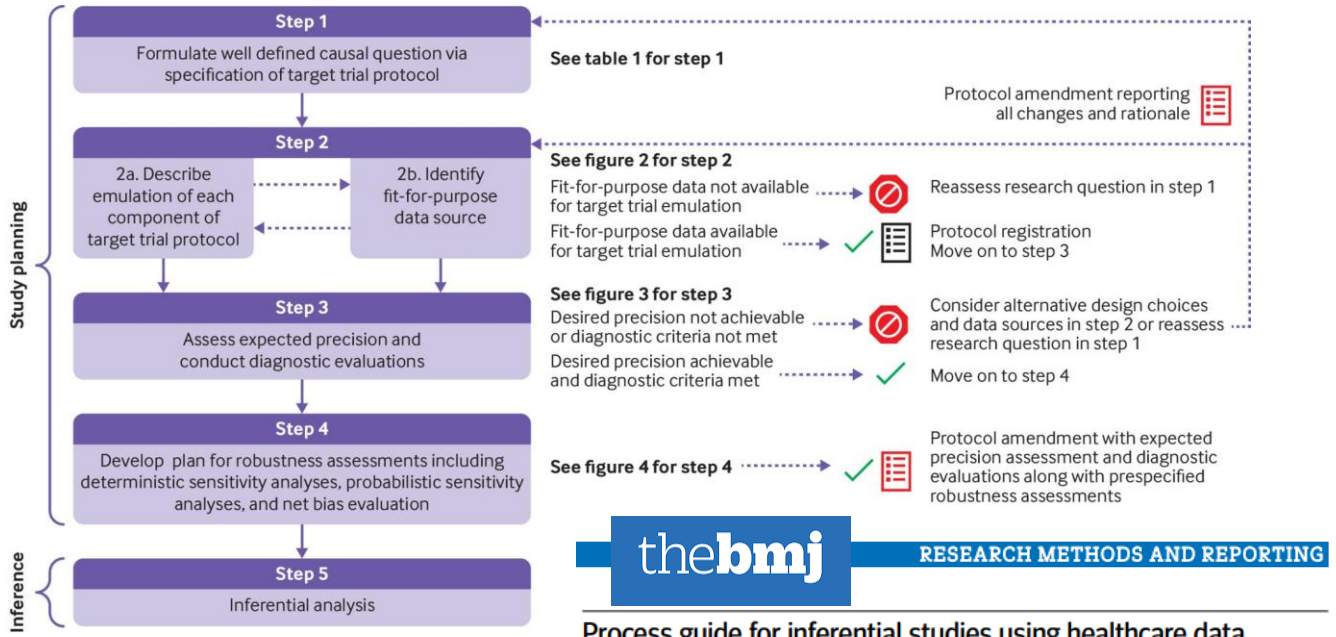
03 Strengthen data fitness assessment

04 Build a resource library

05 Use AI to improve consistency/ implementation

Translate existing guidance into practical application







01 Preserve the core structure but refine

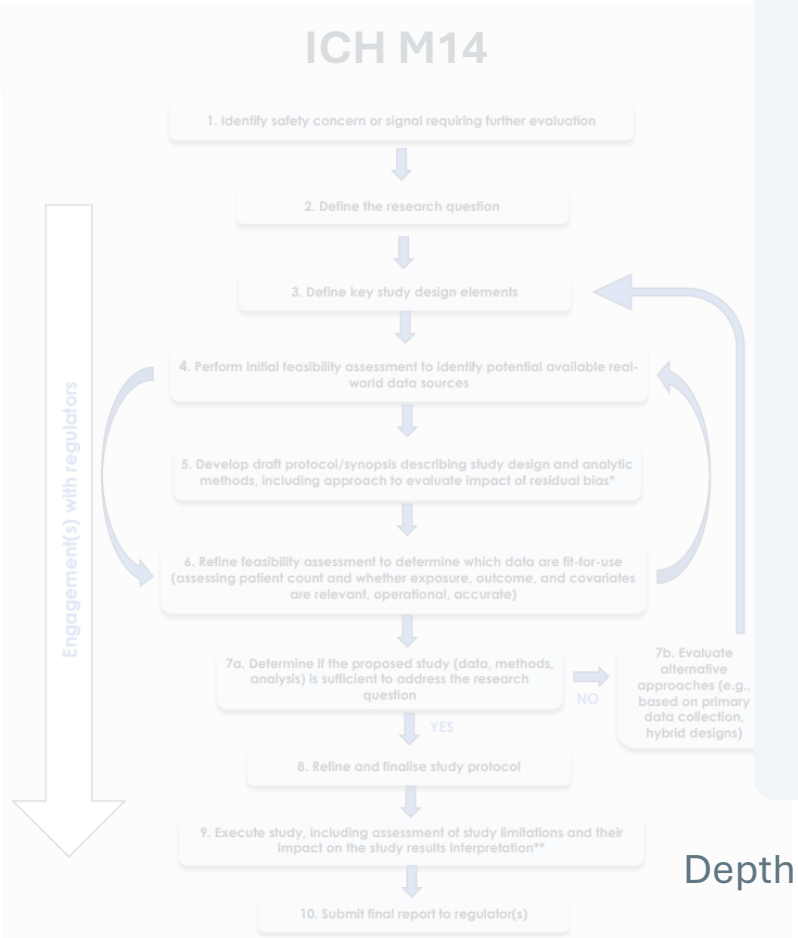
02 Make context and rationale more explicit

03 Strengthen data fitness assessment

04 Build a resource library

05 Use AI to improve consistency/ implementation

Translate existing guidance into practical application



Data and feasibility checks are iterative

Landscape screen

↑↓

Initial protocol

↑↓

Detailed feasibility

↑↓

Full draft

Depth should be proportional to context of use and evidentiary stakes.

01 Preserve the core structure but refine

02 Make context and rationale more explicit

03 Strengthen data fitness assessment

04 **Build a resource library**

05 Use AI to improve consistency/ implementation

Start with 3–5 curated examples[1]



[1]Target trial emulation + external-control studies were suggested as strong early priorities.

01 Preserve the core structure but refine

02 Make context and rationale more explicit

03 Strengthen data fitness assessment

04 Build a resource library

05 **Use AI to improve consistency/ implementation**

Good starting points for automation

Completeness • cross-section consistency • contradictions

Human responsibility remains central

scientific rationale • decisions • expert review

Proposed TWG1 next steps

1	**Refine core structure**	Review each section for conceptual + operational specification and clearer instructions. Include summary with estimand/TARGET reporting. Make clear that N/A can be used as appropriate.
2	**Add decision-relevance context**	Draft concise context-of-use / intended-decision and transportability sections. Strengthen discussion of limitation relevance to context of use.
3	**Strengthen data fitness assessment**	Include structured landscape + study-element-specific fitness assessment and iterative feasibility as instructions/example/annex?
4	**Build resource library**	Build the first 3–5 decision-grade examples and reusable diagrams/resources.
5	**Prototype + test**	Test HARPER AI with users/reviewers.

Today:

→ validate themes

→ identify what's missing

→ prioritize

Session 1: Updating the HARPER template

Brief overview and summary of recommendations



Shirley Wang

Associate Professor of Medicine, Harvard Medical School; Brigham and Women's Hospital

Scan the QR code to join the Q&A on Slido

Panelists



Wei Hua

Director, Division of Epidemiology I, Office of Pharmacovigilance and Epidemiology, Office of Surveillance and Epidemiology, Center for Drug Evaluation and Research, FDA, United States



Wei Liu

Director R&D Data, Data Science & AI, Johnson & Johnson Innovative Medicine



Takashi Waki

Epidemiologist, Pharmaceuticals and Medical Devices Agency (PMDA), Japan



Stephen Motsko

Associate Vice President, Center for Observational Research, Amgen

Moderator



Laura Pizzi

Chief Science Officer, International Society for Pharmacoeconomics and Outcomes Research (ISPOR)

Session 2: Co-development of HARMONIZED review assessment template

Brief overview and summary of recommendations



Shirley Wang
Associate Professor of Medicine, Harvard Medical School; Brigham and Women's Hospital

Panelists



Marie Bradley
Director, Division of Real-World Evidence Analytics (Acting), Office of Innovation and Clinical Trial Modernization, Office of Medical Policy, Center for Drug Evaluation and Research, FDA, United States



Patrice Verpillat
Head of Real-World Evidence, European Medicines Agency (EMA)



Stine Hasling Mogensen
Scientific Senior Advisor, Danish Medicines Agency's Department for Data & Data Analysis



Páll Jónsson
Program Director for Data and Evidence, National Institute for Health and Care Excellence (NICE), United Kingdom



Steve Farmer
Senior Partner, ABIG Health; Former Chief Strategy Officer for Coverage, Centers for Medicare and Medicaid Services (CMS), United States

Moderator



Katsiaryna Bykov
Assistant Professor of Medicine, Harvard Medical School; Brigham and Women's Hospital

Scan the QR code to join the Q&A on Slido

TWG2

From many guidances to a HARMONIZE review approach

What we heard from stakeholders and a proposed path forward



TWG1 makes the science explicit.
TWG2 makes the review explicit.
TWG3 makes both easier to do consistently.

Modified nominal group technique



Kick-off
June

TWG meetings with
22 global stakeholders
July - September

Round-table
October

The central review task is to determine whether the question, data, and design/analysis will generate robust causal evidence to support the intended decision.

Review processes vary across organizations and often involve multiple disciplines and multiple rounds.

Different reviewers may focus on different issues and communicate feedback in different formats, creating fragmented or confusing requests to sponsors.

HARMONIZE should provide a structured way for reviewers to communicate what matters, why it matters, and what important uncertainty remains.

- ✓ Improve the transparency of scientific reasoning.
- ✗ Avoid reducing to a single summary score or checklist exercise.

Connect existing guidance to practical review process

Don’t reinvent the wheel, build from existing workflows

What people are using

ICH M14 + regulator guidance / HTA guidance
Expected study-design and procedural considerations

Tools

HARPER / HARPER+
What a well-specified protocol contains

APPRAISE / ROBINS-I
Bias and validity considerations

ENCePP / TARGET/ STROBE/ SPIFD
Reporting, terminology, data relevance + reliability

HARMONIZE
Holistic Assessment of Real-world evidence for Multistakeholder Objectives and Needs Informing Zonal Evaluations

1. Flag missing/unclear protocol content
2. Map structured assessment items across key domains
3. Link to key references
4. Build a library of examples

Make reviewer reasoning visible for every domain

Every review covers the full framework:

1. **Question/context**
2. **Data fitness**
3. **Design and analysis**
4. **Structured feedback**

Make reviewer reasoning visible for every domain

Every review covers the full framework:

1. **Question/context**
2. **Data fitness**
3. **Design and analysis**
4. **Structured feedback**

Consideration — **Protocol-specific analysis** — **Does it matter?** Why / why not? — **Clarification / change requested** — **Guidance / reference**

Make reviewer reasoning visible for every domain

Every review covers the full framework:

1. **Question/context**
2. **Data fitness**
3. **Design and analysis**
4. **Structured feedback**

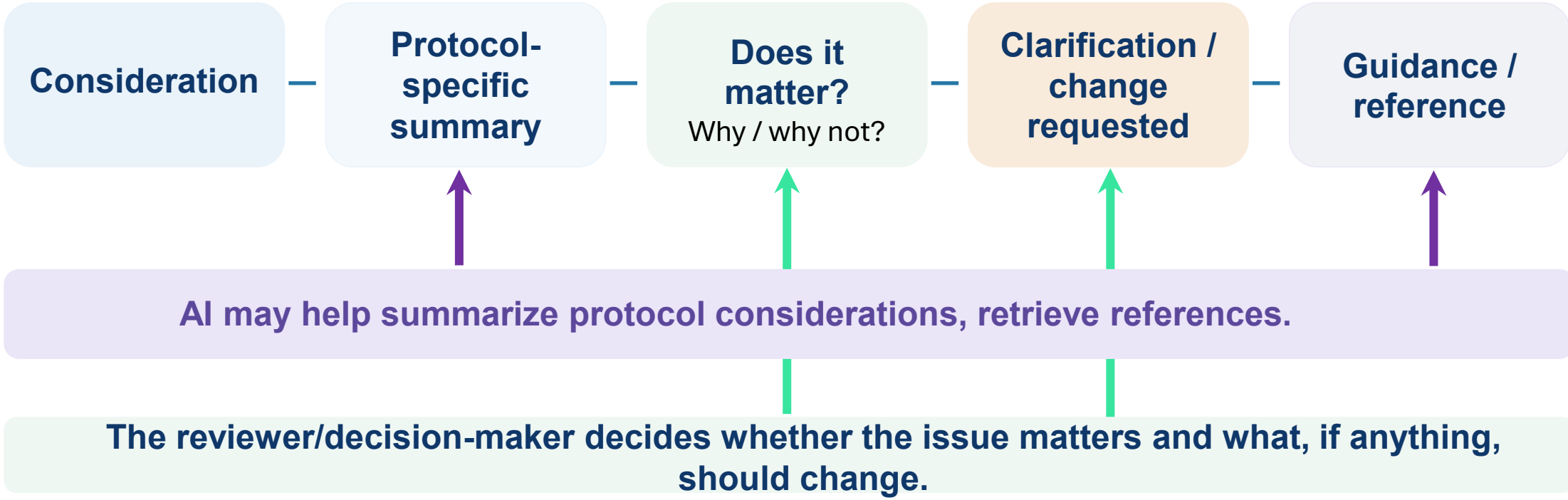
Consideration — **Protocol-specific analysis** — **Does it matter?** Why / why not? — **Clarification / change requested** — **Guidance / reference**

AI may help summarize protocol considerations, retrieve references.

Make reviewer reasoning visible for every domain

Every review covers the full framework:

1. **Question/context**
2. **Data fitness**
3. **Design and analysis**
4. **Structured feedback**



Basic outline

Consideration	No concern	Protocol specific summary analysis	Does it matter? *(remaining uncertainty)*	Clarification/ change requested	Reference
Context of use	■				
…	■				
Question	■				
…	■				
Data fitness	■				
…	■				
Design	■				
…	■				
Analysis	■				
…					

Basic outline

Consideration	No concern	Protocol specific summary analysis	Does it matter? *(remaining uncertainty)*	Clarification/ change requested	Reference
Is surveillance differential between the compared groups?	X				
Is the outcome algorithm valid for this study?		MI/stroke algorithms cite ICD-9 validation in different country's data; mapped ICD-10 codes were not revalidated	Potentially. Applicability of validation performance metrics is uncertain.	Clarify validation evidence for mapped ICD-10 algorithm; discuss implications	APPRAISE 9C; ICH M14

Repeat across every element in each domain.
Question/context • Data fitness • Design • Analysis

Success looks like clearer communications, improved consistency, and fewer iterations while maintaining high standards

Fewer
avoidable review cycles

Lower
reviewer + submitter burden

Clearer
actionable feedback

Proposed TWG2 next steps

1. Draft core question/context of use/data/design structure
2. Map items to HARPER, ICH M14, APPRAISE and other key sources
3. Prototype AI-assisted review that explicitly addresses every domain and consideration
4. Pilot on real / historical protocols with multiple reviewers
5. Measure burden, time, issues identified; refine
6. Build library of examples

Today:

→ identify what’s missing

→ prioritize

Session 2: Co-development of HARMONIZED review assessment template

Brief overview and summary of recommendations



Shirley Wang
Associate Professor of Medicine, Harvard Medical School; Brigham and Women's Hospital

Panelists



Marie Bradley
Director, Division of Real-World Evidence Analytics (Acting), Office of Innovation and Clinical Trial Modernization, Office of Medical Policy, Center for Drug Evaluation and Research, FDA, United States



Patrice Verpillat
Head of Real-World Evidence, European Medicines Agency (EMA)



Stine Hasling Mogensen
Scientific Senior Advisor, Danish Medicines Agency's Department for Data & Data Analysis



Páll Jónsson
Program Director for Data and Evidence, National Institute for Health and Care Excellence (NICE), United Kingdom



Steve Farmer
Senior Partner, ABIG Health; Former Chief Strategy Officer for Coverage, Centers for Medicare and Medicaid Services (CMS), United States

Moderator



Katsiaryna Bykov
Assistant Professor of Medicine, Harvard Medical School; Brigham and Women's Hospital

Scan the QR code to join the Q&A on Slido

LUNCH BREAK

We will reconvene at 12:30 for the next session.

Join the Live Q&A

Scan the QR code to join the Q&A on Slido

OR

1. Go to Slido.com



2. Enter code #2532057

Session 3: Prototype agentic AI tool: Overview and demonstration

Overview and framing



Shirley Wang
Associate Professor, Harvard Medical School; Brigham and Women's Hospital

Agentic AI Prototypes



Jie Yang
Assistant Professor, Harvard Medical School; Brigham and Women's Hospital



Jiageng Wu
Research Fellow, Harvard Medical School, Brigham and Women's Hospital

TWG3

Using AI to support expert protocol development (HARPER) and review (HARMONIZE)

What we are learning from early evaluation and where we aim to go next



• AI-enabled support •
Human review and judgment throughout

TWG1 makes the science explicit.
TWG2 makes the review explicit.
TWG3 makes both easier to do consistently.

TWG3 connects the core structure developed in TWG1 and TWG2 to a practical, AI-enabled expert workflow.

TWG1 | HARPER	TWG2 | HARMONIZE	TWG3 | AI-enabled workflow
Structure for consistently better protocols + clear communication	Structure for consistent review + clear communication	Help experts use both more efficiently

AI may be most useful when it supports specific expert tasks within a structured workflow, rather than trying to replace the expert.

Draft	Check	Retrieve	Summarize
support content + structure	completeness + consistency	guidance + examples	protocol-specific information

Human review of AI-supported work at every step

How does this improve upon ChatGPT/Gemini/Claude?

- ✓ Workflow expedited with built in graphical user interface (GUI)
- ✓ GUI leverages expert engineered prompts based on years of experience to guide the user through each section
- ✓ Section by section design with agents optimized to support specific protocol development/assessment tasks
- ✓ Human-in-the-loop at every step
- ✓ Multi-agent review of protocol quality from different perspectives, e.g. clinician, statistician, regulator, HTA
- ✓ Built in checks for major avoidable biases
- ✓ Built in checks for internal consistency
- ✓ Tuning parameters on local LLM can be controlled to minimize hallucinations and generate near reproducible results from the same prompt
- ✓ Compatible with any LLM, can plug in new LLM versions as they are released - choose one or more to assist in workflow

We started with a bounded task: protocol bias assessment

Early testing focused on APPRAISE[1] evaluation of 6 public HARPER protocols

3 experienced human reviewers

Masters / PhD / PharmD / MBBS
20+ years collective pharmacoepidemiology experience

vs

3 LLMs

Claude
Gemini
OpenAI

What were we testing?

Can structured prompts support consistent, protocol-specific assessment of bias considerations?



APPRAISE: A Tool for Appraising Potential for Bias in Real-world Evidence Studies on Medication Effectiveness or Safety

Katsiaryna Bykov, PharmD, ScD[1], Ashley Jaksa, MPH[2], Jennifer L. Lund, PhD[3], Jessica M. Franklin, PhD[4], Cynthia J. Girman, DrPH[5], Madlen Gazarian, MBBS, MSc[6], Hongbo Yuan, MD, MSc, PhD[7], Stephen Duffield, PhD[8], Seamus Kent, MSc, PhD[9], Elisabetta Patorno, MD, DrPH[1]

[1]APPRAISE is a tool that includes ~25 questions focused on evaluating design, measurement, confounding biases

Both humans and LLMs can struggle

Error patterns were different (and informative)

3 Humans

- Bring contextual expertise (specific to reviewer)
- May differ in judgment
- Expert reasoning may remain implicit
- Can make transcription / coding errors

3 LLMs

- Can bring in external knowledge (reliable sources?)
- Can apply common structured prompts but may differ in judgement (some overly literal or overthink)
- Can make supporting reasoning explicit (sometimes excessively)
- Can hallucinate

Protocol inconsistencies confused both human and LLM reviewers.

Prompt design matters

We used observed errors to add expert-derived nuance and revise prompts.

Initial APPRAISE prompt — **Observe performance** — **Add expert-derived rules** — **Retest**

Source transparency	State what comes from the protocol vs external knowledge.
Use uncertainty	Choose “unclear / unsure” when evidence is insufficient. State what is needed.
Be adversarial	Do not treat what is written in the protocol as automatically true.

More analysis is not always better.

LLMs can identify a possible concern that is not materially important.

Technically possible concern —— **Does it matter here?**

Example: LLM generated true but irrelevant analysis

APPRAISE

8A. Is assessment of exposure/treatment/drug use unreliable and could be missing due to medication use not captured in the data?

Answer: Yes

Analysis: Exposure to oral semaglutide versus sitagliptin is ascertained through prescription claims in Optum Clinformatics, Merative Marketscan, and Medicare. The protocol explicitly states that inpatient medication dispensing and services not covered by insurance are unavailable. Thus, if either study drug is supplied during hospitalization or dispensed outside the captured insurance benefit, that treatment receipt could go unrecorded. Such missing dispensings could obscure prior exposure or create apparent treatment interruptions. This is a potential capture problem, not a concern about adherence after dispensing; the excerpts do not quantify its frequency.

Reasonable people (and LLMs)
may agree on the existence of a potential issue
but disagree on whether it matters.

Governance

Sponsors and reviewers need confidence that sensitive protocols, reviews, and scientific judgments remain within explicitly governed boundaries.

Data + IP control

Secure deployment

Transparency + auditability

Access + organizational control

What to evaluate next

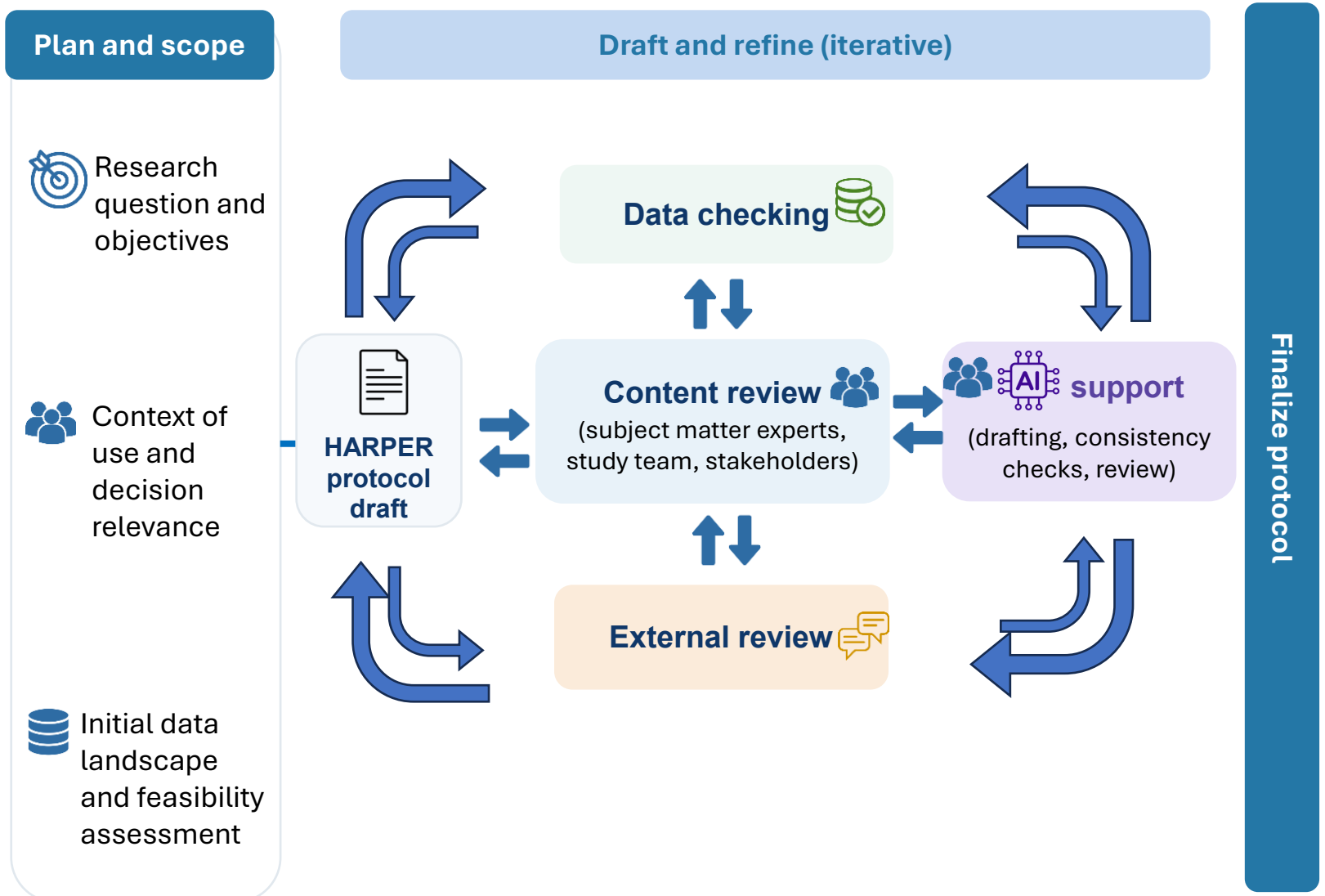
The key comparison is AI-supported experts versus experts working without AI.

- Accuracy
- Completeness
- Relevance
- Consistency
- Efficiency
- Usefulness

Track both output quality and what users actually accept, change, reject, and spend time on.

AI to support consistency/implementation

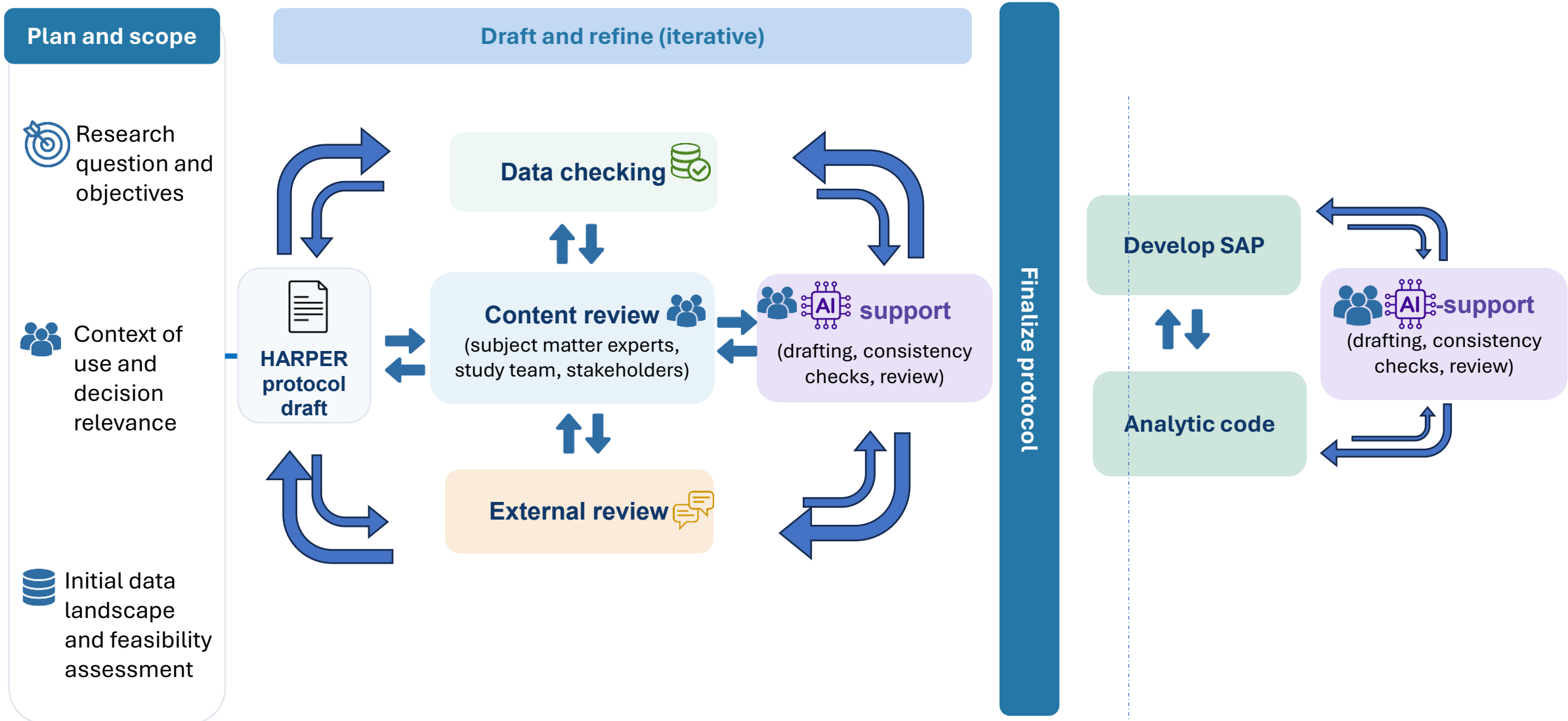
Iterative development and review of HARPER protocol



AI to support consistency/implementation

Iterative development and review of HARPER protocol

Protocol implementation



Today:

→ discuss current design of prototype(s)

→ how are others approaching this?

→ identify what's missing

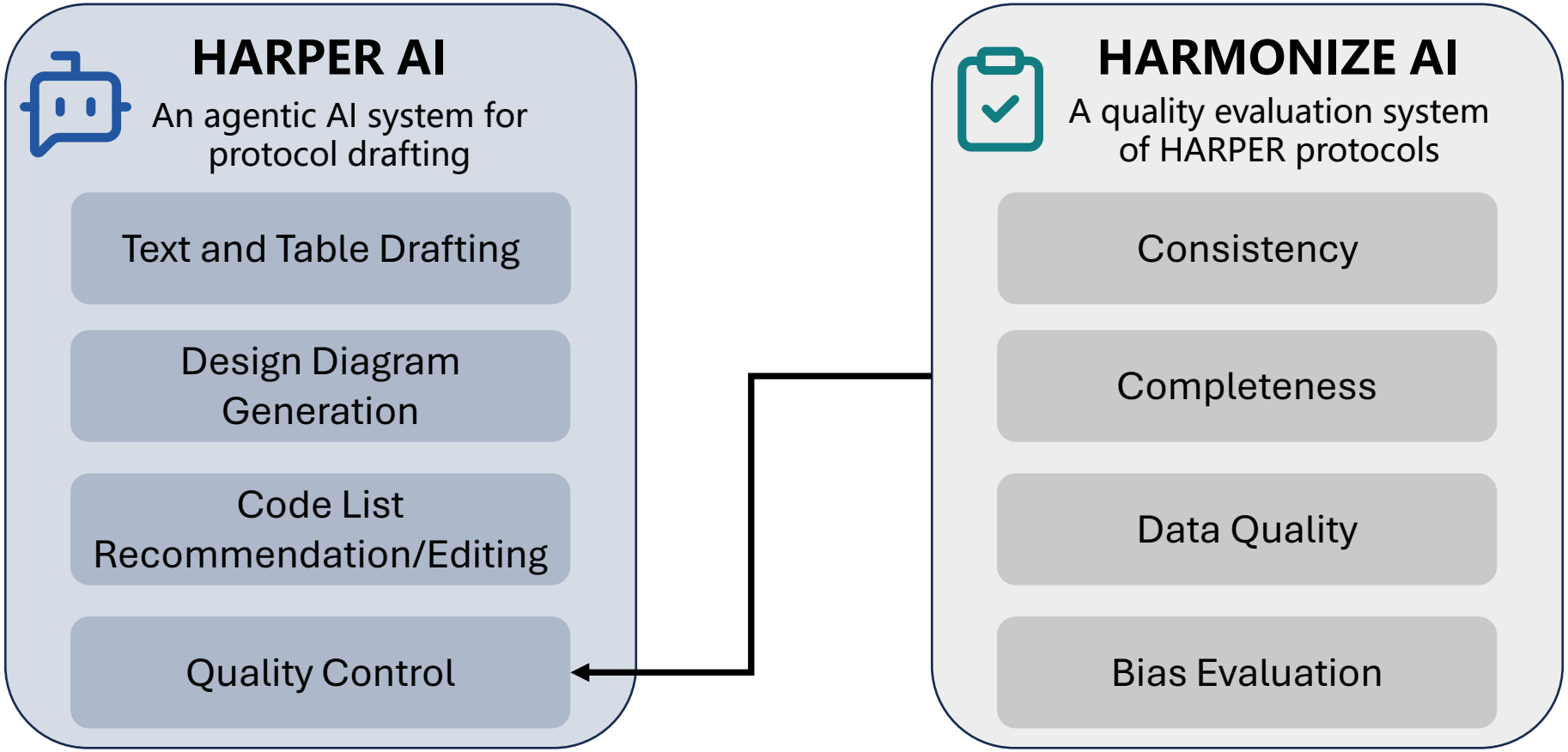
→ prioritize next steps

HARPER AI + HARMONIZE AI Demos

Jie Yang, Assistant Professor
Jiageng Wu, Postdoc Fellow
Division of Pharmacoepidemiology and Pharmacoeconomics
Brigham and Women's Hospital, Harvard Medical School
2026/10/29

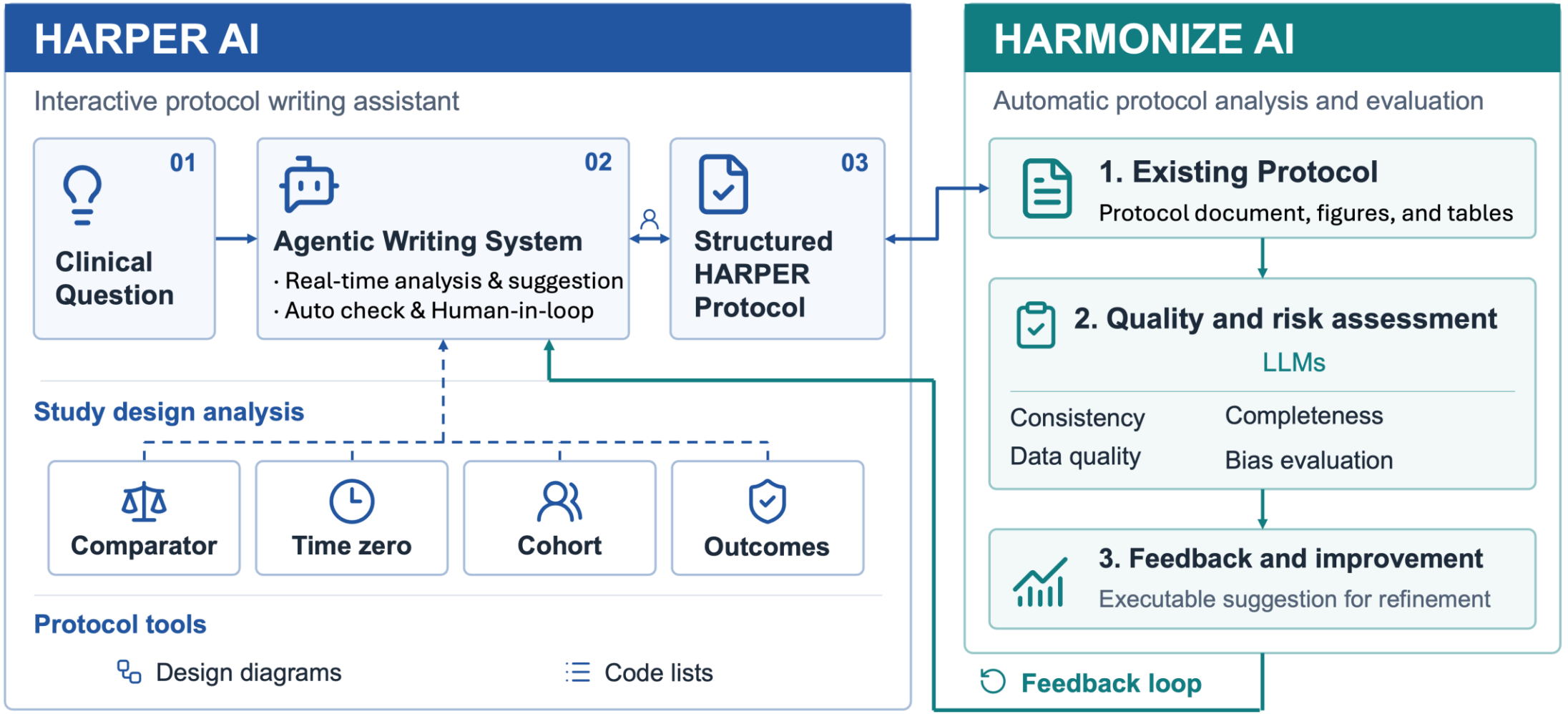
Introduction

- **HARPER AI and HARMONIZE: AI-Powered Protocol Development and Quality Check for RWE**



Overview

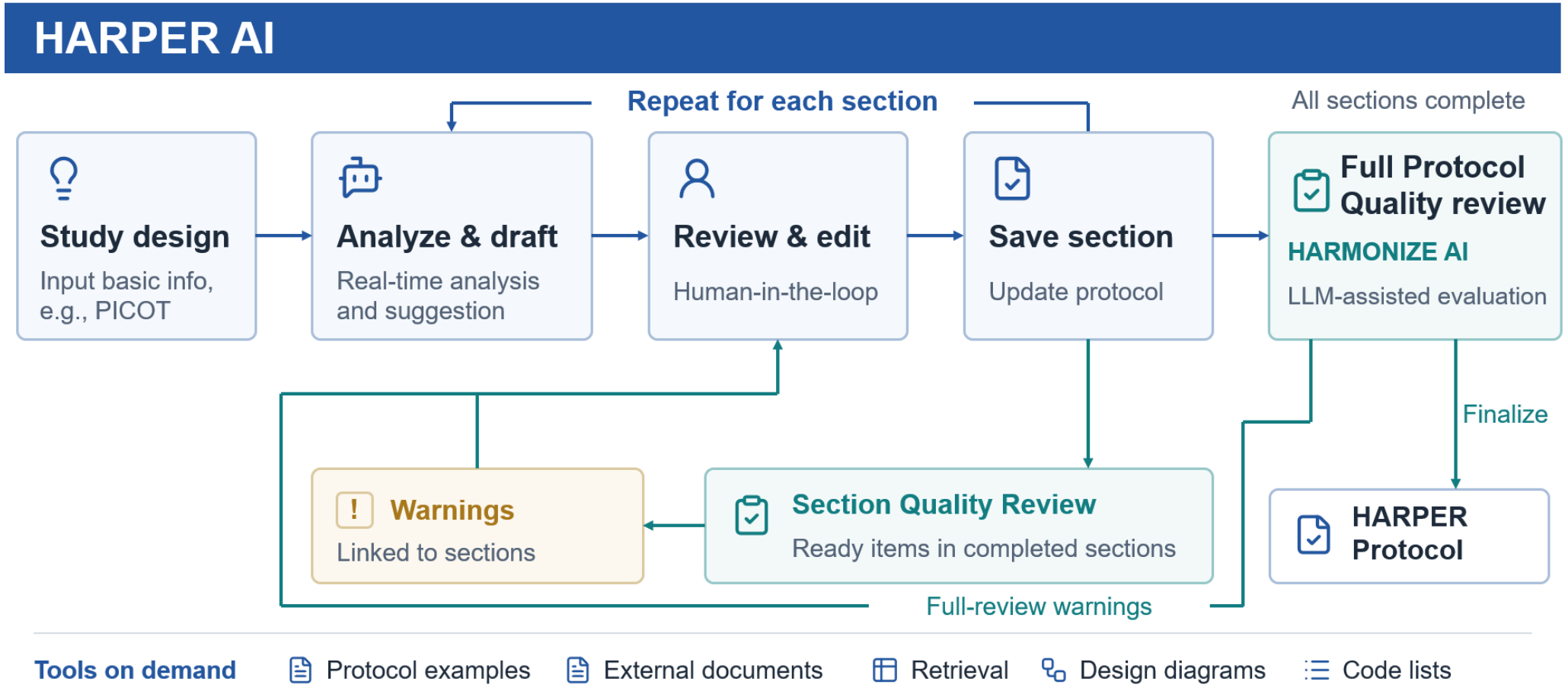
1. HARPER AI: Agentic AI for Protocol Drafting
2. HARMONIZE AI: Efficient Protocol Quality Assessment



Part I – HARPER AI: Framework

➢ An Integrated Workflow for AI-Assisted Protocol Drafting, Human Review, and Quality Assurance

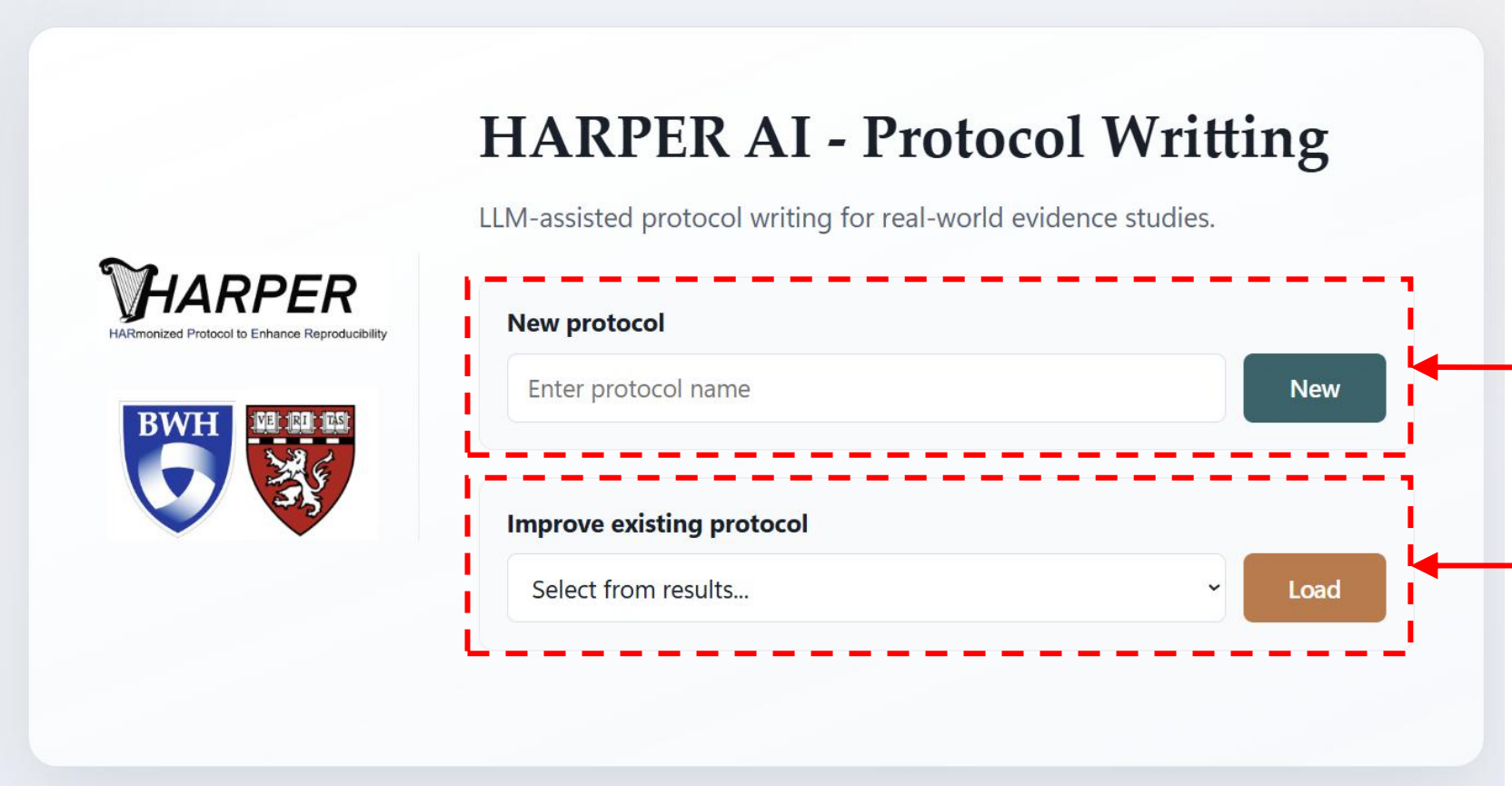
AI accelerates protocol development while keeping researchers in control of critical decisions



Part I – HARPER AI: Demo

➢ Home Page

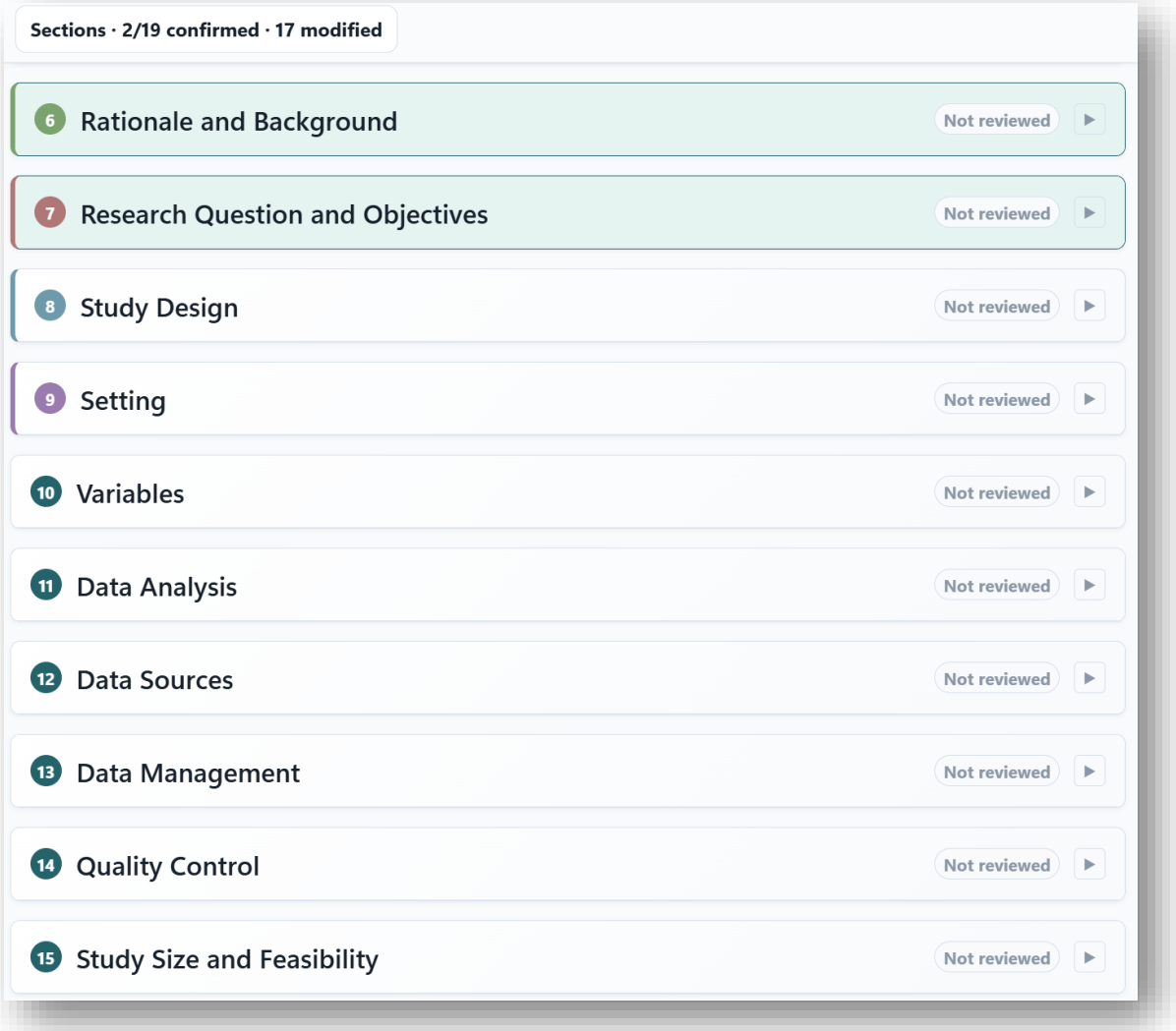
✦ **Supports both drafting new protocols and continually improving existing protocols**



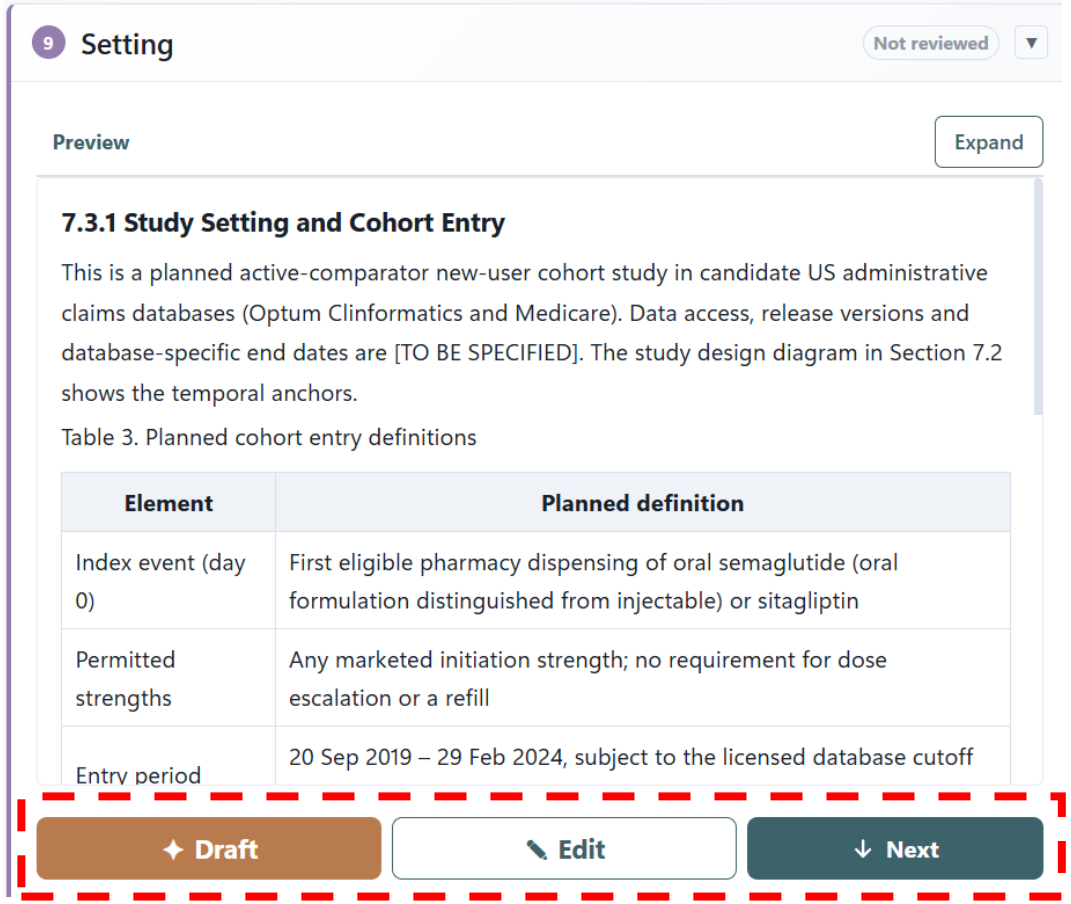
Part I – HARPER AI: Demo

➢ Main Page

- **Human controlled section drafting**



- For each section
  - ✓ **Draft**: draft this section with LLMs
  - ✓ **Edit**: review and edit the current draft
  - ✓ **Next**: confirm and move to next section

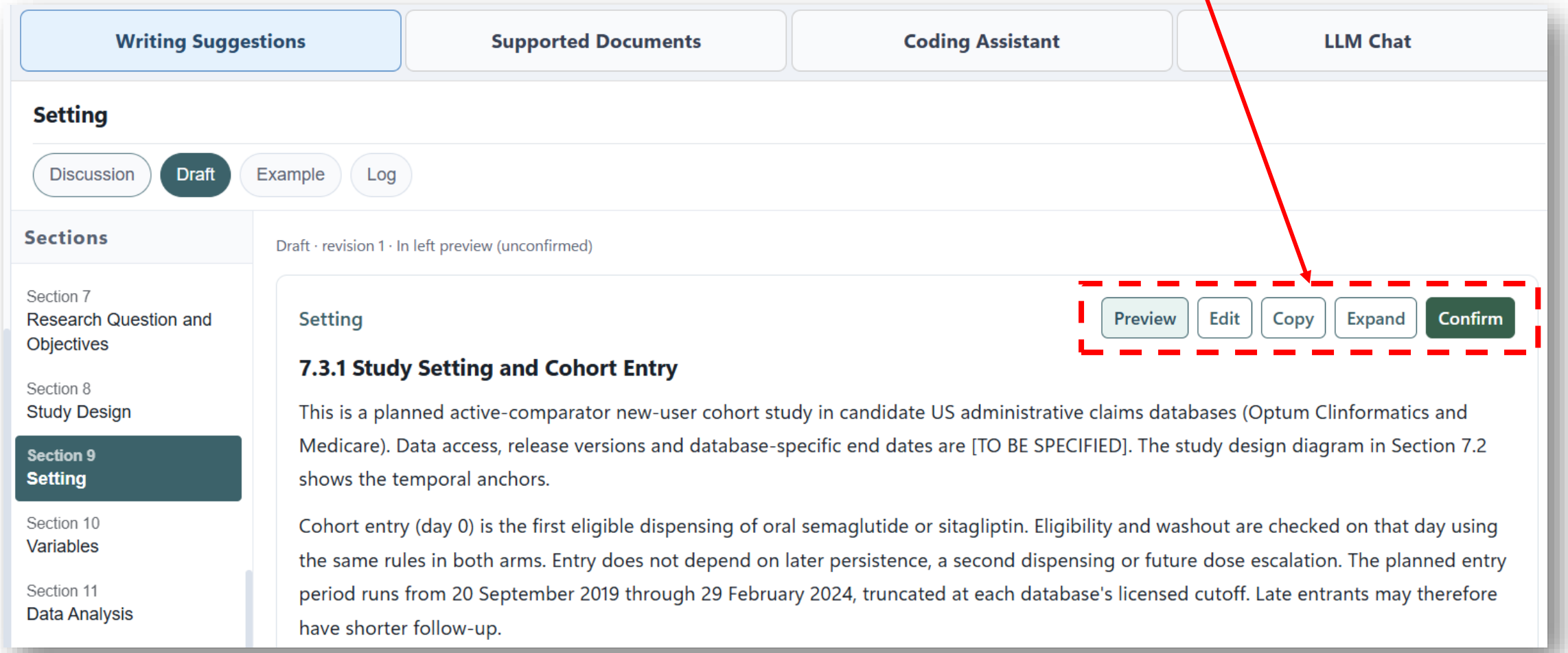


Part I – HARPER AI: Demo

➢ Protocol draft

✦ Human can review and edit LLM-generated suggestions at every step

LLM suggestion -> human review and edit



Part I – HARPER AI: Demo

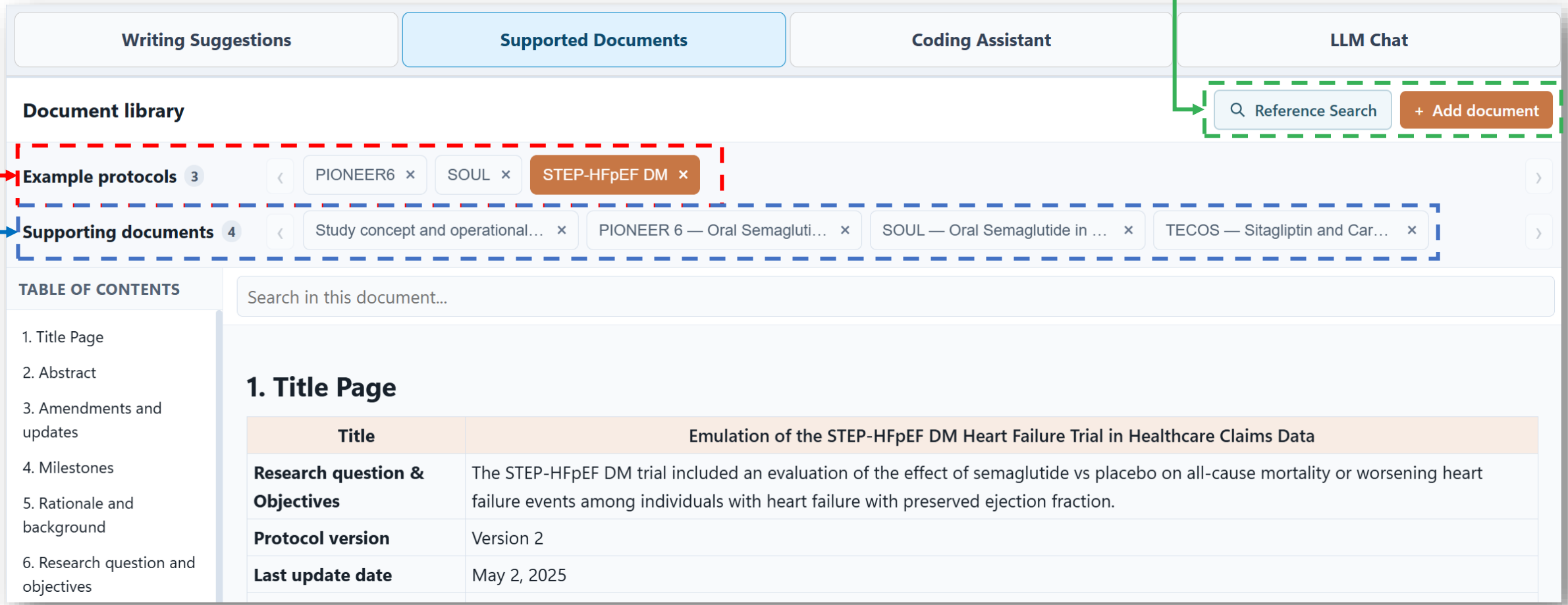
➢ **Supported documents:**

✦ **Guide AI-generated content with existing evidence.**

1. Example protocols: Automatically retrieve relevant protocols

2. Supporting document: manually added materials

3. Search / Add more

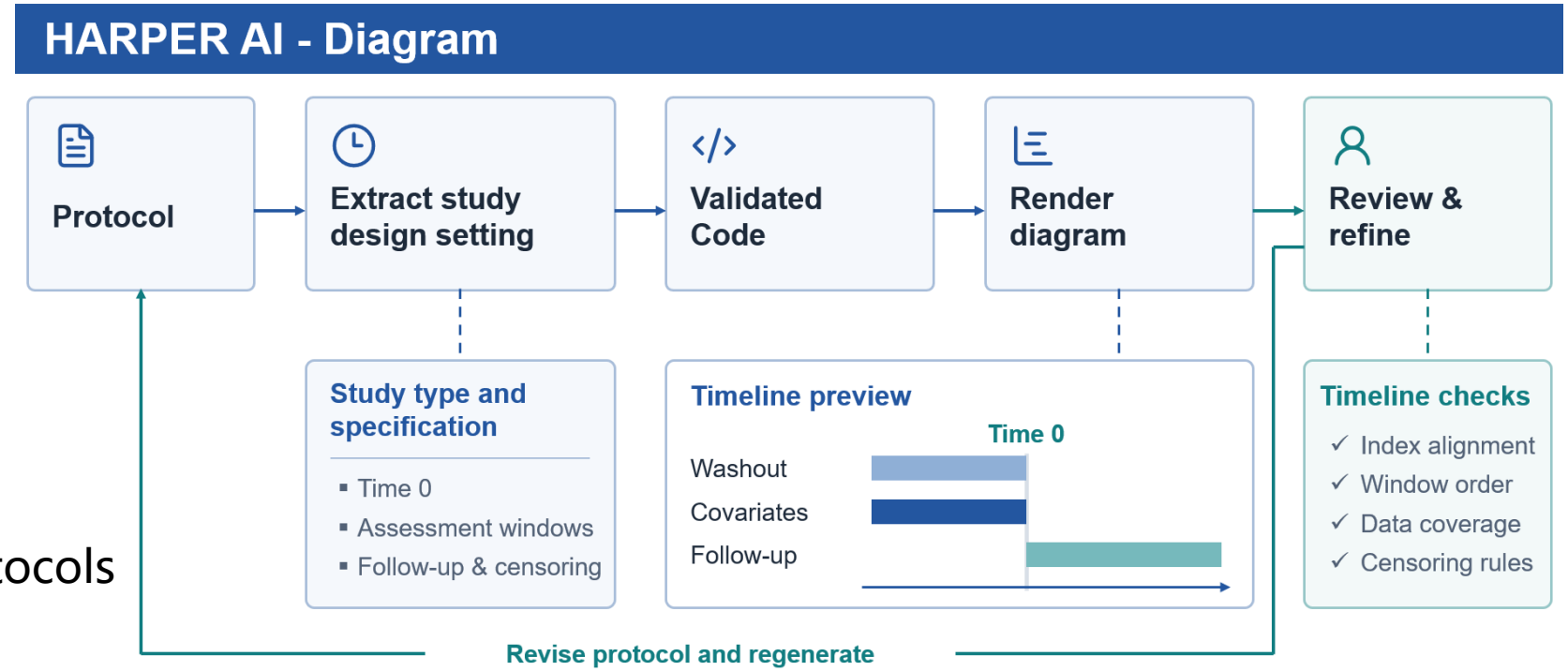


Part I – HARPER AI: Study Design Diagram

- **Transforming Protocol Text into Traceable and Editable Study Design Diagrams**

> Protocol design -> key design elements-> validated visualization code -> traceable design diagrams

- **Advantages**:
  - traceable
  - controllable
  - easy to edit
  - less hallucination
  - reproducible across protocols

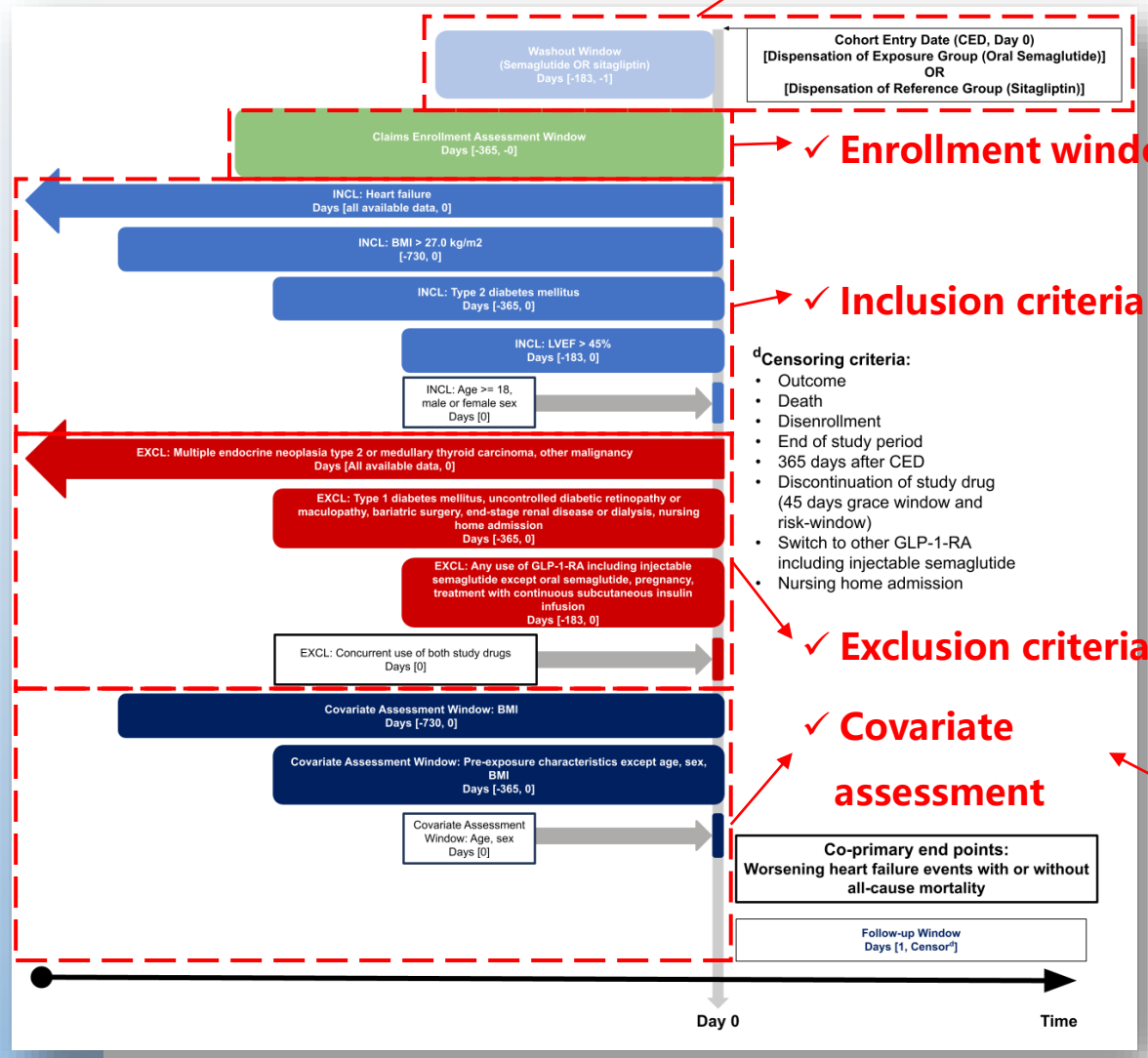


- **Output**: time zero, inclusion/exclusion windows, covariates, follow-up, outcomes

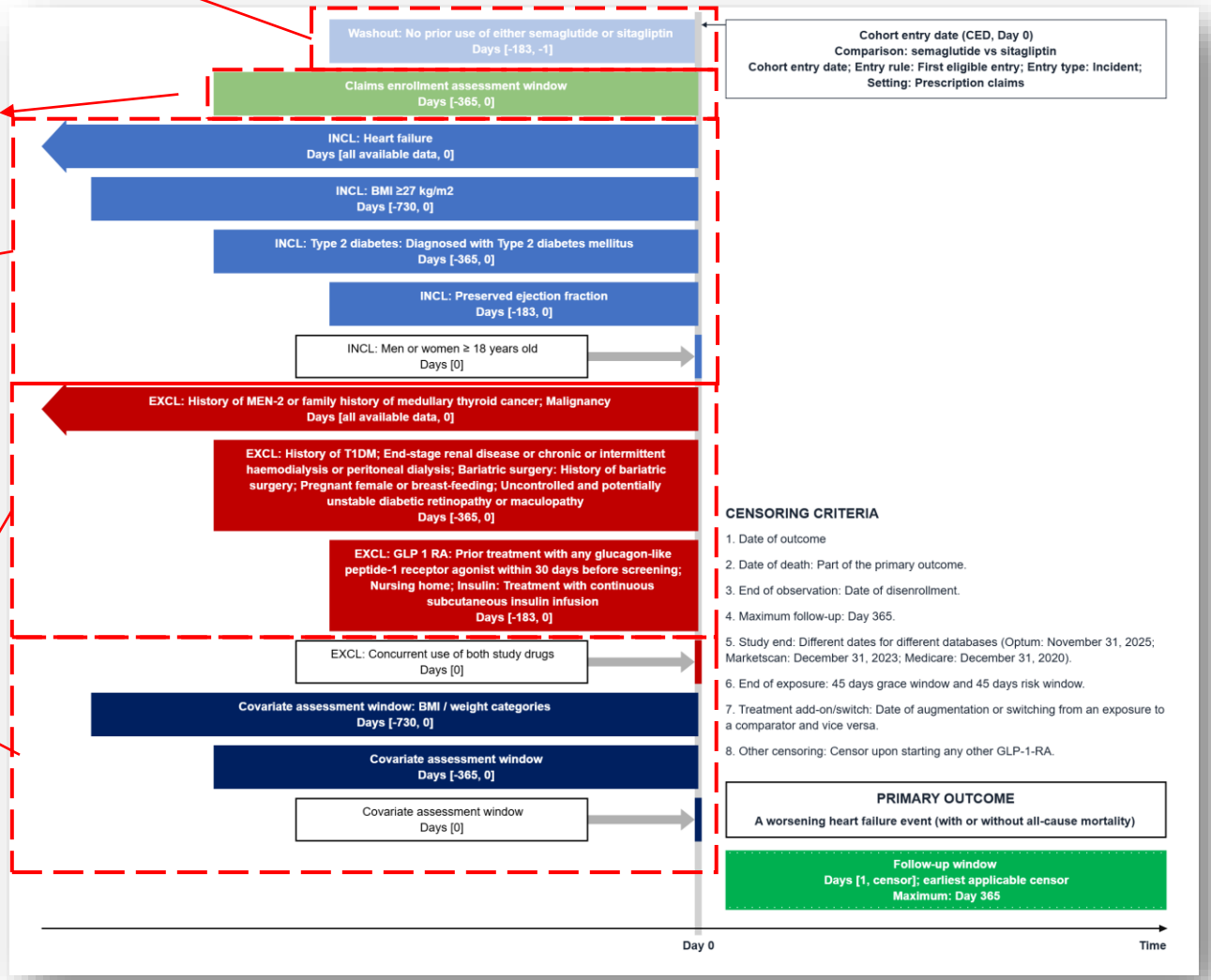
Part I – HARPER AI: Study Design Diagram Demo

- LLM reads the full protocol and extracts structured design elements to organize diagram

Human created



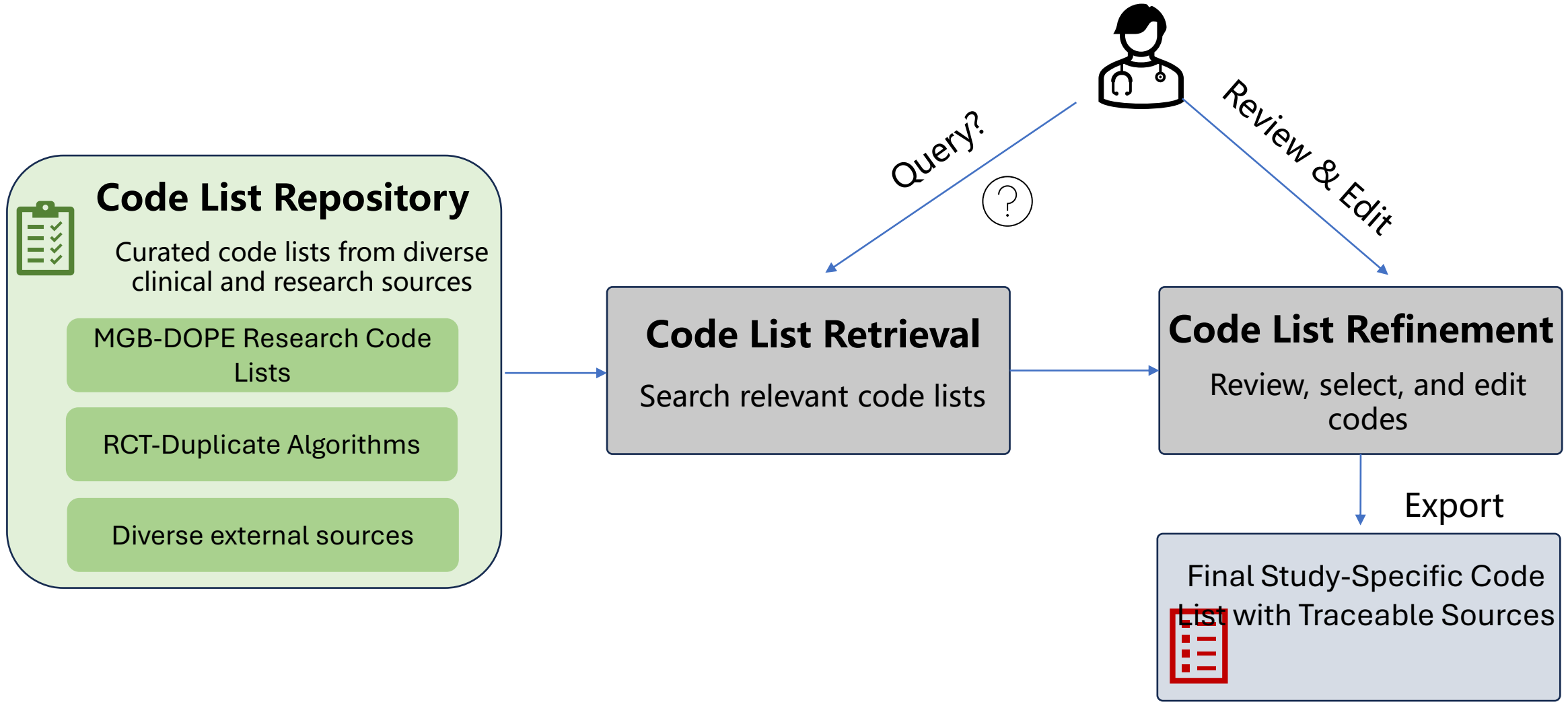
Organized diagram from protocol info



Part I – HARPER AI: Code List Recommendation/Editing

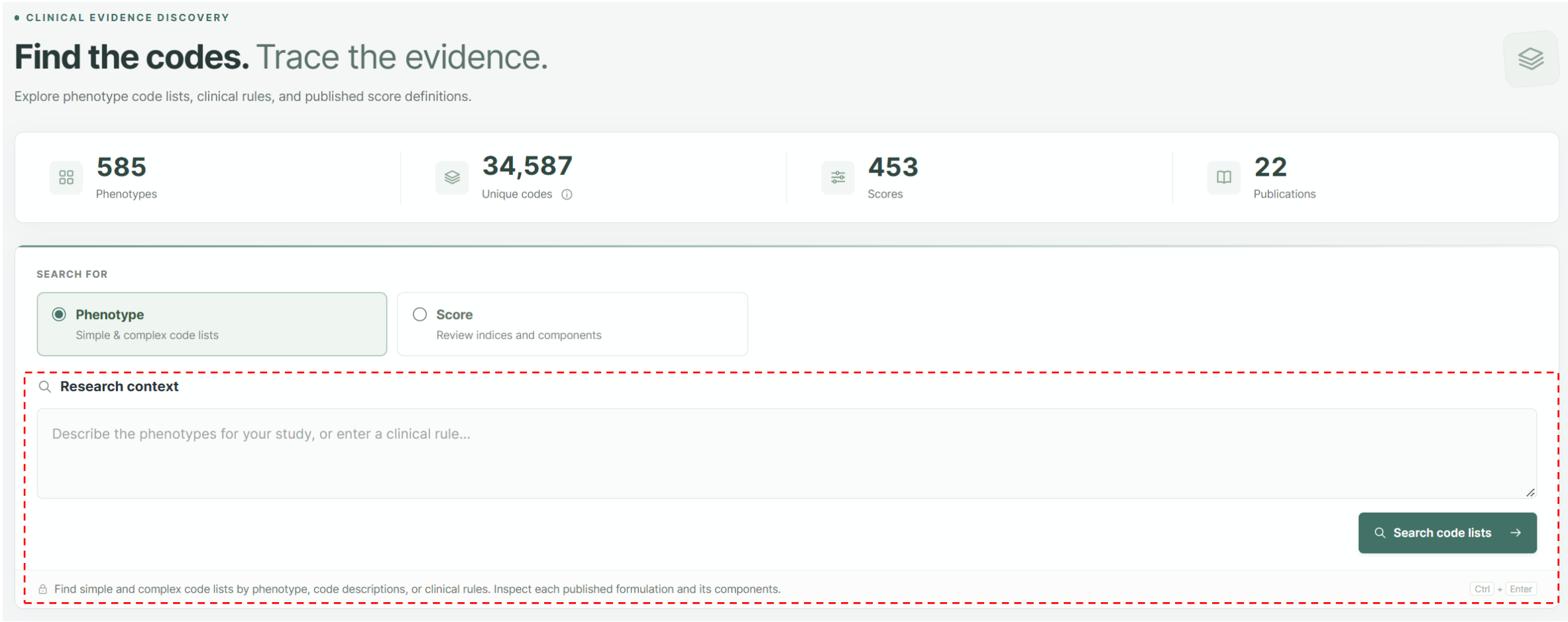
➢ Reusing Existing Research Resources to Accelerate Code List Development

Efficient code list development through querying, selecting, and editing codes from a large curated code pool.



Part I – HARPER AI: Code List Demo

Natural-language search for phenotype code lists and clinical scores in a curated code list repository.

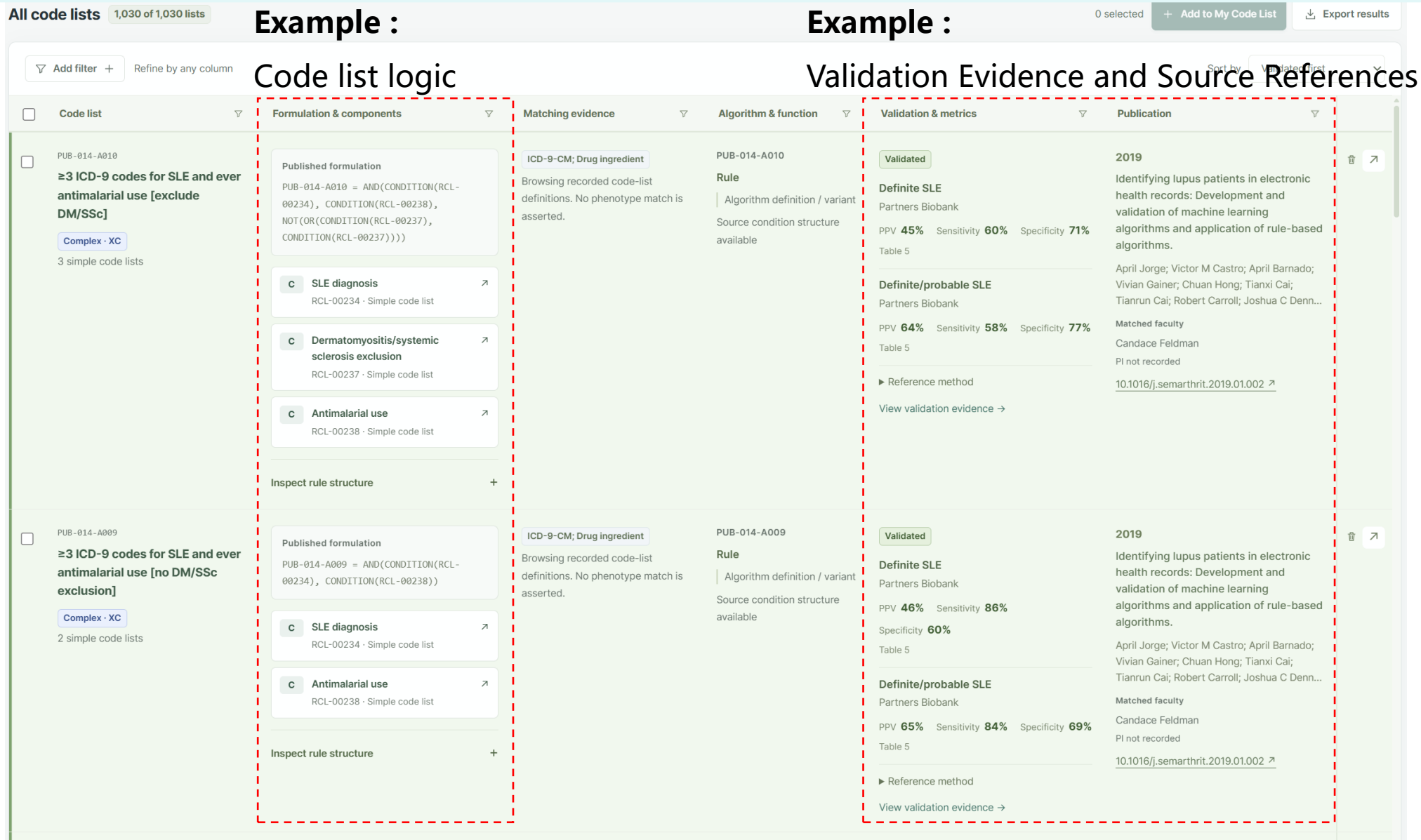


Part I – HARPER AI: Code List Demo

The system retrieves relevant code lists with their logic, validation results, and supporting references.

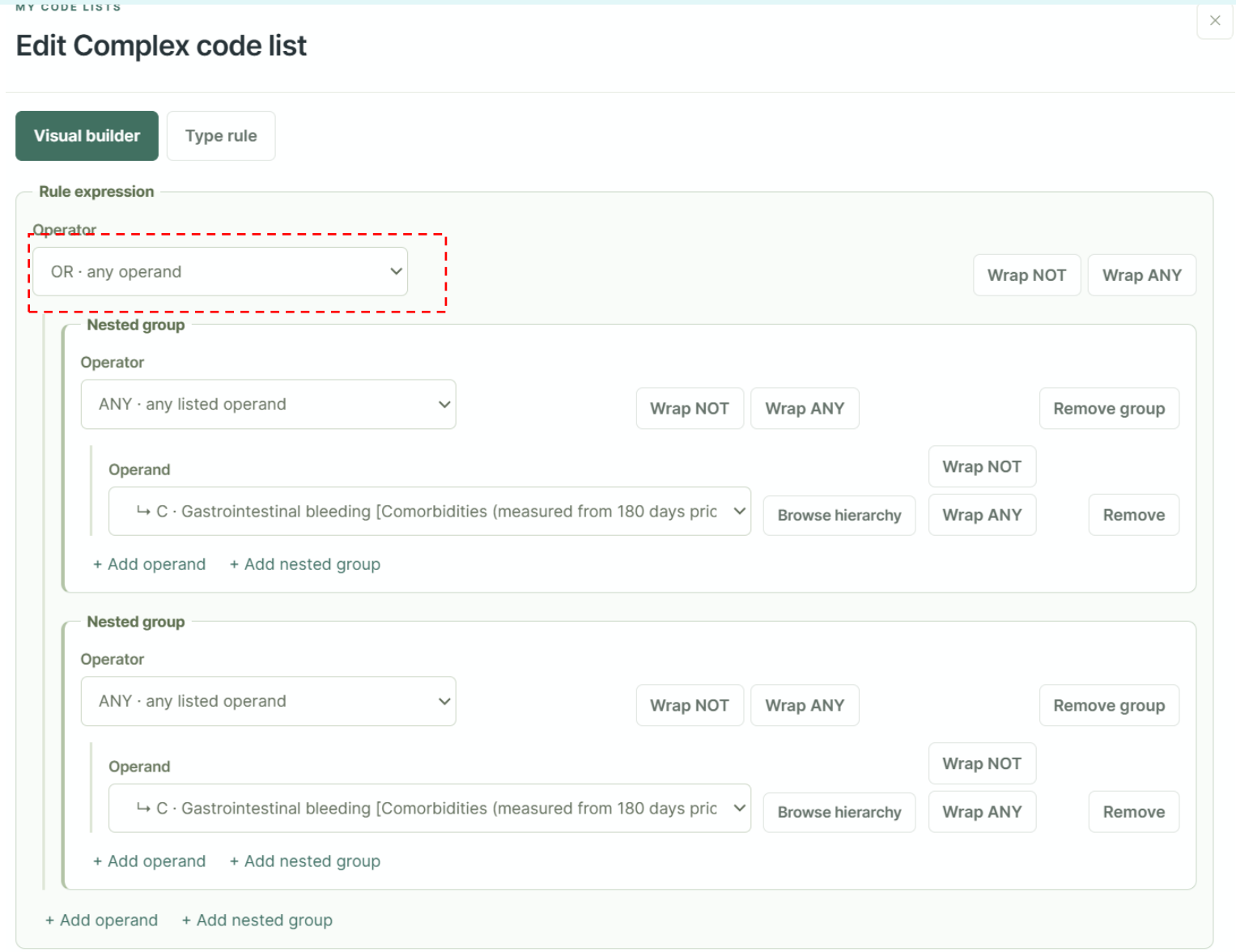
Example : Code list logic

Example : Validation Evidence and Source References



Part I – HARPER AI: Code List Demo

Users can review and edit complex code list logic through an interactive interface.



Part I – HARPER AI: Code List Demo

Users can refine code lists by searching, selecting, and adding individual codes from existing code lists.
Export human-edited code lists with traceable sources.

Codes in this Simple list

Add code

Edit each code, code type, or English interpretation. Source ranges and expressions keep their notation.

Find an entry

Code, type, or English interpretation

Code type	Code / source definition	English interpretation	Actions
ICD-10-CM	Exact code I80	Phlebitis and thrombophlebitis	Remove
ICD-10-CM	Exact code I800	Phlebitis and thombophlb of superficial vessels of low extrm	Remove
ICD-10-CM	Exact code I8000	Phlbts and thombophlb of superfic vessels of unsp low extrm	Remove
ICD-10-CM	Exact code I8001	Phlebitis and thombophlb of superfic vessels of r low extrem	Remove

409 entries · Page 1 of 17 Previous Next

Add codes

Choose codes from a Simple list, or search the Master code base directly by code or description.

Search by: Master code base

Search context: e.g., E11.9 or diabetes

Search

Browsing all 124,566 codes in the Master code base.

Master code base

Code / type	English interpretation
☑ 0016070 ICD-10-PCS Exact code	Bypass Cerebral Ventricle to Nasopharynx with Autologous Tissue Substitute, Open Approach
☑ 0016071 ICD-10-PCS Exact code	Bypass Cerebral Ventricle to Mastoid Sinus with Autologous Tissue Substitute, Open Approach
☑ 0016072 ICD-10-PCS Exact code	Bypass Cerebral Ventricle to Atrium with Autologous Tissue Substitute, Open Approach
☐ 0016073 ICD-10-PCS Exact code	Bypass Cerebral Ventricle to Blood Vessel with Autologous Tissue Substitute, Open Approach
☐ 0016074 ICD-10-PCS	Bypass Cerebral Ventricle to Pleural Cavity with Autologous Tissue Substitute, Open Approach

124,566 entries · Page 1 of 4983 Previous Next

3 codes selected

← Add selected codes

Part I – HARPER AI: Evaluation

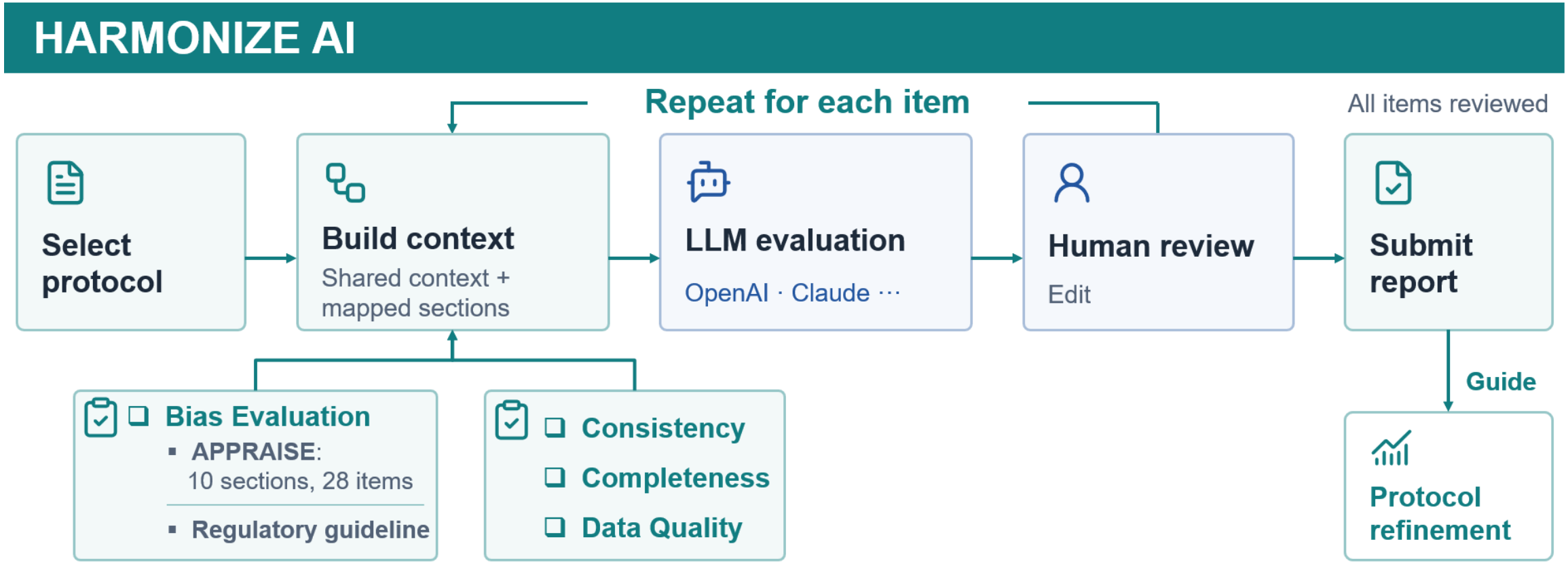
Combining system-level evaluation with expert-validated protocol quality assessment to enable efficient, reliable, and human-centered AI-assisted protocol development.

		Evaluation Focus	Methods / Metrics
1 Usability Evaluation	◔	Efficiency, usability, and human–AI interaction	Task completion time, edit records, revision frequency, usability surveys
2 Protocol Completeness	✓	HARPER compliance and internal consistency	LLM-based assessment of sections, tables, definitions, and code lists
3 Methodological Validity	◆	Study design bias, data suitability, implementation feasibility	APPRAISE, DataSAT, code generation
4 Reviewer Consistency	A / B	Reliability and accuracy of protocol quality assessment (HARMONIZE AI)	Bias injection, multi-LLM comparison, agreement with human experts
5 Stakeholder Acceptance	♙	Satisfaction, perceived usefulness, and protocol acceptability	Stakeholder surveys, Likert scores, expert feedback

Part II – HARMONIZE AI

➢ Integrating Efficient Quality Assessment into HARPER AI

✦ **Robust, multidimensional protocol quality assessment with AI + Human.**



*Bykov, K., Jaksa, A., Lund, J.L., Franklin, J.M., Girman, C.J., Gazarian, M., Yuan, H., Duffield, S., Kent, S. and Patorno, E., 2025. APPRAISE: A tool for appraising potential for bias in real-world evidence studies on medication effectiveness or safety. Value in Health.

Part II – HARMONIZE AI: Settings

➢ Prompt setting

Detailed query for LLMs

- Role: ···· ···
- Background: ···· ···
- Task: ···· ···
- Rules: ···· ···

General system prompt

- APPRAISE Item
- Relevant Protocol Sections
- Important Guidelines

Detailed Item prompt

- **Example:** the relevant sections of similar protocols (medication/design relevant)
- **References:** added materials, such as key descriptions, relevant papers, etc

Supplementary information



Large Language Model

Model Response

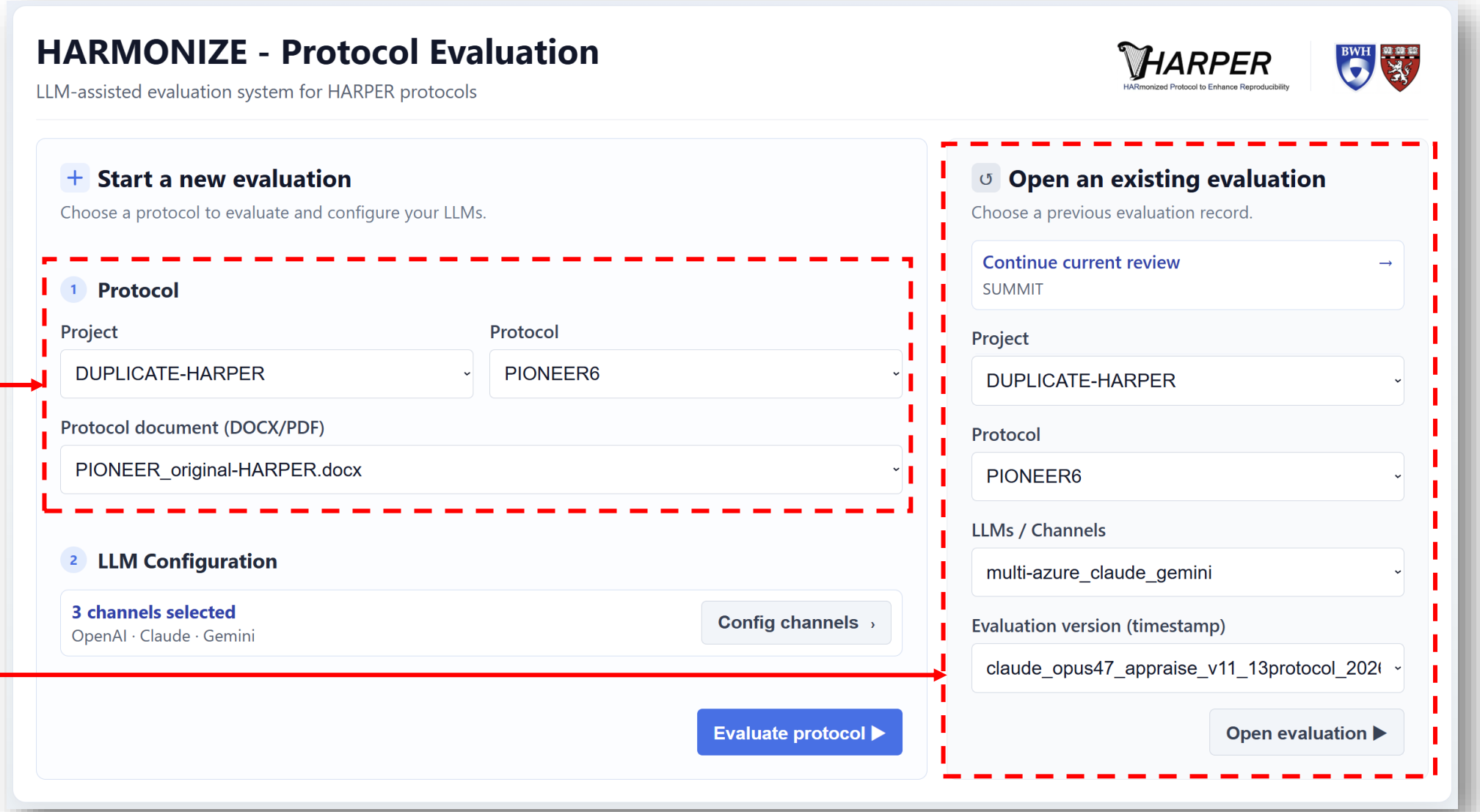
Answer: Yes
Analysis: why the model chose the answer
Evidence: the original text supports the analysis
Confidence: the confidence level of answering this item

Part II – HARMONIZE AI: Demo

➢ Home Page

Select a new protocol for evaluation

Review an existing evaluation

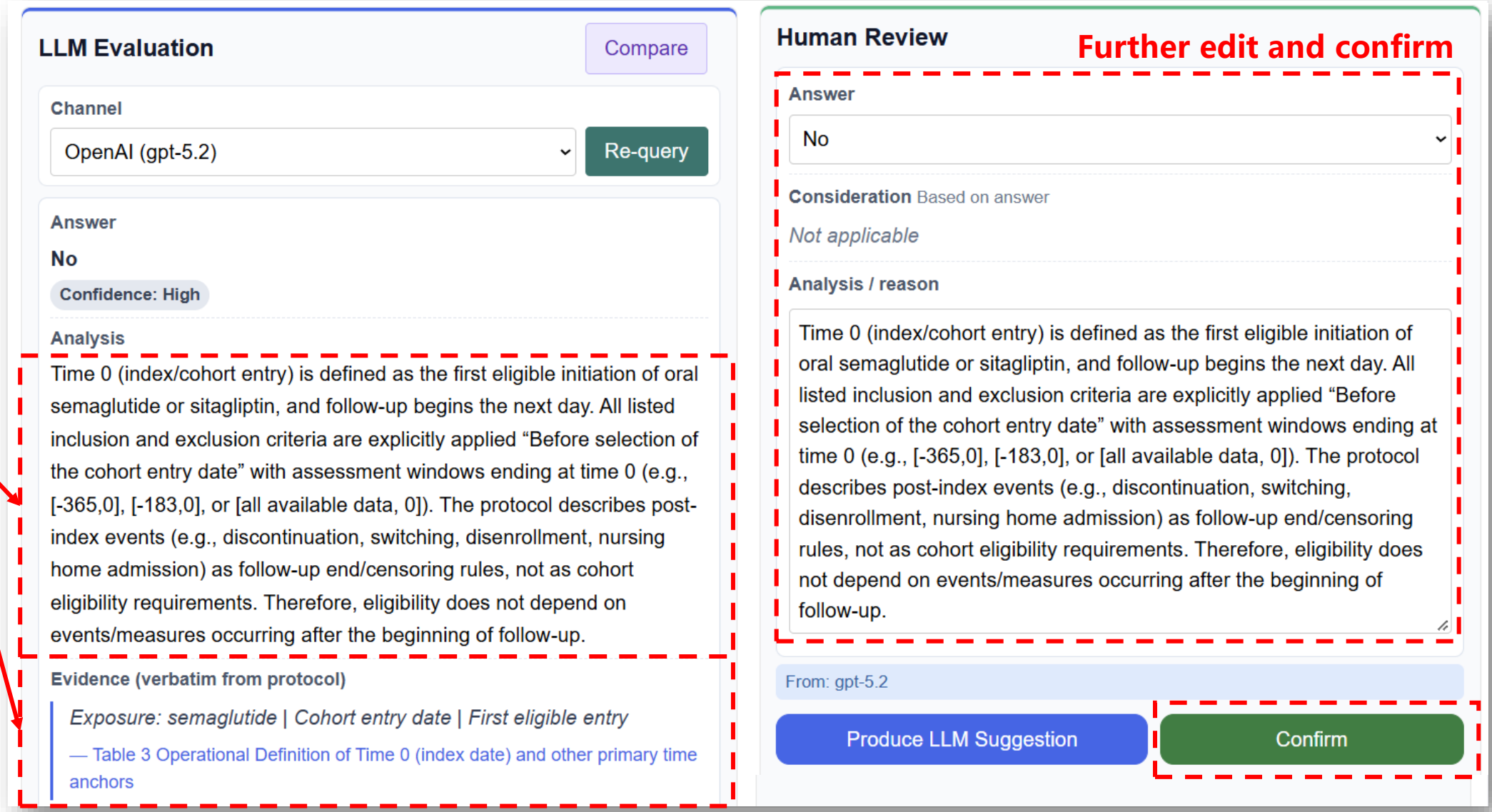


Part II – HARMONIZE AI: Demo

➢ Section Page

LLM:

- Answer
- Confidence
- Analysis
- Evidence



Part II – HARMONIZE AI: Benchmark

➢ Benchmark construction

✦ **Human reviewer agreement of 87.8%–90.4% serves as a benchmark for evaluating AI performance**

✓ **Guideline refinement (APPRAISE to assess protocol bias):**

10 sections, 28 questions

Question Example	9A. Can information on outcome events be missing from the data due to care outside of the data capture system?
Options	Yes | No | Unsure

✓ **Protocol evaluation:**

20 existing HARPER protocols

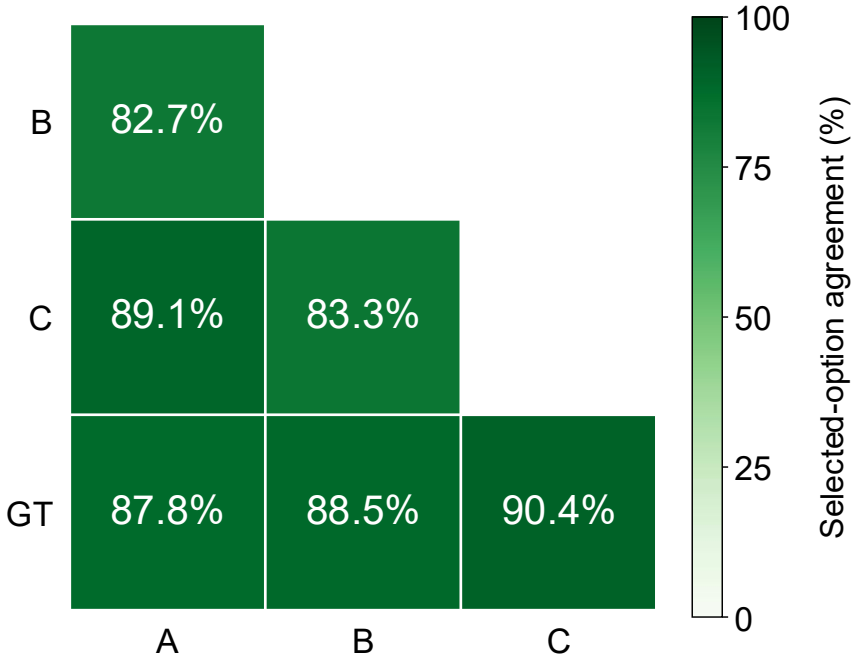
6 protocols completed

✓ **Human validation:**

3 expert reviewers' independent annotation

+ 1 senior expert discussion

} Ground-truth [GT]



(Reviewers: A/B/C and GT[Ground-truth])

Part II – HARMONIZE AI: Evaluation

Prompts with detailed expert guidelines help AI to analyze and review the protocol

➢ Basic prompt

Question description from APPRAISE

Section: 2. Time-related bias (immortal person-time)
Question: 2C. Were individuals in the treatment group or the comparator selected in hierarchical order (e.g., were individuals in the treatment group or the comparator selected first)?

Response options:
- Yes – No - Unclear

Option-specific considerations:
- Yes: Time-related bias due to systematic exclusion ⋯ ⋯
- No: ⋯ ⋯

Examples/guidance (from the APPRAISE definition):
For example, comparing "ever-users" or "initiators" of a drug to "never-users", OR selecting patients initiating the drug of interest from the data first and then selecting comparators ⋯ ⋯

➢ Expert prompt

Question description from APPRAISE

Section: 2. Time-related bias (immortal person-time)
Question: 2C. Were individuals in the treatment group or the comparator selected in hierarchical order (e.g., were individuals in the treatment group or the comparator selected first)?

Response options:
⋯ ⋯

Option-specific considerations:
⋯ ⋯

Examples/guidance (from the APPRAISE framework):
⋯ ⋯

Important Guidelines (Highest Priority)

Step 1. Determine how treatment groups are constructed.
⋯ ⋯

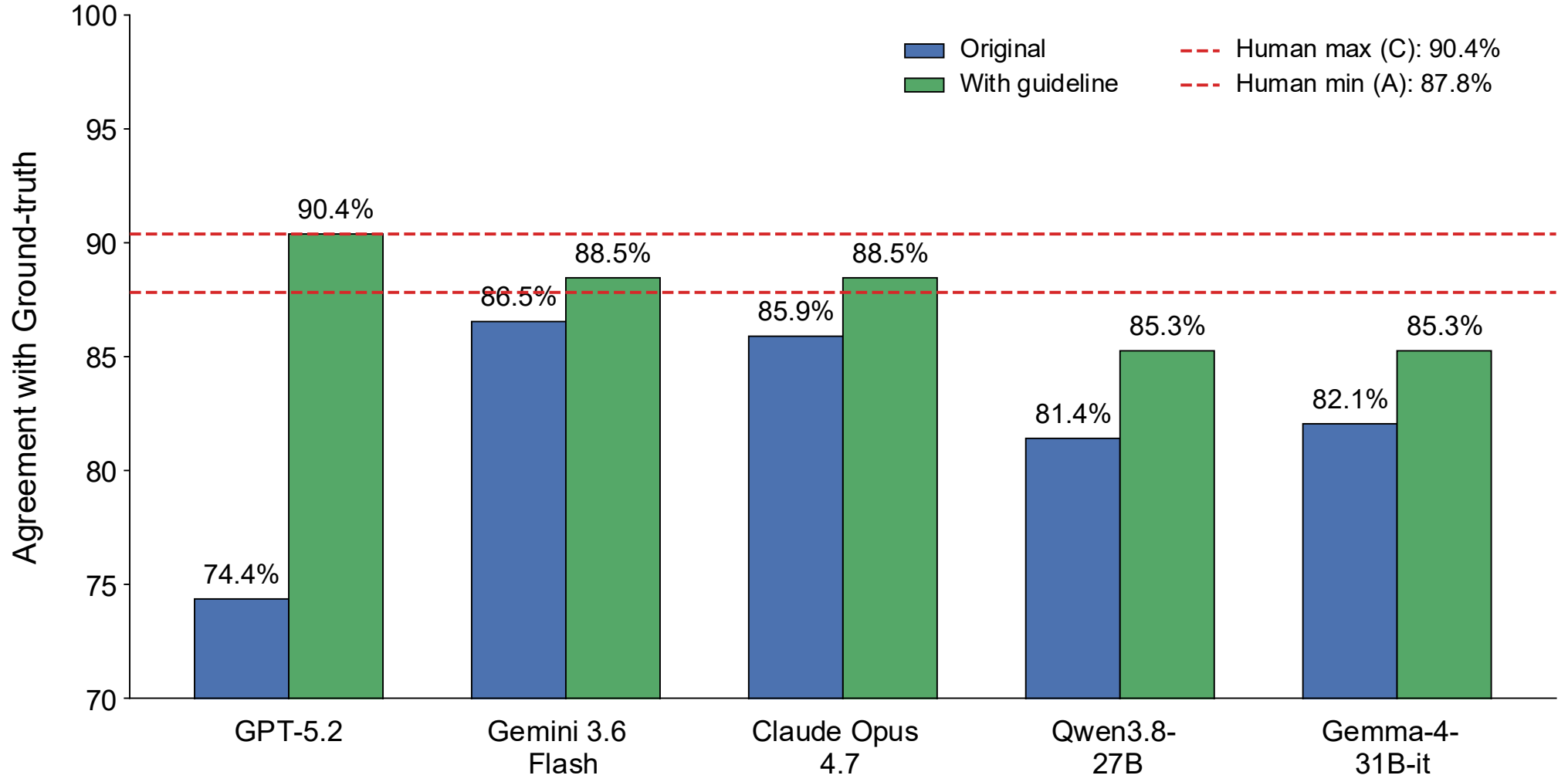
Step 2. Determine whether treatment-group assignment is hierarchical. ⋯ ⋯

Part II – HARMONIZE AI: Evaluation

➢ Performance comparisons (annotator vs ground-truth)

Expert guidelines improved agreement across all five LLMs (up to +16%)

Proprietary models achieved human-level agreement, while open-weight models remained slightly behind

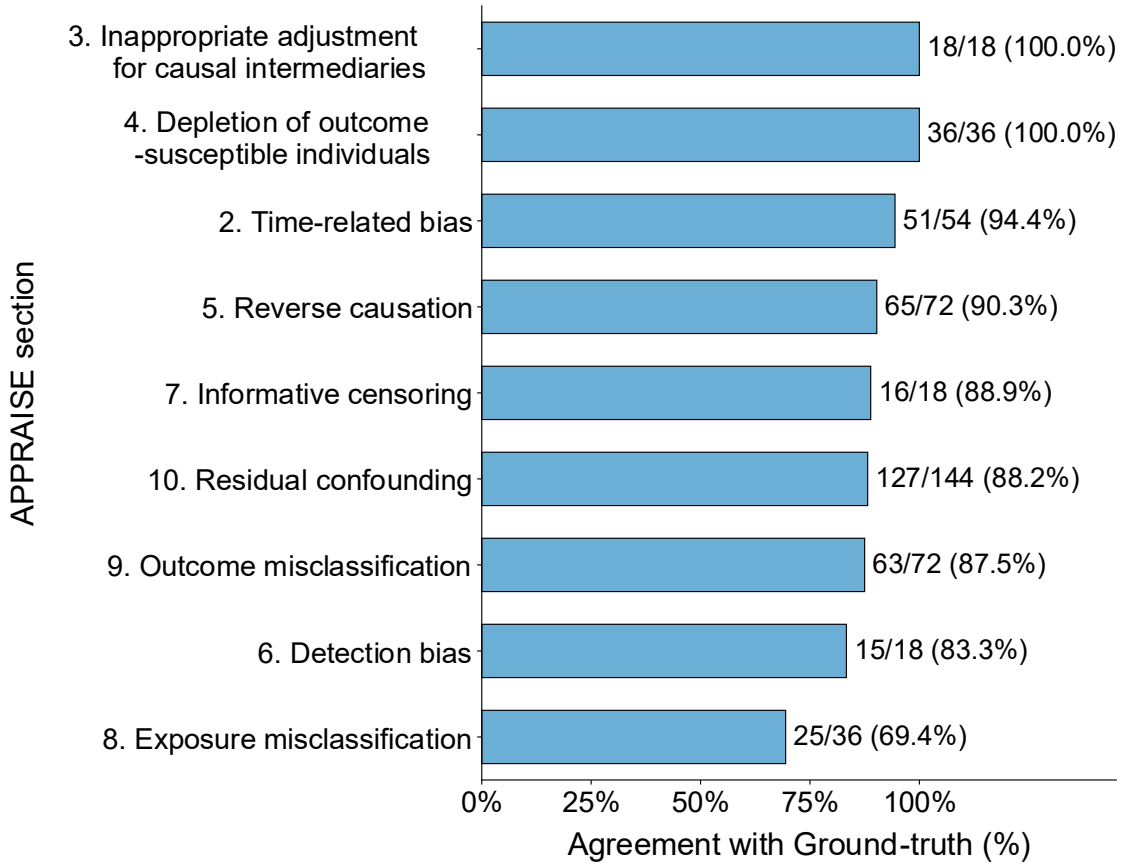


Part II – HARMONIZE AI: Evaluation

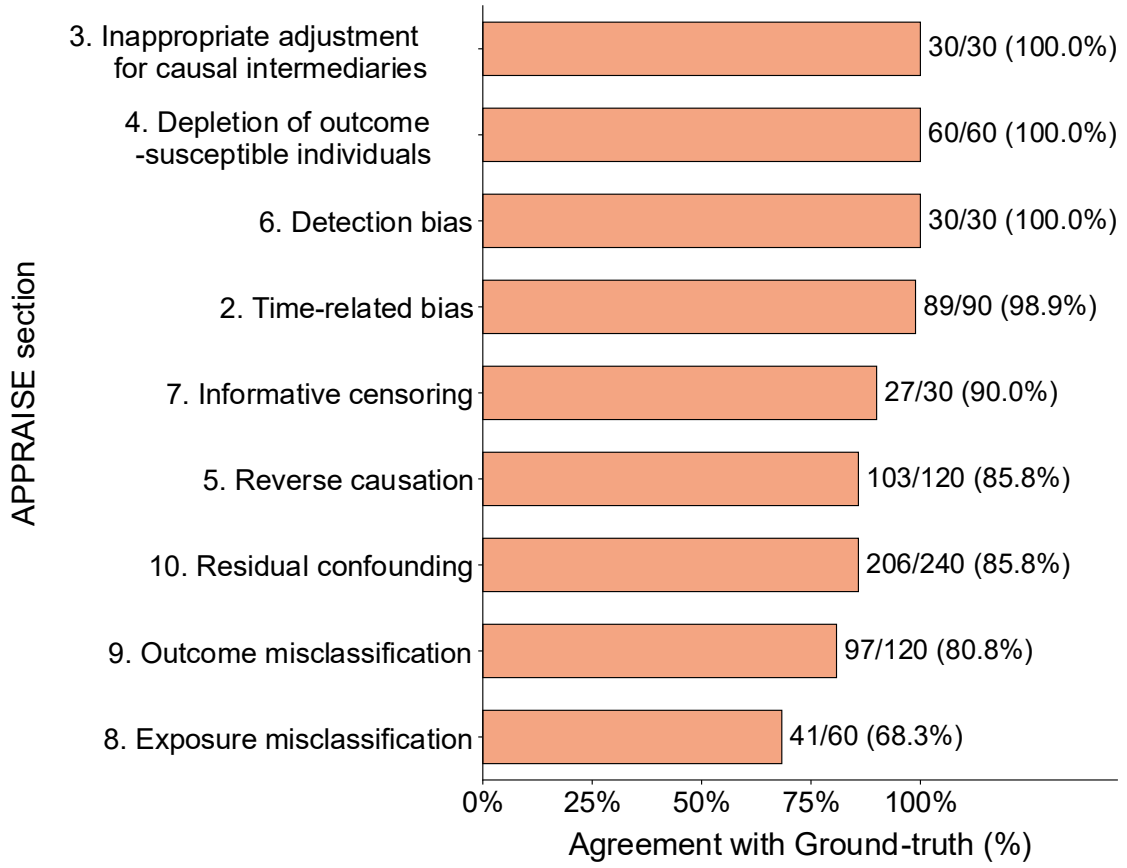
➢ Agreement at APPRAISE section level

✦ **The performance varies across different sections, “exposure misclassification”->highest disagreement rate**

Agreement among human reviewers



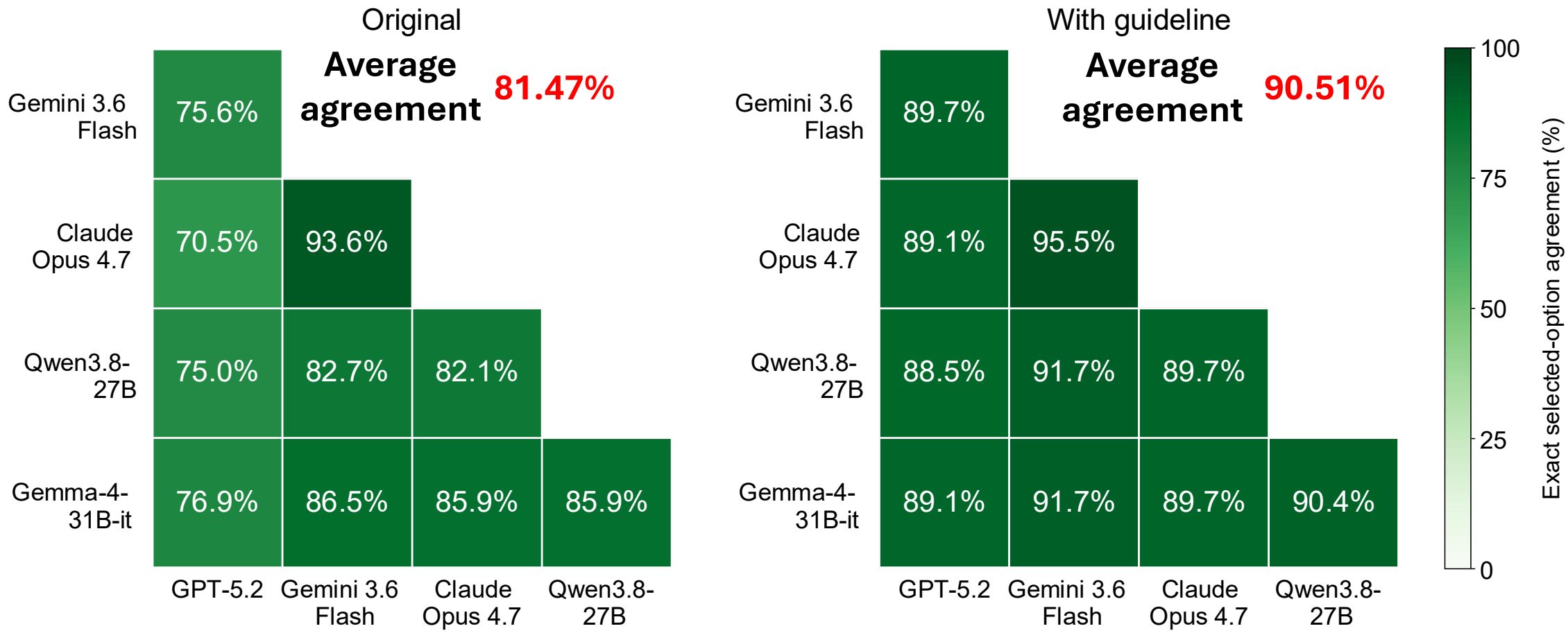
Agreement among LLMs



Part II – HARMONIZE AI: Evaluation

➢ Internal agreement analysis

✦ The expert guideline constraint model thinking and achieve better agreement across models



Conclusion

➢ Summary

1 HARPER AI: Assisted protocol drafting

Human-in-the-loop text drafting, editable design diagrams, and source-traceable code lists.

2 HARMONIZE AI: Systematic quality review

Structured evaluation of completeness, consistency, data quality, and potential bias.

3 Initial evaluation: Encouraging agreement in HARMONIZE AI

Human reviewer agreement: 87.8%–90.4%.

Proprietary LLMs achieved human-level agreement on protocol quality assessment

Session 3: Prototype agentic AI tool: Overview and demonstration

Panelists



Carol Koro
Vice President, Head of Global Epidemiology and Real-World Evidence, GlaxoSmithKline



Amy Mulick
Biostatistician and Real-World Evidence Researcher, Bayer



Daniel Prieto-Alhambra,
Professor of Regulatory Health Data Science, University of Oxford, United Kingdom



Arup Pramanik
Head of Global Real-World Evidence and Health Economics and Outcomes Research Strategy, Immunology, UCB



Fatma M. Shebl
Epidemiologist, Sentinel Core Team, Office of Surveillance and Epidemiology, Center for Drug Evaluation and Research, FDA, United States

Moderator



Mina Tadrous
Associate Professor, Drug Safety Health Services and Policy Research, Leslie Dan Faculty of Pharmacy, University of Toronto

Scan the QR code to join the Q&A on Slido





Thank you!