phuse·share
connect·share·advance

Single Day Event

From Protocol to SDTM: Automating Trial Design with Large Language Models

Haris Khan
Arpitha Hanumanthaiah
PointCross Life Sciences

phuse.global

Disclaimer

Date: [22-11-2025]
Presenter: Arpita Hanumanthaiah, Haris Khan
Organization: PointCross Life sciences

Intellectual Property:

This presentation may include materials and information protected by copyright, trademarks, or other intellectual property rights belonging to PointCross Life Sciences. Use of such material is made under the fair use doctrine for educational, informational, and non-commercial purposes only. Any reproduction, redistribution, or use of the material beyond fair use may require the permission of the copyright holders.

Confidentiality Notice:

This presentation may contain confidential or proprietary information intended only for the use of the individuals or entities to whom it is addressed. If you are not the intended recipient, you are hereby notified that any dissemination, distribution, or copying of this presentation is strictly prohibited.

Questions and Contact Information:
For questions or more information about this presentation, please contact [haris@pointcross.com & arpitha@pointcross.com].

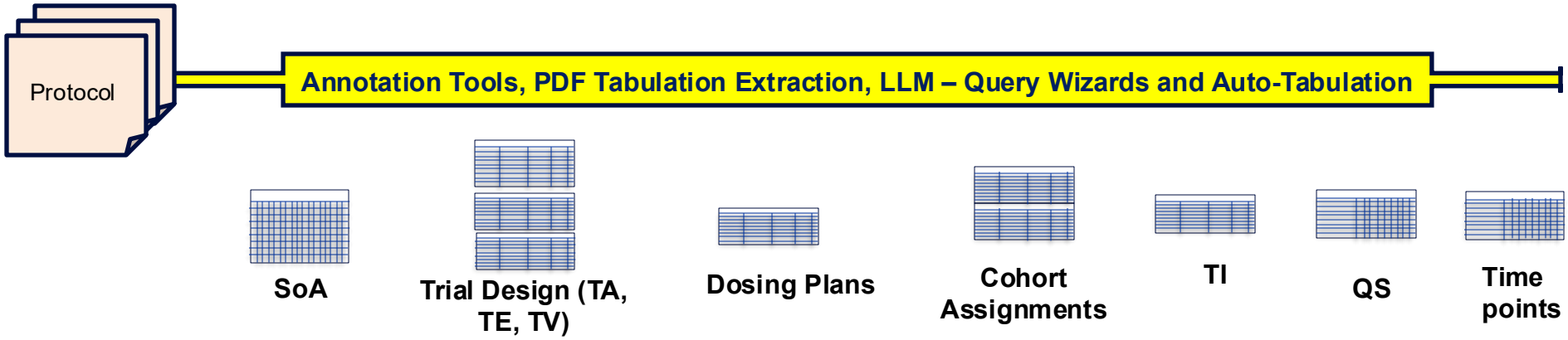
Agenda

- ✓ Protocol Digitization
- ✓ Why & How Large Language Models are used?
- ✓ Experience & Learnings
- ✓ Our Approaches & Methodologies
- ✓ Few Examples of Targeted Prompt Execution
- ✓ Human in Loop for Review & Approvals
- ✓ Automatic generation of SDTM Trial Design Domains
- ✓ Conclusion

Protocol Digitization

The protocol is the central source of study truth, guiding data collection and analysis. Manual extraction into systems is slow and error-prone, and frequent amendments often lead to inconsistencies across platforms.

Digitization converts the protocol from a narrative document into a structured, machine-readable model. This allows automated configuration of downstream systems (e.g., EDC, CDISC outputs), improving both accuracy and consistency.



Key study elements (TS, TE, TA, TV, TI, Dosing, DM, Cohorts, SoA) become machine-readable objects.

Why & How Large Language Models are Used?

Understanding Domain-Specific Language

Interprets clinical terminology and varied protocol phrasing to consistently extract accurate study concepts.

Handling Variability

Clinical documents vary vastly in styles and LLMs adapt to these variations.

Few-shot Learning

LLMs can be guided with just a few examples to extract specific information, making them much faster to deploy.

Chunking & Context Management

Breaks long protocols into logical sections while preserving relationships across design, visits, and endpoints.

Extracting Unstructured/Tabular Data

Converts narrative text and complex tables into clean, structured data formats like JSON or CSV.

Summarizing Complex Study Design Logic

Simplifies dose escalation, randomization, and visit rules into clear summaries or rule-based representations.

Building Traceability

Links extracted data back to exact protocol locations for transparency, validation, and audit readiness.

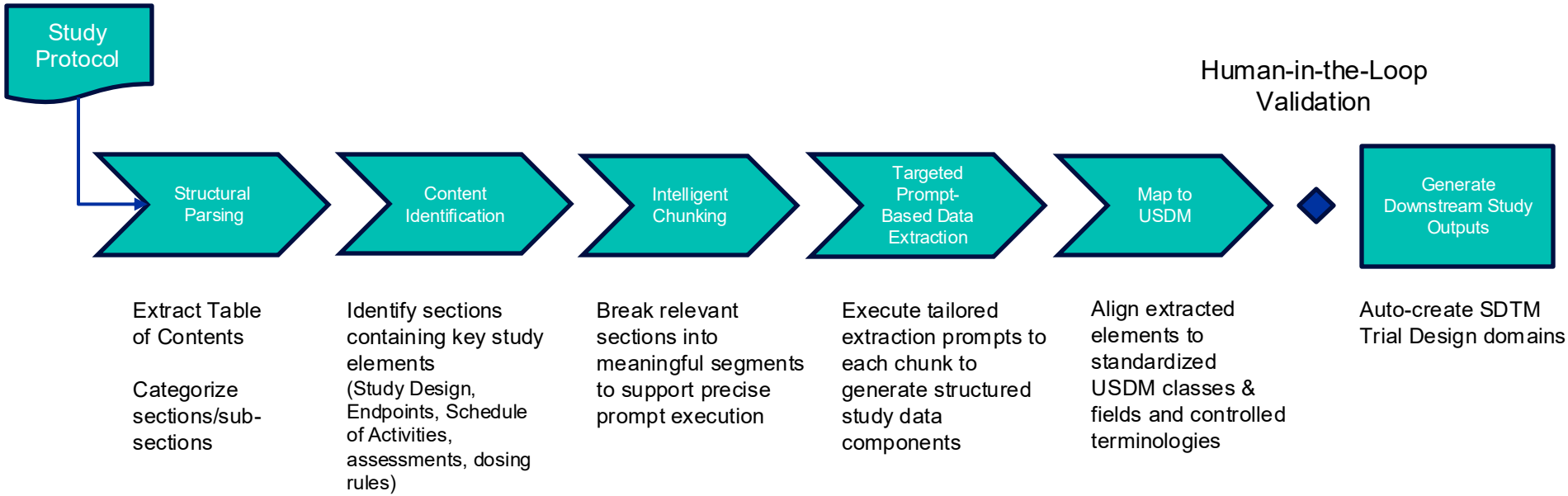
Experience & Learnings

Using public or proprietary AI LLM providers with clinical study information raises significant regulatory risks. Concerns center on PHI/PII security, Intellectual Property exposure, and the inherent lack of GxP/regulatory compliance within these AI models.

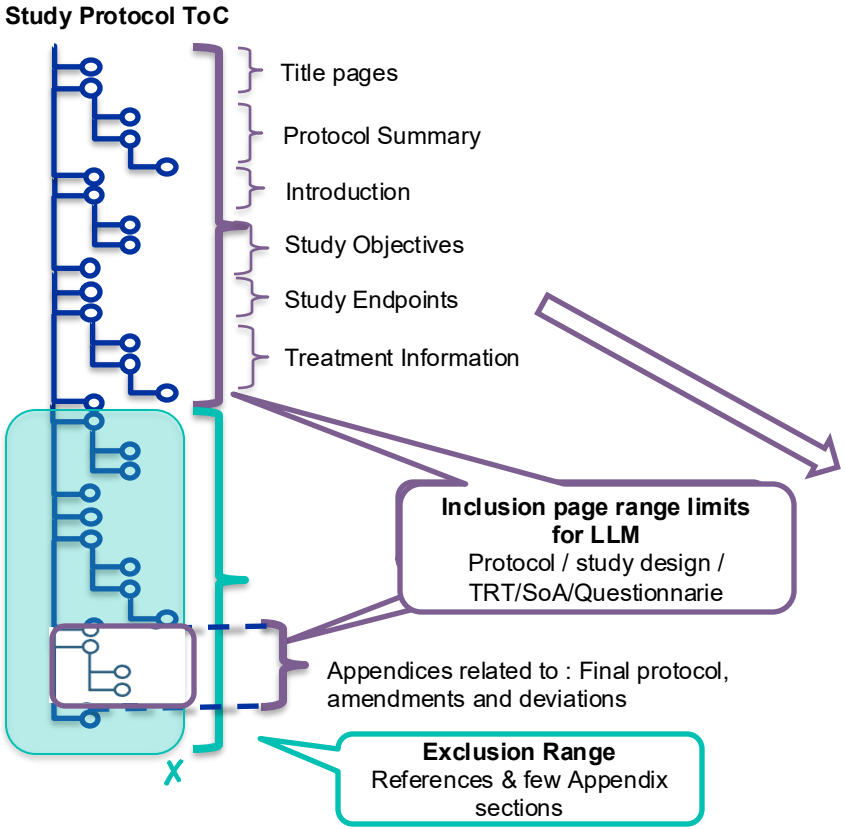
To address this, we shifted to experimenting with secure private models:

- Nov 2024: Initial LLaMA training on study protocols produced unreliable, hallucinated outputs, not viable for production.
- Jan 2025: Switched to DeepSeek inference, resulting in a major improvement in quality and reliability.
- Mar 2025: Transitioned to Gemma, achieving stable, accurate, and consistently better performance.
- Expanded to fully private deployments on AWS SageMaker and Bedrock using large models like Anthropic Sonnet to ensure compliance, scalability, and secure enterprise-grade AI operations.

Our Approach



Intelligent Parsing of Table of Contents



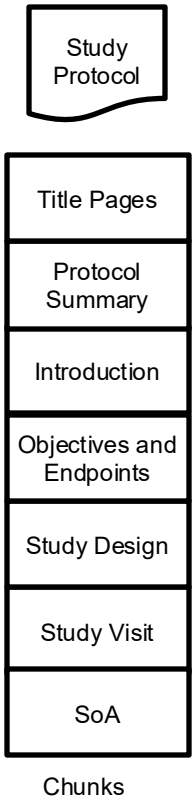
Matrix of **Parameters vs Curated section** names where parameter information likely to be found

Param	Title Pages	Protocol Summary	Introduction	Objectives and Endpoints	Study Design	Study Intervention
DOSE		✓	✓		✓	✓
DOSFRQ	✓		✓		✓	✓
OBJPRIM			✓	✓		
OBJSEC			✓	✓		
ROUTE	✓		✓		✓	✓
TRT	✓		✓		✓	

ToC extracted page ranges

		Start	End
1	Title Pages	1	3
2	Protocol Summary	6	8
3	Introduction	16	20
4	Study Objectives	22	22
5	Study Endpoints	23	24
6	Study Design	25	26
7	…		

Trial Summary Prompts Execution



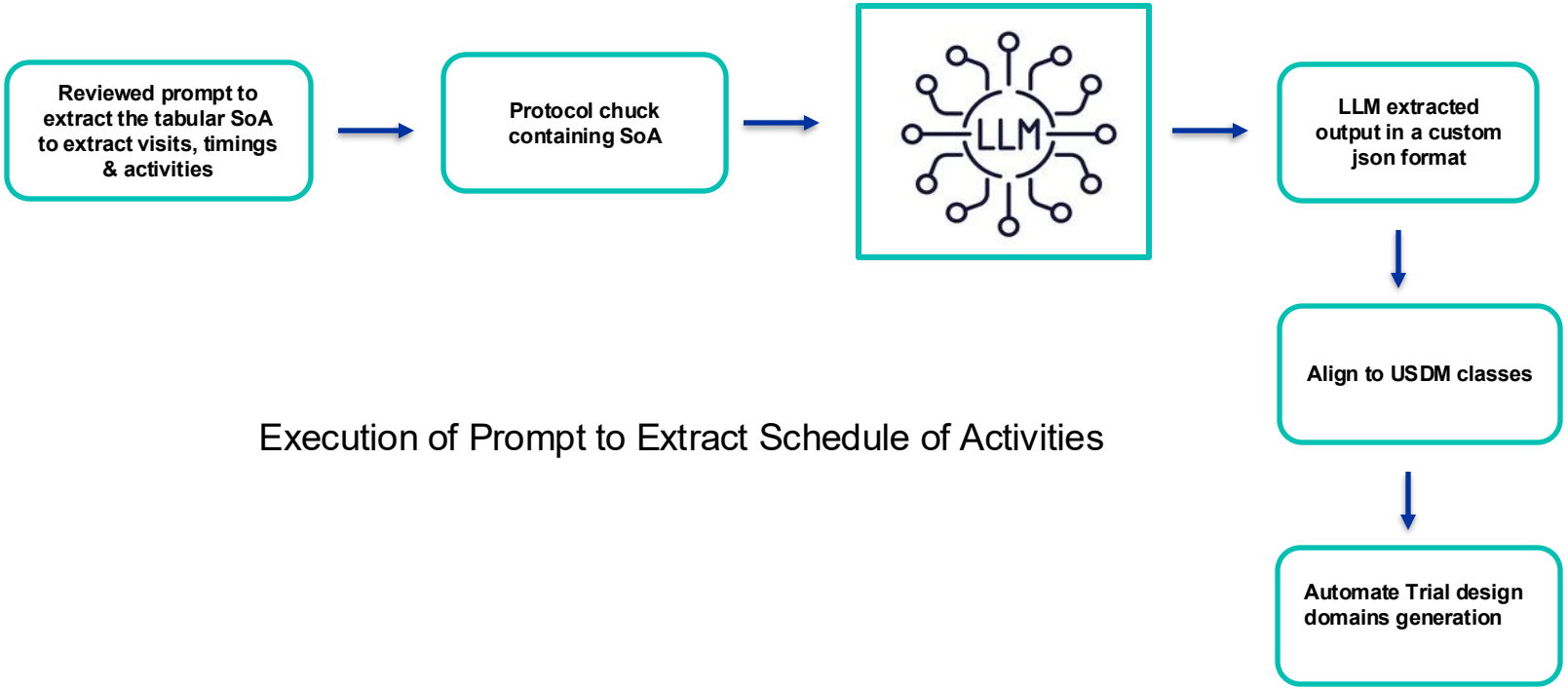
Steps to execute the prompts:

1. Collect the parameters likely to found in a section (as per the predefined matrix)
2. Consolidate all the prompts of the selected parameters into one prompt
3. Loop through all the chunks belongs to that section and execute the consolidated prompt

Param	Title Pages	Protocol Summary	Introduction	Objectives and Endpoints	Study Design	Study Intervention
DOSE		✓	✓		✓	✓
DOSFRQ	✓		✓		✓	✓
OBJPRIM			✓	✓		
OBJSEC			✓	✓		
ROUTE	✓		✓		✓	✓
TRT	✓		✓		✓	

```
Extract the following from the provided document and give response in json format:
{
   "DOSFRQ": {
      "Description": "Extract the dosing frequency of the study drug from dosing instructions or treatment plan. Do not extract or
      interpret dose cycle information or route of administration as dosing frequency. If the dosing frequency is not mentioned,
      return 'NA'.",
      "Value": "",
      "Sentences": "Return the complete sentence(s) from the text that directly support the inferred value. If the value is blank
      or no relevant supporting information is found, return 'N/A'"
   },
   "OBJPRIM": {
      "Description": "Extract only the primary objective of the clinical study from the provided text. Return only the objective
      explicitly designated as primary or described as the main purpose of the study. Exclude any secondary, exploratory,
      pharmacokinetic, safety, or procedural objectives. If no primary objective is clearly stated, return 'NA'.",
      "Value": "",
      "Sentences": "Return the complete sentence(s) from the text that directly support the inferred value. If the value is blank
      or no relevant supporting information is found, return 'N/A'"
   }
}
```

Schedule of Activities Prompt Execution



Execution of Prompt to Extract Schedule of Activities

Schedule of Activities Prompt Execution

Precise prompt to extract the tabular SoA to extract visits, timings & activities

```
Extract the Schedule of Activities data from t[illegible] image and return it in tabulated format. Capture all relevant information from tables,
footnotes, and inline citations—even if multiple tables[illegible] different treatment arms, dose escalation phases, or varied schedules. Ensure all
activity-note pairs are extracted accurately and completely. Org[illegible] extracted data under the following columns, based on the availability,

Table – The title of the table.
Activity – This refers to the specific tasks, assessments, or procedures that are conduc[illegible] during a study visit.
Cycle – A repeating period of treatment and evaluation in a clinical trial.
Cycle Window – The total length of a single treatment cycle. It defines the start and end days for that specific block of the study.
Visit – The specific label assigned to each clinical trial visit.
Visit Day – This specifies the exact day or range of days when a visit is scheduled to happen, relative to the start of the study treatment. This provides
clarity on how long after enrollment a specific visit is planned.
Visit Window – This column defines the permissible timeframe around a target visit day, allowing for some flexibility in scheduling. If not explicitly
provided, skip this field.
Epoch – A period of time that serves a purpose in the trial as a whole. This refers to a named study segment that reflects a distinct period in the study
timeline. E.g., Screening, Baseline, etc.
Notes – Extract all footnotes linked to the Schedule of Activities table in the clinical protocol by identifying superscript/subscript markers (e.g., [1], [a],
*, †, etc) that are directly tagged to the activity names, visit timepoints (column headers), or specific cell values. For each activity that has a
superscript/subscript, consolidate all related footnotes—including inline notes, cell-level citations, column-level notes, and table-level footnotes—into a
single entry formatted as: "Activity name = • [Note Marker 1, Note 1] • [Note Marker 2, Note 2] ...". If no note is present or no superscript/subscript is
tagged to the activity, return: "Activity name = None". Ensure each superscript/subscript is accurately mapped to its corresponding footnote, and avoid
duplicating identical notes across activities unless they are explicitly referenced.

Do not miss any notes attached as they are critical information to understand the activities specifically performed during the visits. Do not infer missing
values, only return explicitly stated values. Ensure accurate mapping of each extracted value to the appropriate column and maintain consistent formatting
across all fields. Follow the instructions provided above strictly and return only the tabulated response.
```

SoA

Study Period	Screen	Treatment Period															Post-treatment		
		Cycle 1						Cycle 2			C3+						EOT	SFU	
Study Day	-28 to -1	1[b]	2	3	5	8	15	1	8	15	1	2	3	5	8	15	≤7d post last dose	>28d post last dose	LTFU[a]
Window					+1d	±1d	±1d	n/a	±1d	±1d	n/a				±1d	±1d	+7d	+7d	
Informed Consent[c]	X																		
Inclusion/Exclusion	X																		
Enrollment /Dose Assignment	X																		
Demography/Medical History	X																		
ECOG	X							X			X						X		
Weight, Height[d]	X							X			X						X		
Physical Examination[e]	X	X				X	X	X	X	X	X				X[c]	X[c]	X		
Vital Signs[f]	X	X	X	X		X	X	X	X	X	X				X	X	X		
Pulse Oximetry[h]	X	X				X	X	X	X	X	X				X	X	X		
12-Lead ECG[g]		X						X			X						X		
Hematology[i]	X	X	X	X		X	X	X			X						X		
Chemistry[j]	X	X	X	X		X	X	X			X						X		
Coagulation[l]	X	X						X			X						X		
Urinalysis[k]		X						X			X						X		
Pregnancy Test[c]	X	X						X			X						X		
Serum Cytokines[l]		X	X	X		X	X	X			C3, C6[l]								
Immunophenotyping (blood)[m]		X	X	X		X	X				C3, C6[m]	C6	C6						
Serum BCMA[n]		X					X	X		X	X[n]						X		
PK Samples[o]		X	X	X	X	X	X	X	X	X	X[o]	C6	C6	C6			X		
Anti-Drug Antibodies[p]		X					X	X		X	X[p]						X		
Bone Marrow Biomarkers[q]		X									C3, C6+[q]						X		
Disease Assessment[c]	X	X					X			X						X	X[c]		
Hospital Admission[s]		X																	

Align to USDM classes to automate Trial design domains generation

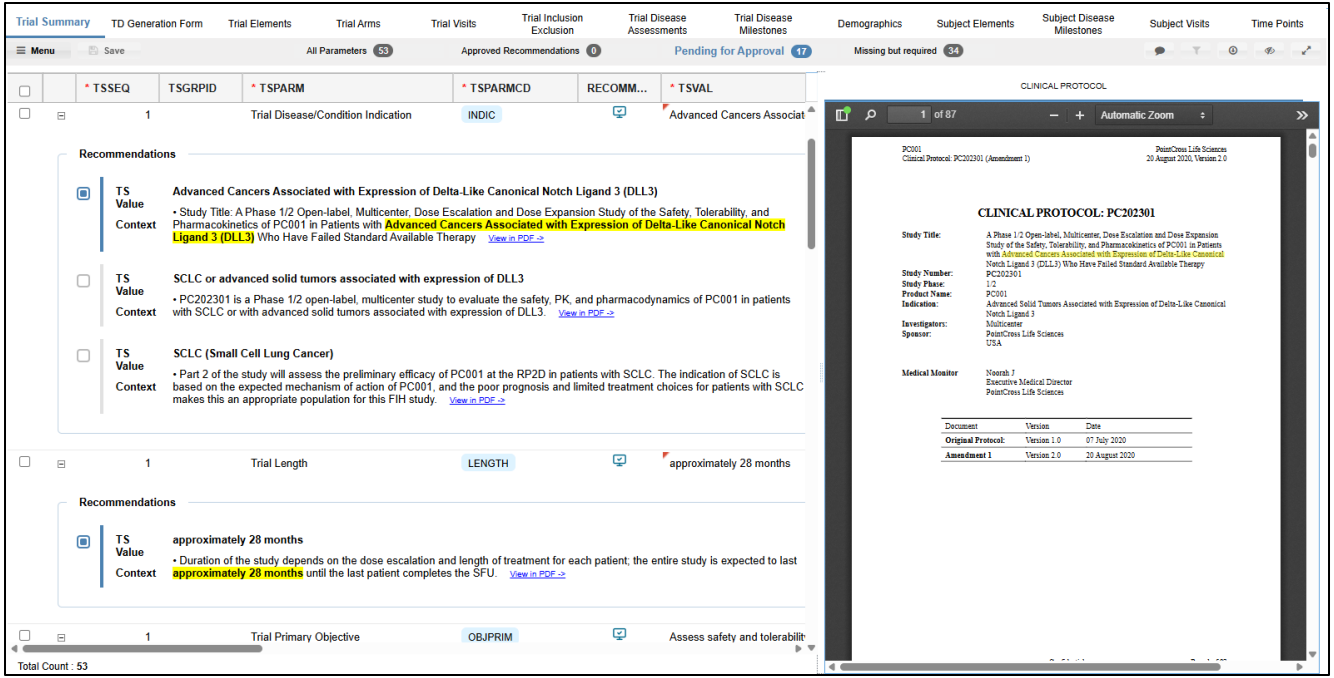
Name	Description	Label	From	Type	To	ToFrom	Value	Value Label	Window Label	Window Lower	Window Upper
TIM1	Screening period – informed consent s	Screening	SCREEN	BEFORE	C1D1	S2S	P28D	Day -28 to -1	D -28 to -1	P0D	P27D
TIM2	Cycle 1 Day 1 – First fixed dose admin	C1D1 – First Dose	C1D1	FIXED	C1D1	S2S	P1D	Day 0			
TIM3	Cycle 1 Day 2 – Early post-dose	C1D2	C1D1	AFTER	C1D2	S2S	P1D	Day 2			
TIM4	Cycle 1 Day 3 – Continued safety and	C1D3	C1D2	AFTER	C1D3	S2S	P1D	Day 3			
TIM5	Cycle 1 Day 5 – Safety and lab evaluat	C1D5	C1D3	AFTER	C1D5	S2S	P2D	Day 5	1 day after planned visit	P1D	P1D
TIM6	Cycle 1 Day 8 – Second weekly fixed d	C1D8	C1D5	AFTER	C1D8	S2S	P3D	Day 8	1 day before to 1 day after pla	P1D	P1D
TIM7	Cycle 1 Day 15 – End-of-cycle fixed do	C1D15	C1D8	AFTER	C1D15	S2S	P7D	Day 15	1 day before to 1 day after pla	P1D	P1D
TIM8	Cycle 2 Day 1 – Start of next cycle dos	C2D1	C1D15	AFTER	C2D1	S2S	P7D	Day 22			
TIM9	Cycle 2 Day 8 – Mid-cycle fixed dose	C2D8	C2D1	AFTER	C2D8	S2S	P7D	Day 29	1 day before to 1 day after pla	P1D	P1D
TIM10	Cycle 2 Day 15 – End-of-cycle dose	C2D15	C2D8	AFTER	C2D15	S2S	P7D	Day 36	1 day before to 1 day after pla	P1D	P1D
TIM11	Cycle 3+ Day 1 – Maintenance weekly	C3D1	C2D15	AFTER	C3D1	S2S	P7D	Day 43			
TIM12	Cycle 3 Day 2- Early post-dose	C3D2	C3D1	AFTER	C3D2	S2S	P7D	Day 49			
TIM13	Cycle 3 Day 3- Continued safety and P	C3D3	C3D2	AFTER	C3D3	S2S	P7D	Day 55			
TIM14	Cycle 3 Day 5- Safety and lab evaluatic	C3D5	C3D3	AFTER	C3D5	S2S	P7D	Day 61			
TIM15	Cycle 3 Day 8- Third weekly fixed dose	C3D8	C3D5	AFTER	C3D8	S2S	P7D	Day 67	1 day before to 1 day after pla	P1D	P1D

LLM extracted output to flat structure

Table	Activity	Cycle	Cycle Wi	Visit	Visit Day	Visit Win	Epoch	Notes
Schedule of Assessments – Study Visits: Single-F	Informed	-	-	Screen	-28 to -1	-	Screen	med Consent • • [Superscript 1] Long-term follow-up telephone calls or visits to occur monthly (±7 days) for 6 months, the
Schedule of Assessments – Study Visits: Single-F	Inclusion	-	-	Screen	-28 to -1	-	Screen	Inclusion/Exclusion • • [Superscript 2] Brackets designate visits/assessments that be completed within ±1d screening assessr
Schedule of Assessments – Study Visits: Single-F	Enrollme	-	-	Screen	-28 to -1	-	Screen	Enrollment /Dose Assignment • None
Schedule of Assessments – Study Visits: Single-F	Demogra	-	-	Screen	-28 to -1	-	Screen	Demography/Medical History • None
Schedule of Assessments – Study Visits: Single-F	ECOG	-	-	Screen	-28 to -1	-	Screen	ECOG • None
Schedule of Assessments – Study Visits: Single-F	ECOG	Cycle 1	-		1	-	Treatmen	ECOG • None
Schedule of Assessments – Study Visits: Single-F	ECOG	Cycle2+	-		1	n/a	Treatmen	ECOG • None
Schedule of Assessments – Study Visits: Single-F	ECOG	-	-	EOT	≤7d post	-	Post-treat	ECOG • None
Schedule of Assessments – Study Visits: Single-F	Weight, H	-	-	Screen	-28 to -1	-	Screen	Weight, Height • • [Superscript 3] Height is measured at Screening only.
Schedule of Assessments – Study Visits: Single-F	Weight, H	Cycle 1	-		1	-	Treatmen	Weight, Height • • [Superscript 3] Height is measured at Screening only.
Schedule of Assessments – Study Visits: Single-F	Weight, H	Cycle2+	-		1	n/a	Treatmen	Weight, Height • • [Superscript 3] Height is measured at Screening only.
Schedule of Assessments – Study Visits: Single-F	Weight, H	-	-	EOT	≤7d post	-	Post-treat	Weight, Height • • [Superscript 3] Height is measured at Screening only.
Schedule of Assessments – Study Visits: Single-F	Physical I	-	-	Screen	-28 to -1	-	Screen	Physical Examination • • [Superscript 4, 5] Full physical examination & Screening: abbreviated physical exams at indicated vis
Schedule of Assessments – Study Visits: Single-F	Physical I	Cycle 1	-		1	-	Treatmen	Physical Examination • • [Superscript 4, 5] Full physical examination & Screening: abbreviated physical exams at indicated vis
Schedule of Assessments – Study Visits: Single-F	Physical I	Cycle 1	-		8	±1d	Treatmen	Physical Examination • • [Superscript 4, 5] Full physical examination & Screening: abbreviated physical exams at indicated vis
Schedule of Assessments – Study Visits: Single-F	Physical I	Cycle 1	-		15	±1d	Treatmen	Physical Examination • • [Superscript 4, 5] Full physical examination & Screening: abbreviated physical exams at indicated vis
Schedule of Assessments – Study Visits: Single-F	Physical I	Cycle2+	-		1	n/a	Treatmen	Physical Examination • • [Superscript 4, 5] Full physical examination & Screening: abbreviated physical exams at indicated vis
Schedule of Assessments – Study Visits: Single-F	Physical I	C3+	-		5	±1d	Treatmen	Physical Examination • • [Superscript 4, 5] Full physical examination & Screening: abbreviated physical exams at indicated vis
Schedule of Assessments – Study Visits: Single-F	Physical I	C3+	-		8	±1d	Treatmen	Physical Examination • • [Superscript 4, 5] Full physical examination & Screening: abbreviated physical exams at indicated vis
Schedule of Assessments – Study Visits: Single-F	Physical I	-	-	EOT	≤7d post	-	Post-treat	Physical Examination • • [Superscript 4, 5] Full physical examination & Screening: abbreviated physical exams at indicated vis
Schedule of Assessments – Study Visits: Single-F	Vital Sign	-	-	Screen	-28 to -1	-	Screen	Vital Signs • • [Superscript 6] Vital signs: blood pressure, heart rate, respiratory rate, and body temperature call or text as at dr
Schedule of Assessments – Study Visits: Single-F	Vital Sign	Cycle 1	-		1	-	Treatmen	Vital Signs • • [Superscript 6] Vital signs: blood pressure, heart rate, respiratory rate, and body temperature call or text as at dr
Schedule of Assessments – Study Visits: Single-F	Vital Sign	Cycle 1	-		2	-	Treatmen	Vital Signs • • [Superscript 6] Vital signs: blood pressure, heart rate, respiratory rate, and body temperature call or text as at dr
Schedule of Assessments – Study Visits: Single-F	Vital Sign	Cycle 1	-		3	-	Treatmen	Vital Signs • • [Superscript 6] Vital signs: blood pressure, heart rate, respiratory rate, and body temperature call or text as at dr
Schedule of Assessments – Study Visits: Single-F	Vital Sign	Cycle 1	-		8	±1d	Treatmen	Vital Signs • • [Superscript 6] Vital signs: blood pressure, heart rate, respiratory rate, and body temperature call or text as at dr
Schedule of Assessments – Study Visits: Single-F	Vital Sign	Cycle 1	-		15	±1d	Treatmen	Vital Signs • • [Superscript 6] Vital signs: blood pressure, heart rate, respiratory rate, and body temperature call or text as at dr
Schedule of Assessments – Study Visits: Single-F	Vital Sign	Cycle2+	-		1	n/a	Treatmen	Vital Signs • • [Superscript 6] Vital signs: blood pressure, heart rate, respiratory rate, and body temperature call or text as at dr

Human in Loop - Review & Approvals

- Compare LLM-extracted data side-by-side with source protocol text and line references
- Study team reviews and approves extracted elements before processing
- Approved data automatically supports SDTM Trial Design generation, and study build outputs



Conclusion

Targeted prompt-driven NLP and semantic modeling enable precise extraction of key study elements directly from the protocol.

Human-in-the-loop review safeguards scientific intent while reducing manual effort and interpretation variability.

This digital solution accelerates workflows, enhances data quality, and enables full automation via USDM structure for study build and SDTM trial design with full traceability.

Thank you

Xbiom™ makes data useful

The Global Healthcare Data Science Community

Contact Channels
UK +44 1843 609600
office@phuse.global
phuse.global

Social Media
@phusetwitta
/phusebook
/phusetube
/company/phuse

Questions and Contact Information:

For questions or more information about this presentation, please contact

[Haris Khan, haris@pointcross.com]

[Arpitha Hanumanthaiah, arpitha@pointcross.com]