

Automated Annotated CRF Generation Using Large Language Models

Constantin Weberpals
PHUSE Single Day Event 2026



Agenda

1. The Naïve Approach & Learnings
2. LLM-Driven Pipeline
3. Results & Hard Problems
4. Takeaways & Questions

Can we just ask Claude?

Significant Medical History

Significant Medical History

Note: Any Significant Medical Condition existing on or before the participant signs Informed Consent should be listed on the Significant Medical History

Note: Record the condition or diagnosis (preferred)

MH number
(Auto-populated)

Medical history term

Start date
(dd/MM/yyyy OR UN/MMMM/yyyy OR UN/UNK/yyyy)

Ongoing ☐ Yes ☐ No

Stop date
(dd/MM/yyyy OR UN/MMMM/yyyy OR UN/UNK/yyyy)

Is condition currently being treated or was previously treated ☐ Yes ☐ No

If yes, please indicate whether treatment was with a concomitant medication and/or a procedure (Check all that apply) ☐ Concomitant Medication ☐ Procedure ☐ Other

Concomitant Medication ID
(Check all that apply)

Procedure
(Note: Please report the most recent procedure. Any additional procedures should be added to the Concomitant Procedures Summary.)

Procedure date
(dd/MM/yyyy OR UN/MMMM/yyyy OR UN/UNK/yyyy)

Were there any additional relevant procedures associated with this medical history ☐ Yes ☐ No

Concomitant Procedure ID
(Check all that apply)

- Single medical history page
- Capable model

Let's try

Significant Medical History

Significant Medical History

Note: Any Significant Medical Condition existing on or before the participant signs Informed Consent should be listed on the Significant Medical History

Note: Record the condition or diagnosis (preferred)

MH number
(Auto-populated)

Medical history term

Start date
(dd/MM/yyyy OR UNIMMM/yyyy OR UN/UNK/yyyy)

Ongoing ☐ Yes ☐ No

Stop date
(dd/MM/yyyy OR UNIMMM/yyyy OR UN/UNK/yyyy)

Is condition currently being treated or was previously treated ☐ Yes ☐ No

If yes, please indicate whether treatment was with a concomitant medication and/or a procedure (Check all that apply)

Concomitant Medication ID (Check all that apply)

Procedure
(Note: Please report the most recent procedure. Any additional procedures should be added to the Concomitant Procedures Summary.)

Procedure date
(dd/MM/yyyy OR UNIMMM/yyyy OR UN/UNK/yyyy)

Were there any additional relevant procedures associated with this medical history ☐ Yes ☐ No

Concomitant Procedure ID (Check all that apply)



Significant Medical History

MH (Medical History)

Significant Medical History

Note: Any Significant Medical Condition existing on or before the participant signs Informed Consent should be listed on the Significant Medical History

Note: Record the condition or diagnosis (preferred)

MH number
(Auto-populated)

Medical history term

Start date
(dd/MM/yyyy OR UNIMMM/yyyy OR UN/UNK/yyyy)

Ongoing ☐ Yes ☐ No

Stop date
(dd/MM/yyyy OR UNIMMM/yyyy OR UN/UNK/yyyy)

Is condition currently being treated or was previously treated ☐ Yes ☐ No

If yes, please indicate whether treatment was with a concomitant medication and/or a procedure (Check all that apply)

Concomitant Medication ID (Check all that apply)

Procedure
(Note: Please report the most recent procedure. Any additional procedures should be added to the Concomitant Procedures Summary.)

Procedure date
(dd/MM/yyyy OR UNIMMM/yyyy OR UN/UNK/yyyy)

Were there any additional relevant procedures associated with this medical history ☐ Yes ☐ No

Concomitant Procedure ID (Check all that apply)

What did Claude do?

Significant Medical History MH (Medical History)

Significant Medical History

Note: Any Significant Medical Condition existing on or before the participant signs Informed Consent should be listed on the Significant Medical History

Note: Record the condition or diagnosis (preferred)

MH number (derived - sequence number) 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99 100 101 102 103 104 105 106 107 108 109 110 111 112 113 114 115 116 117 118 119 120 121 122 123 124 125 126 127 128 129 130 131 132 133 134 135 136 137 138 139 140 141 142 143 144 145 146 147 148 149 150 151 152 153 154 155 156 157 158 159 160 161 162 163 164 165 166 167 168 169 170 171 172 173 174 175 176 177 178 179 180 181 182 183 184 185 186 187 188 189 190 191 192 193 194 195 196 197 198 199 200 201 202 203 204 205 206 207 208 209 210 211 212 213 214 215 216 217 218 219 220 221 222 223 224 225 226 227 228 229 230 231 232 233 234 235 236 237 238 239 240 241 242 243 244 245 246 247 248 249 250 251 252 253 254 255 256 257 258 259 260 261 262 263 264 265 266 267 268 269 270 271 272 273 274 275 276 277 278 279 280 281 282 283 284 285 286 287 288 289 290 291 292 293 294 295 296 297 298 299 300 301 302 303 304 305 306 307 308 309 310 311 312 313 314 315 316 317 318 319 320 321 322 323 324 325 326 327 328 329 330 331 332 333 334 335 336 337 338 339 340 341 342 343 344 345 346 347 348 349 350 351 352 353 354 355 356 357 358 359 360 361 362 363 364 365 366 367 368 369 370 371 372 373 374 375 376 377 378 379 380 381 382 383 384 385 386 387 388 389 390 391 392 393 394 395 396 397 398 399 400 401 402 403 404 405 406 407 408 409 410 411 412 413 414 415 416 417 418 419 420 421 422 423 424 425 426 427 428 429 430 431 432 433 434 435 436 437 438 439 440 441 442 443 444 445 446 447 448 449 450 451 452 453 454 455 456 457 458 459 460 461 462 463 464 465 466 467 468 469 470 471 472 473 474 475 476 477 478 479 480 481 482 483 484 485 486 487 488 489 490 491 492 493 494 495 496 497 498 499 500 501 502 503 504 505 506 507 508 509 510 511 512 513

What actually happened

 Install PyMuPDF for PDF annotation

Script



 Rasterize the uploaded CRF page to understand its exact layout

Script



 Get exact page dimensions

Script



 View the rasterized CRF page

 Now I have a clear view of the CRF page. Let me map each field to SDTM variables and create annotations.

The page dimensions are 612 x 792 (standard letter, 8.5" x 11" at 72dpi).

Let me identify the approximate positions of each field (in PDF coordinates, where origin is bottom-left):

Looking at the image (200dpi render, 935px wide x ~1210px tall):

- Scale factor: $612/935 = 0.655$ for x, $792/?$ for y
- The PDF is 792pt tall (11 inches, letter size), so I need to work from the PDF coordinates

Show more



 Get text block positions to understand layout

Script

 Now I have the exact positions of all text elements. Let me also find the positions of the input boxes (rectangles/form fields).

Now I can create the annotated PDF. Let me plan the annotations:

MH Domain annotations:

Standard MSG v2.0 approach:

- Domain label: "MH (Medical History)" placed at top of the form section
- Solid border boxes = collected variables

Show more



 Get rectangle positions (input fields and form borders)

Script

 Now I have the positions of all the rectangles. Let me understand the coordinate system:

- PDF coordinates: (0,0) is at bottom-left, y increases upward
- But PyMuPDF uses top-left origin, y increases downward

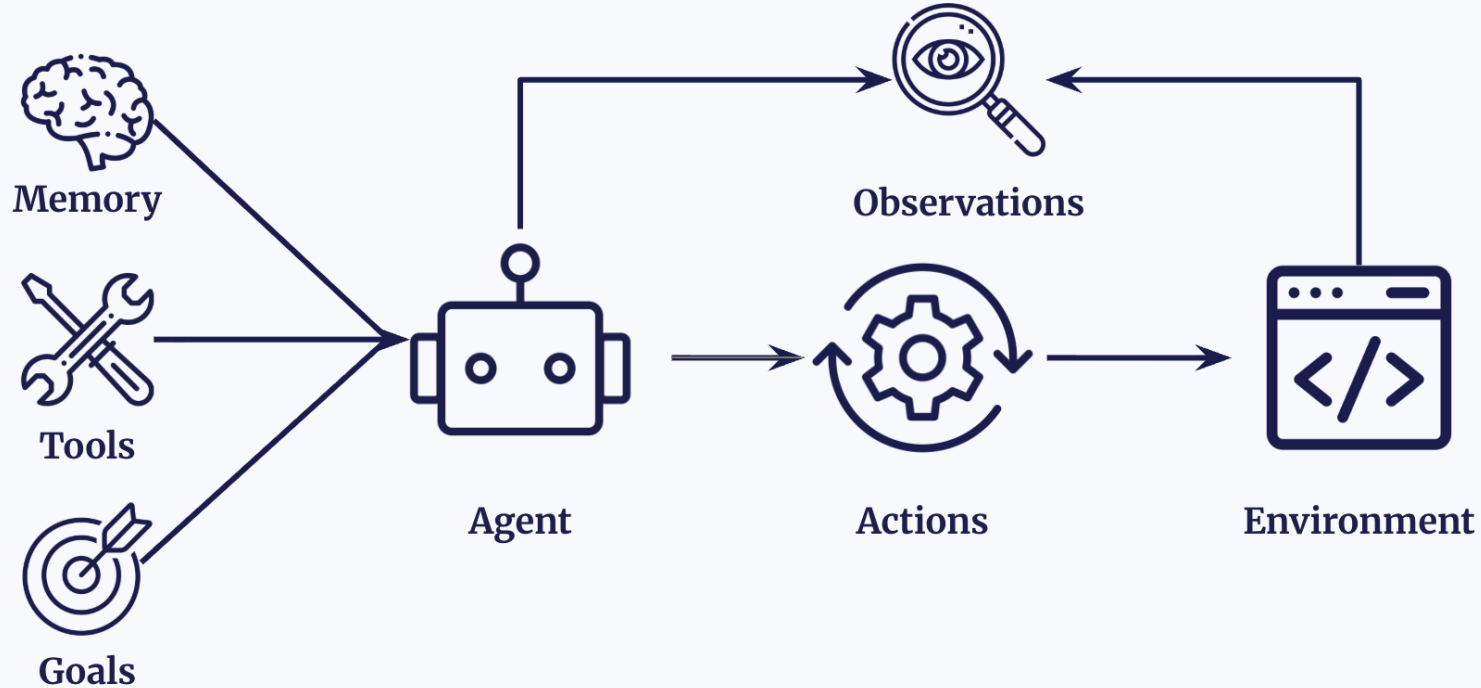
Wait, in PyMuPDF, the coordinates are in the PDF coordinate space where (0,0) is top-left? Let me check... actually PyMuPDF uses the device coordinate space where (0,0) is at top-left and y increases downward. But the PDF page height is 792.

Looking at the y-values from the text blocks:

Show more

 Done

How does an AI Agent work?



Naïve approach – Let's analyze the results

The image shows a screenshot of a 'Significant Medical History' form. The form is titled 'Significant Medical History' and includes a sub-header 'MH (Medical History)'. It contains several sections with text input fields and checkboxes. The form is annotated with various boxes and lines, indicating areas of interest or error. These annotations include:

- A red box around the 'Medical history term' field.
- A red box around the 'Start date' field.
- A red box around the 'Ongoing' checkbox.
- A red box around the 'Stop date' field.
- A red box around the 'Is condition currently being treated or was previously treated' checkbox.
- A red box around the 'Concomitant Medication' section.
- A red box around the 'Procedure' section.
- A red box around the 'Were there any additional relevant procedures associated with this medical history' checkbox.
- A red box around the 'Concomitant Procedure ID' field.

The form also includes a footer note: '(NOT SUBMITTED) EDC images (consent, history, CM and PR data are captured on dedicated CRF pages)'.

What didn't work:

- Other domains on this form just labeled [NOT SUBMITTED] -> no recognition of cross-domain relationship or target variable
- Unexpected/excessive derived variables
- Wrong formatting of annotations and overlapping placement
- Hallucinations

-> Not useful like this

Naïve approach – Let's analyze the results

Significant Medical History MH (Medical History)

Significant Medical History

Note: Any Significant Medical Condition existing on or before the participant signs Informed Consent should be listed on the Significant Medical History

Note: Record the condition or diagnosis (preferred)

MH number (Auto-populated) MHSEQ (derived sequence number) STUDYID (derived) DOMAIN > MH

Medical history term MHTERM (collected) MHSTDTCT (derived)

Start date (dd/MM/yyyy OR UN/MMM/yyyy OR UN/UNK/yyyy) MHSTDTCT (collected) MHDECODE (derived) MHBODSYS (derived) MHENRF (derived)

Ongoing ☐ No MHENRF (derived) MHSTDTCT (derived)

Stop date (dd/MM/yyyy OR UN/MMM/yyyy OR UN/UNK/yyyy) MHSTDTCT (collected) MHDECODE (derived) MHBODSYS (derived) MHENRF (derived)

Is condition currently being treated or was previously treated ☐ No SUPPXX (derived)

If yes, please indicate whether treatment was with a concomitant medication and/or a procedure (Check all that apply)

Concomitant Medication ID (Check all that apply) ☐ Concomitant Medication NOT SUBMITTED

Procedure (Note: Please report the most recent procedure. Any additional procedures should be added to the Concomitant Procedures Summary.) ☐ Procedure NOT SUBMITTED

Procedure date (dd/MM/yyyy OR UN/MMM/yyyy OR UN/UNK/yyyy) ☐ Other NOT SUBMITTED

Were there any additional relevant procedures associated with this medical history ☐ Yes NOT SUBMITTED

Concomitant Procedure ID (Check all that apply) ☐ No NOT SUBMITTED

(NOT SUBMITTED) EDC image - cross-reference fields, CM and PR data are captured on dedicated CRF pages

What worked:

- Domain label "MH (Medical History)"
parenthetical style is MSG v2.0-compliant
- Border type logic — solid line for collected variables (MHTERM, MHSTDTCT, ...), dashed for derived/assigned (MHSEQ, STUDYID, MHDECODE, MHBODSYS, MHENRF)
- SUPPXX identification: SUPPMH: MHTREAT
- [NOT SUBMITTED] handling

Without structure, it is like handing a **smart intern a blank CRF and saying "annotate this"** — without giving them the SDTMIG, a mapping template, or a review checklist.

Where LLMs add value and where they don't

LLM Strengths

- They are strong at semantic interpretation
- They can handle noisy CRF text
- They are good at tool calling

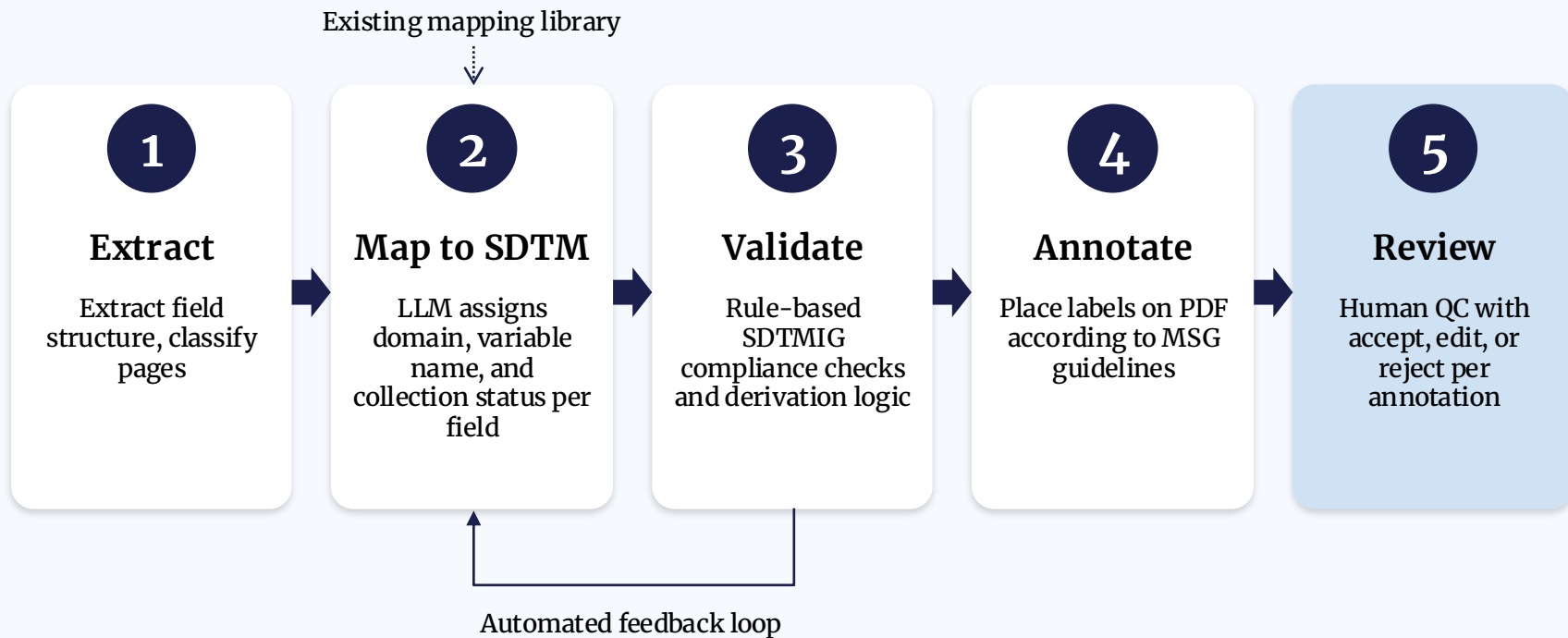
Structured Logic

- They are weak at strict geometry and coordinates
- They are weak at repeated-group and multi-page continuity logic
- They are not reliable enough for standards compliance alone

Hybrid Approach

- LLM-guided annotation inside a rule-based, validated pipeline.
- Deterministic code is needed for validation, normalization, and reproducibility.

CRF annotation can be decomposed into clear steps



Giving the agent structure, tools and guardrails

Significant Medical History MH (Medical History)

Significant Medical History

Note: Any Significant Medical Condition existing on or before the participant signs Informed Consent should be listed on the Significant Medical History

Note: Record the condition or diagnosis (preferred)

MH number
(Auto-populated)

Medical history term

Start date
(dd/MM/yyyy OR UN/MM/yyyy OR UNUNK/yyyy)

Ongoing

Stop date
(dd/MM/yyyy OR UN/MM/yyyy OR UNUNK/yyyy)

Is condition currently being treated or was previously treated

If yes, please indicate whether treatment was with a concomitant medication and/or a procedure
(Check all that apply)

Concomitant Medication ID
(Check all that apply)

Procedure
(Note: Please report the most recent procedure. Any additional procedures should be added to the Concomitant Procedures Summary.)

Procedure date
(dd/MM/yyyy OR UN/MM/yyyy OR UNUNK/yyyy)

Were there any additional relevant procedures associated with this medical history

Concomitant Procedure ID
(Check all that apply)



MH (Medical History) PR (Procedures)
Significant Medical History CM (Concomitant Medications)

Significant Medical History

Note: Any Significant Medical Condition existing on or before the participant signs Informed Consent should be listed on the Significant Medical History

Note: Record the condition or diagnosis (preferred)

MH number MHSPID

Medical history term MHTERM

Start date
(dd/MM/yyyy OR UN/MM/yyyy OR UNUNK/yyyy) MHSTDTC

Ongoing ☐ Yes MHENRTPT
☐ No

Stop date
(dd/MM/yyyy OR UN/MM/yyyy OR UNUNK/yyyy) MHENDTC

Is condition currently being treated or was previously treated ☐ Yes MHTRT in SUPPMH
☐ No

If yes, please indicate whether treatment was with a concomitant medication and/or a procedure
(Check all that apply)

Concomitant Medication ID CMSPID

Procedure PRTERM

Procedure date
(dd/MM/yyyy OR UN/MM/yyyy OR UNUNK/yyyy) PRSTDTC

Were there any additional relevant procedures associated with this medical history ☐ Yes MHADDP in SUPPMH
☐ No

Concomitant Procedure ID
(Check all that apply) PRSPID

Another example

VS (Vital Signs)
Vital Signs

Vital Signs

Were vital signs collected ☐ Yes **VSSTAT**
☐ No

If no, explain **VSREASND**

Evaluator initials **VSEVAL** in SUPPVS
(XXX: Enter a '-' for no middle initial)

Blood Pressure

Note: Measure after 5 minutes of rest in sitting position.

Was blood pressure assessed ☐ Yes **VSSTAT**
☐ No

If no, explain **VSREASND**

Systolic blood pressure (mmHg) **VSRORRES / VSORRESU when VSTESTCD=SYSBP**
(Results: XXX)

Diastolic blood pressure (mmHg) **VSRORRES / VSORRESU when VSTESTCD=DIABP**
(Results: XXX)

Pulse Rate

Note: Measure after 5 minutes of rest in sitting position.

Was pulse rate assessed ☐ Yes **VSSTAT**
☐ No

If no, explain **VSREASND**

Pulse rate (Beats/minute) **VSRORRES / VSORRESU when VSTESTCD=PULSE**
(Results: XXX)

Repeat Vital Signs

Were vital signs repeated ☐ Yes **VSREP** in SUPPVS
☐ No

Blood Pressure

Note: Measure after 5 additional minutes of rest in sitting position.

Was repeat blood pressure assessed ☐ Yes **VSSTAT**
☐ No

VS (Vital Signs)

If no, explain **VSREASND**

Systolic blood pressure (mmHg) **VSRORRES / VSORRESU when VSTESTCD=SYSBP**
(Results: XXX)

Diastolic blood pressure (mmHg) **VSRORRES / VSORRESU when VSTESTCD=DIABP**
(Results: XXX)

Pulse Rate

Note: Measure after 5 additional minutes of rest in sitting position.

Was repeat pulse rate assessed ☐ Yes **VSSTAT**
☐ No

If no, explain **VSREASND**

Pulse rate (Beats/minute) **VSRORRES / VSORRESU when VSTESTCD=PULSE**
(Results: XXX)

Benefits of an AI enabled approach to aCRF



Generic to any CRF structure

Handles non-standard layouts and freeform text where purely deterministic rules break down



Draft in minutes, not days

Turns multi-day manual annotation work into a near-instant first pass ready for review



Traceable annotation decisions

Each mapping decision can be linked back to the LLM's reasoning, making QC easier/transparent



Continuously improving

LLM capabilities are advancing rapidly — accuracy gains can compound over time

The hard problems: Where LLM automation struggles

Multi-visit detection

Tracking variable groups reliably across large context windows

Dense checkbox and criteria blocks

Hierarchical inclusion/exclusion grids are difficult to parse and annotate consistently

Output consistency across re-runs

LLM non-determinism causes annotation drift; seeding and temperature control only partially help

Ambiguous domain selection

Sparse CRF context makes domain assignment uncertain without protocol knowledge

Missing context on the CRF

Forms often omit study-specific coding conventions that a human annotator would know

Annotation placement on crowded pages

Avoiding overlapping labels and connector clutter on dense multi-column forms

Key Takeaways

- Naïve LLM prompting is not enough: structure, tools, and guardrails are essential
- A hybrid approach can work: LLMs for semantic reasoning, deterministic code for validation
- Decomposing the pipeline into clear steps is what makes automation possible
- Current capabilities can already produce useful first drafts, but human review remains essential
- The technology is improving rapidly

Please reach out if you have any questions!



Constantin Weberpals
constantin.weberpals@cdtm.com

