



PHUSE · SINGLE-DAY EVENT

Building a Persona Stack

Four layers that turn a general-purpose model into a
Persona assistant

Kanishk Singh – J&J Innovative Medicine

Agent Prototypes

what exactly?

MCP + API Layer

what's true now?

Agent Skill

how, repeatably?

Claude Project

who am I?

Disclaimer: The views and opinions expressed in this presentation are my own and do not necessarily reflect those of Johnson & Johnson. Shared for informational and educational purposes only; it does not constitute official company guidance.

The gap is not capability. It is context.

Ask a standalone model to apply functional area standards, cite your SOPs, or reason through validation rules (such as CDISC) — the answers may be vague



Standalone LLM

generic knowledge only

- Talks about functional area or persona needs in the abstract
- No functional area standard(such as GCDMP) to anchor to
- Limited knowledge to sponsor specific SOPs and templates
- Confident, potentially not ready to use output



Persona-Stacked Agent

same model + four layers

- ✓ Cites the functional standards it used
- ✓ Separates Must Have, Optional or Best Practice
- ✓ Knows what it must not answer
- ✓ Output a persona can review & build on

The hypothesis: the model is not the variable. What you stack around it is — and each layer is separable, testable, and reusable.

The persona stack

Each layer answers one question about the agent — and contributes one thing to its identity.



Built bottom-up. Each layer is only as good as the one beneath it.

Four concepts, in plain terms

Ground terms referred during presentation ,referring Clinical Data Management functional area



Identity / Knowledge base

The documents you choose to give an agent so it can look things up. Not training data — files it retrieves from at answer time and can cite back to you.

Layer 1 — GCDMP chapters, DMP templates, SOPs



Skill

A packaged procedure: a folder of instructions, templates and scripts the agent loads when a task matches its description. Same job, same way, every run.

Layer 2 — the DMP drafting procedure



MCP

Model Context Protocol — an open standard for connecting a model to outside tools and live data through one common interface, instead of hard-wiring each system.

Layer 3 — CT.gov, FAERS, PubMed



Agent

A model given a defined scope, a set of tools and a job to finish. It works through the steps rather than answering one question — and knows what to escalate.

Layer 4 — DMP Draft Agent, Protocol Reviewer

One sentence version: a knowledge base is what it knows, a skill is how it works, MCP is what it can reach, and an Agent is what does the work.

Layer 1 — Writing the persona

The Identity is the system prompt: who the agent is, how it speaks, what it assumes. Specificity is a constraint — a vague identity produces vague output.



CDM Agent

the persona, as written

Certification

SCDM DM Certifications

Experience

5 years ·

Therapeutic areas

Oncology · Immunology · CNS

Phases

I (FIH) · II (adaptive) · III (pivotal)

Settings

Sponsor-side and CRO oversight

Team

Mentors 2–3 junior CDAs

Why the detail earns its place



It sets a difficulty floor

"5 years, Phase I–III, oncology" makes the agent answer at a working practitioner's level — not an intro-to-CDM level.



It creates a default posture

Risk-averse on data integrity. Compliance-first. When convenience and data quality conflict, it picks quality without being asked.



It forces the vocabulary

eCRF not "form". Query not "question". Database lock not "closing the database". Wrong words tell a CDM the output is not theirs.

Why it matters: without the persona, the agent can still cite the right sources — it just loses the judgement to apply them well.

Layer 1 — Scope and boundaries

Scope is written into the same persona: what it owns, what it flags, what it hands off. What an agent refuses to answer is part of its credibility.



In scope

full expertise

- GCDMP-covered topics
- CDISC standards application
- EDC config & best practice
- Medical coding guidance
- Data quality strategy & metrics
- Vendor / CRO oversight



Working knowledge

will flag its limits

- Biostatistical concepts
- Regulatory submission process
- Pharmacovigilance workflows
- Clinical operations process
- DCT data management
- RWD / RWE — emerging



Out of scope

redirects, does not guess

- Patient safety judgement → Medical Monitor
- Analysis methodology → Biostatistics
- Filing strategy → Regulatory Affairs
- Privacy opinions → Legal
- Infrastructure → IT / Systems

The CDM reflex: escalate rather than speculate. An agent that names the right function is more useful than one that improvises an answer.

Layer 1 — Curating the knowledge base

The knowledge base is the agent's reference library — documents it retrieves from and cites. Not everything you have access to but what a CDM would actually reach for.

P1	Core reference	GCDMP chapters (28) · SCDM position papers & topic briefs · CDISC CDASH / SDTM / ADaM quick references
P2	Templates	DMP template (15 sections) · edit check specification · database lock checklist · query text library
P3	Regulatory	ICH E6(R2/R3) · E8(R1) · E9(R1) · 21 CFR Part 11 · FDA eSource & standardized study data guidance
P4	Therapeutic area	RECIST 1.1 & iRECIST · DAS28 / SLEDAI · ADAS-Cog, CDR-SB, EDSS



The unlock

Every GCDMP chapter splits its guidance two ways:

Minimum Standards

required for data integrity

Best Practices

efficiency and lower risk

That structure hands the agent a severity ladder — what you must do before what you should do.

Curation beats volume: a smaller, well-named corpus retrieves better than everything you own. Filenames are part of the design.

Layer 2 — Anatomy of a skill

A skill is a packaged, reusable procedure — a folder, not a prompt. Four parts, each doing a different job.

```
oncology-sdtm-mapping / SKILL.md

---
name: oncology-sdtm-mapping
description: Generate an EDC-to-SDTM
  mapping matrix for oncology solid-
  tumor trials. Use when the user needs
  a CRF/EDC-to-SDTM specification...
---

## Procedure
1. Read the CRF / form inventory
2. Map Rave Form OID → SDTM domain
3. Derive variable-level rows + logic
4. Emit workbook, then self-check
## Resources
templates/ · references/ · scripts/
## What it produces
Five-tab workbook + CSV · 36 domains, 16 at variable level
```

The four parts

1

Trigger *name + description*

It decides when the skill fires — so it names the task, the inputs and the output, not the topic.

2

Procedure *the numbered steps*

The sequence you would give a new hire. Order matters: it is what stops the agent improvising a different route each run.

3

Bundled resources *templates, references, scripts*

Loaded only when needed — a mapping spec, a controlled-terminology list, a helper script. Keeps the instruction file short.

4

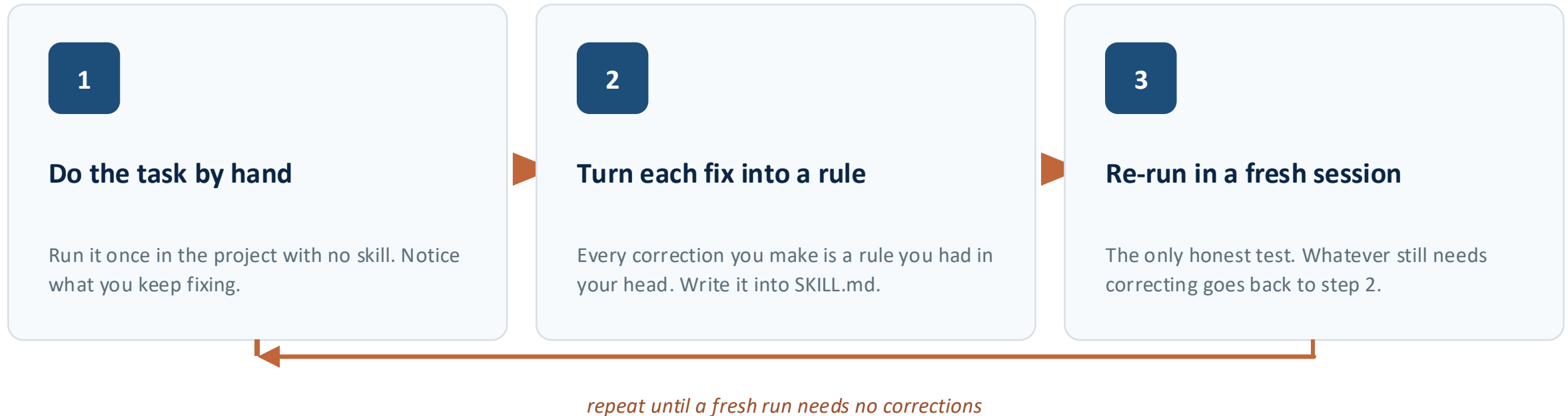
Output contract *the shape of the deliverable*

Tabs, columns, file format, naming. Without it you get the right content in an unusable shape.

All four in place: the skill fires on the right task, runs the same steps every time, pulls in what it needs, and returns the result in a usable shape.

Layer 2 — How a skill gets built

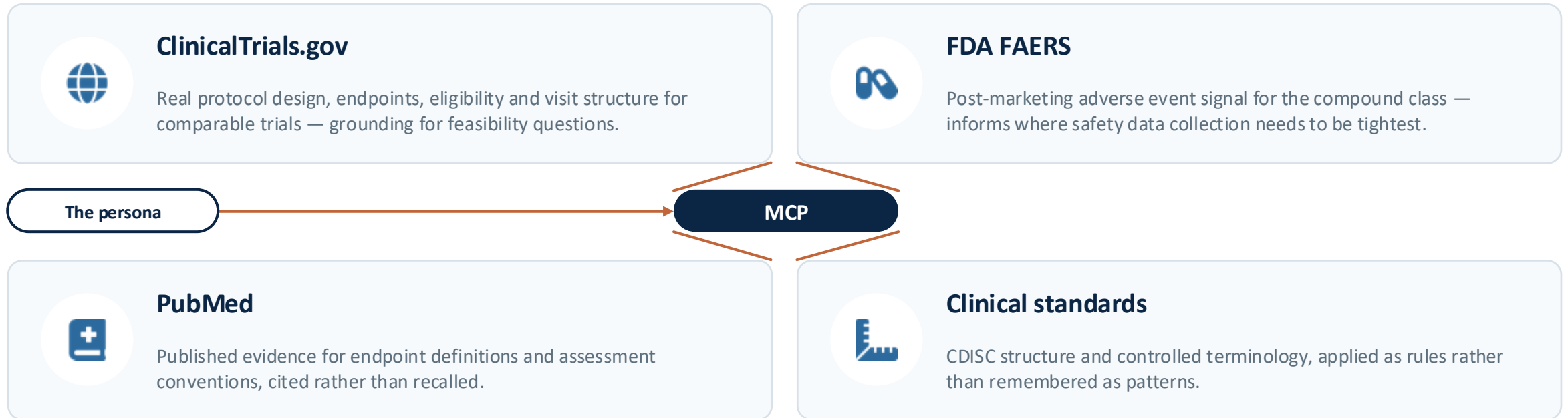
Skills are not authored in one pass. They are harvested from the corrections you were already making by hand.



The shortcut nobody takes: step 3. Your current chat already knows too much — only a clean session tells you what the skill actually carries on its own.

Layer 3 — Wiring in live evidence

MCP is a standard interface between the model and outside systems. Training data has a cutoff; clinical practice does not — this is how the persona checks its facts.



The shift this creates

From *"a Phase 3 NSCLC trial usually collects..."* → **"these registered NSCLC trials collect exactly this — here is the record."**

Layer 4 — One persona, many jobs

An agent is the stack plus a defined job. Breaking work into focused agents beats one massive prompt — same foundation, narrower scope on top.

LAYERS 1–3 · ONE SHARED PERSONA STACK — every agent inherits it unchanged



DMP Draft Agent

Writes the document

Section structure, house style, study-specific tailoring. Optimised for coverage and consistency. Today's worked example.



Protocol Reviewer

Reads before others write

Scans a protocol for DM feasibility: visit schedule, complex assessments, external data sources.



...and others

same stack, narrower scope

Each new agent inherits Layers 1–3 unchanged and adds only its own task definition.

...and the handoff is explicit: each agent knows what it owns and which function to escalate to. Ownership boundaries are designed, not assumed.



Why splitting wins: a narrow scope means fewer competing instructions, a shorter path to the right knowledge, and a failure you can actually locate.

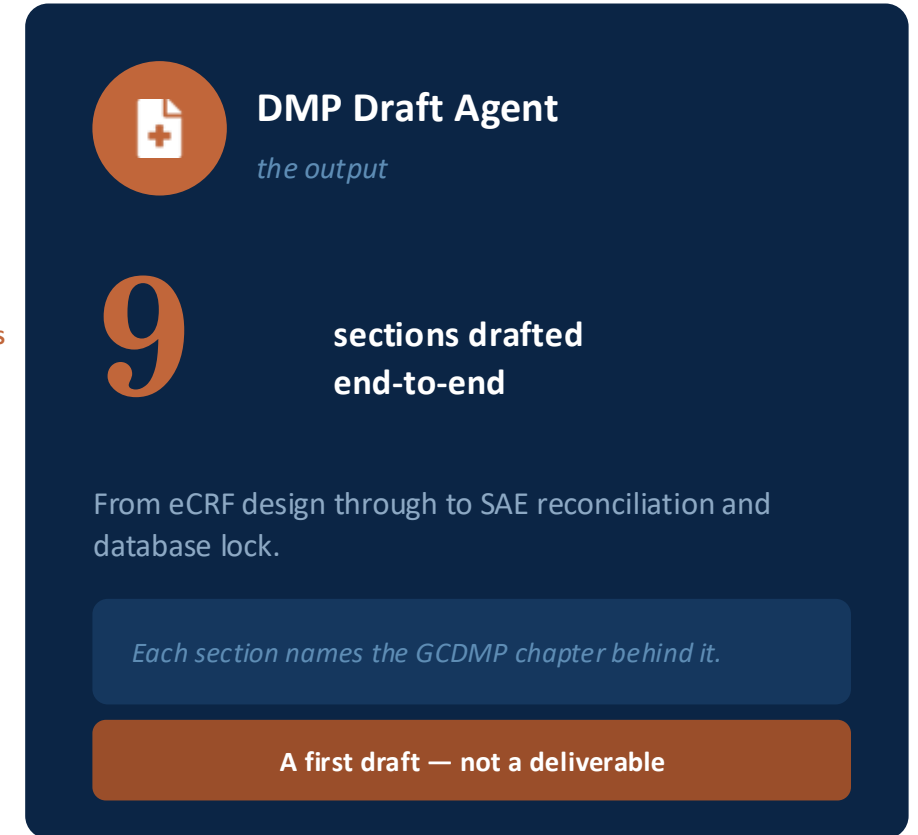
Putting the stack to work

A mock Phase 3 NSCLC trial. The brief: draft a Data Management Plan.

What each layer contributed to this one task

- 1 IDENTITY + KNOWLEDGE**
Section structure from the DMP template; GCDMP chapter cited per section; oncology conventions assumed, not asked for.
- 2 REPEATABLE PROCEDURE**
A fixed section order and output contract — so a second run is comparable to the first.
- 3 LIVE EVIDENCE**
Comparable registered NSCLC trials consulted for visit structure and endpoint definitions.
- 4 TASK SPECIALIZATION**
Only the DMP Draft Agent ran. Narrow scope, no competing instructions.

PRODUCES
→



Four layers, one agent → nine DMP sections drafted in a single pass. The stack did the remembering, not the prompt.

DMP components and the artifacts they drive

A DMP is not a standalone document. Each section is the specification for something downstream.

DMP SECTION	GCDMP ANCHOR	DOWNSTREAM ARTIFACT
■ Study overview & design	<i>Data Management Plan</i>	DM feasibility assessment
■ Roles & responsibilities	<i>Project Management for CDMs</i>	RACI matrix
■ Data collection & eCRF design	<i>Design & Development of DCIs</i>	eCRF spec · CRF Completion Guidelines
■ Database design & validation	<i>Database Validation & Programming</i>	UAT test plan
■ Data entry & query management	<i>Data Entry Processes</i>	Query text library
■ Edit check strategy	<i>Edit Check Design Principles</i>	Edit Check Specification
■ External data management	<i>External Data Transfers</i>	Data Transfer Agreement outline
■ Medical coding	<i>Coding Dictionary Management</i>	Coding conventions & version log
■ SAE reconciliation & database lock	<i>SAE Reconciliation · Database Closure</i>	Reconciliation tracker · pre-lock checklist

Why this matters: the DMP is where the agent earns its keep — one grounded draft seeds nine downstream specifications a CDM would otherwise start cold.

What I learned

Four layers, four lessons



Persona sets the altitude

A specific identity — certification, years, therapeutic areas — decides whether you get a textbook answer or a practitioner's one.



Curation beats volume

A smaller, well-named corpus retrieves better than everything you own. GCDMP's Minimum vs Best Practice split gives a severity ladder for free.



Context beats recall

Wired to CT.gov, FAERS and PubMed, claims come with a record behind them instead of a confident guess.



Narrow agents win

Splitting the work beat one massive prompt — better output, and failures you can actually locate.



Human-in-the-loop, always. This gets a data manager past the blank page. It does not get them to a deliverable — expert review is not optional.

Thank you , Keep Building !!!