IMEN891K / IDSC501 / AXGS-501 · DATA ENGINEERING · FALL 2026

15.

LLM Systems

Sangwoo Mo · POSTECH

Industrial & Management Engineering

Where We Are



Part III connects stored data to predictions, answers, and actions

Outline: LLM Systems

01

Inputs and Evidence

a question, the supplied material, and a candidate answer

02

Documents and Search

prepare readable pieces and find the current rule

03

Answers and Checks

connect evidence to the answer and check support

The P-17 Question

What is the pressure alarm threshold for P-17?



The answer needs the current P-17 manual

01 · Inputs and Evidence

01

Inputs and Evidence

a question, the supplied material, and a candidate answer

02

Documents and Search

prepare readable pieces and find the current rule

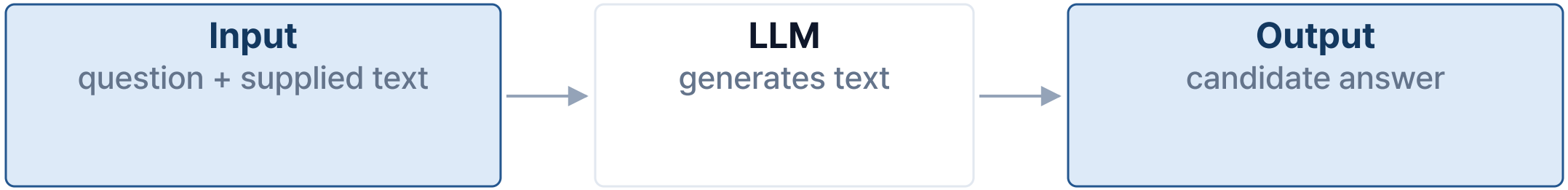
03

Answers and Checks

connect evidence to the answer and check support

Large Language Models

A large language model (LLM) generates text from its input

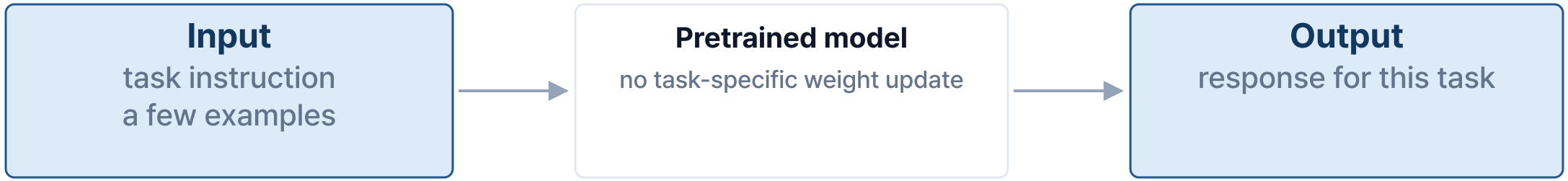


The input is visible; the generated answer still needs checking

Language Models: A Turning Point

In 2020, GPT-3 showed how task instructions and examples could guide one pretrained model

2020 · GPT-3: Language Models are Few-Shot Learners



Changing the request is different from training a new model

The input can describe a task without retraining the model for each request

Question Without Evidence

A question alone does not supply the current manual

Input

P-17 pressure alarm threshold?
No manual text supplied



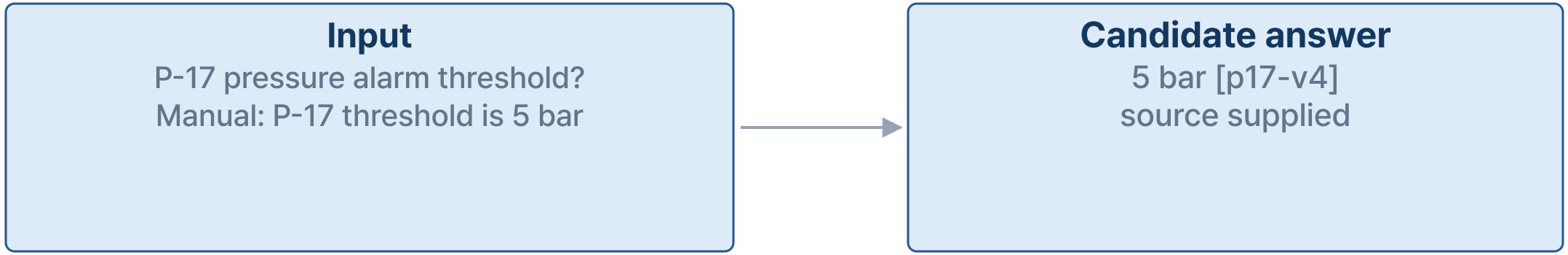
Candidate answer

7 bar
no source to inspect

7 bar is an unsupported example answer

Question With Evidence

The manual gives the model a source it can use



The supplied P-17 rule states 5 bar

Prompt

The prompt is the input sent to the model

PROMPT: INPUT SENT TO THE MODEL

Use only evidence. Cite the source. Say unknown if missing.

Question: P-17 pressure alarm threshold?

Context: no manual text yet

Instructions and the question share one input

Context

Context includes the material supplied for this request

PROMPT: INPUT SENT TO THE MODEL

Use only evidence. Cite the source. Say unknown if missing.

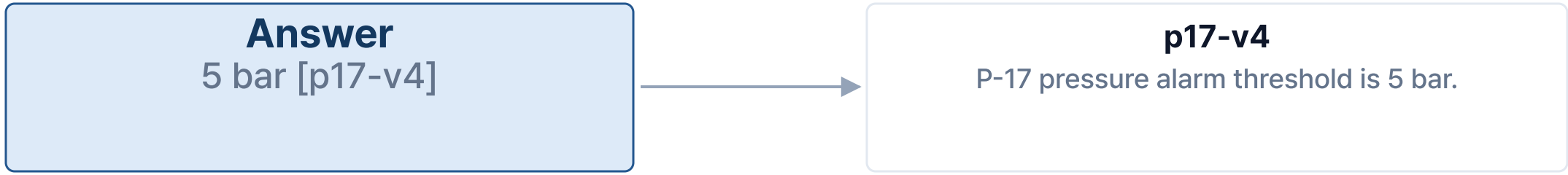
Question: P-17 pressure alarm threshold?

Context: [p17-v4] P-17 pressure alarm threshold is 5 bar.

The manual text becomes part of the model input

Answer With a Source

A source reference lets a reader inspect the claim



The source label gives the reader a place to check

5 bar points back to p17-v4

A Collection of Manuals

The collection contains similar rules for different cases

Source	Text	Eligibility
p17-v4	P-17 pressure alarm threshold is 5 bar. Stop if pressure exceeds 5 bar.	P-17, approved
p17-v2	P-17 pressure alarm threshold is 7 bar. This obsolete version must not guide the current alarm.	P-17, obsolete
p18-v3	P-18 pressure alarm threshold is 3 bar. This rule applies to P-18 only.	P-18, approved
p17-vibration	P-17 vibration inspection: record vibration and temperature before changing the operating speed.	P-17, approved

P-17 has a current rule, an obsolete rule, and a neighbouring pump

02 · Documents and Search

01

Inputs and Evidence

a question, the supplied material, and a candidate answer

02

Documents and Search

prepare readable pieces and find the current rule

03

Answers and Checks

connect evidence to the answer and check support

Why Document Preparation

The answer needs the rule and the equipment name together

Source table

Equipment	Threshold
P-17	5 bar
P-18	3 bar

Flattened text

P-17 P-18 5 bar 3 bar

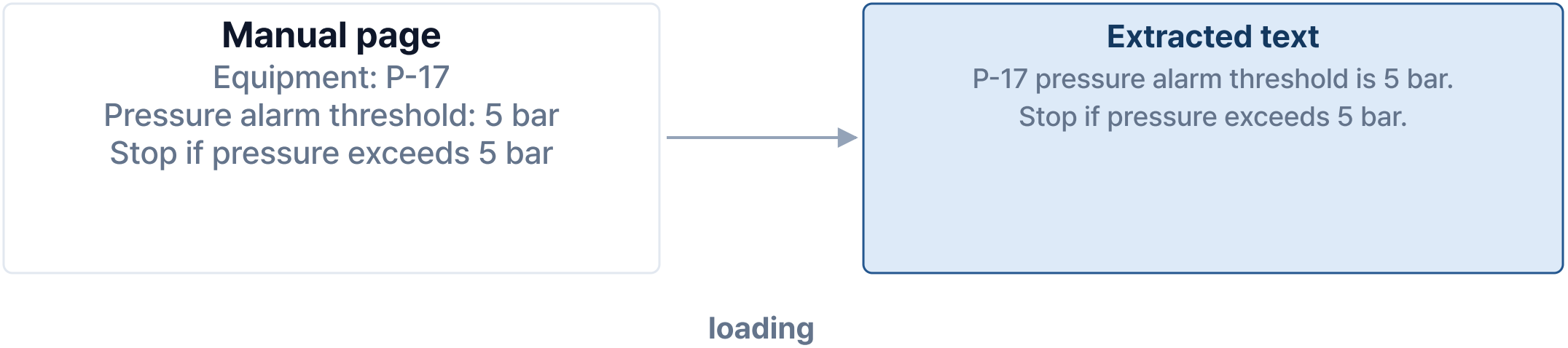
Which value belongs to P-17?

Preserve row membership during extraction

A value without its row can describe the wrong pump

Loading the Manual

Loading extracts readable text from the source



The equipment name, value, and unit survive extraction

Document Pieces

A chunk is a small piece that can be retrieved

p17-v4 / Pressure

Pressure alarm threshold is 5 bar.

Stop if pressure exceeds 5 bar.

Use Celsius for temperature.

split

chunk 0 / p17-v4

5 bar rule + stop condition

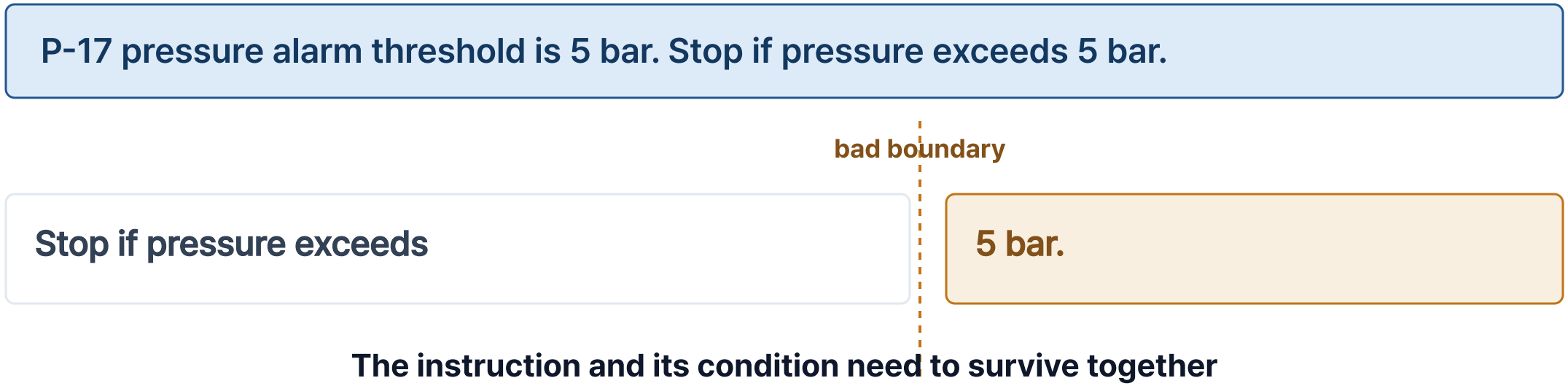
chunk 1 / p17-v4

temperature unit

Each piece keeps a route back to its source

Caveat: Broken Conditions

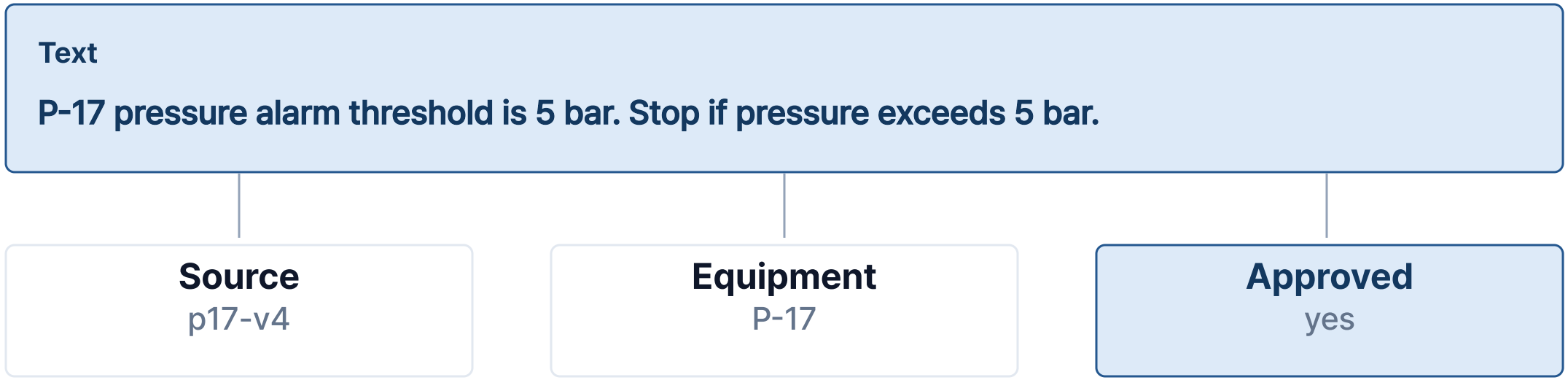
A cut can separate an instruction from its condition



The stop instruction loses its threshold after a bad cut

Source Labels

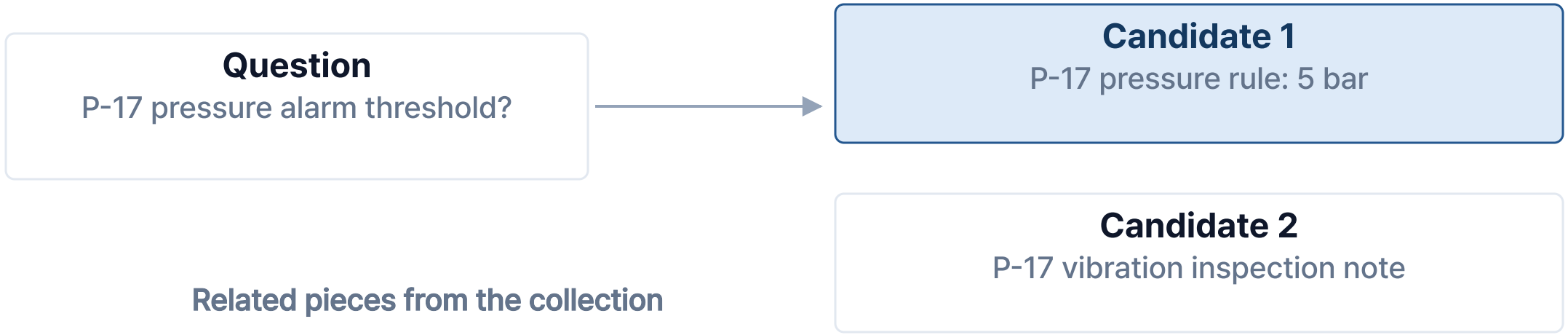
A piece keeps its equipment and version alongside the text



The text and its labels travel together

Retrieval

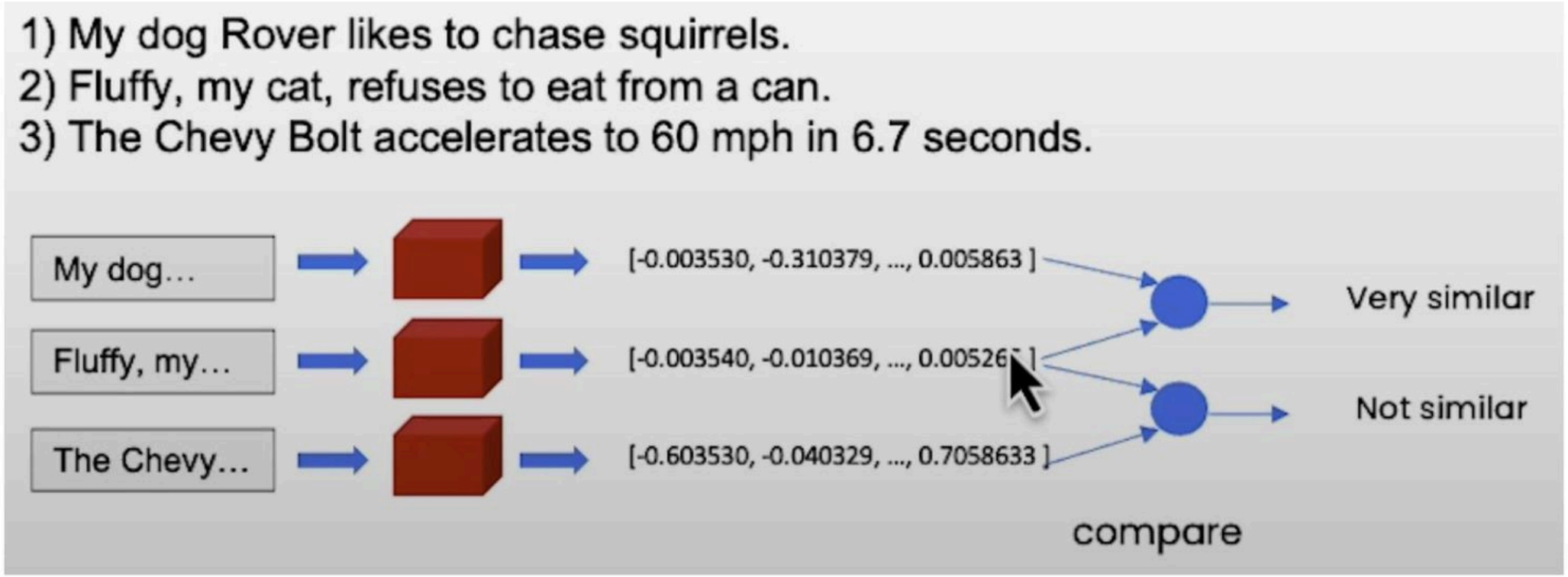
Retrieval finds pieces related to the question



A search returns evidence candidates rather than a finished answer

Embeddings: Text Becomes a Vector

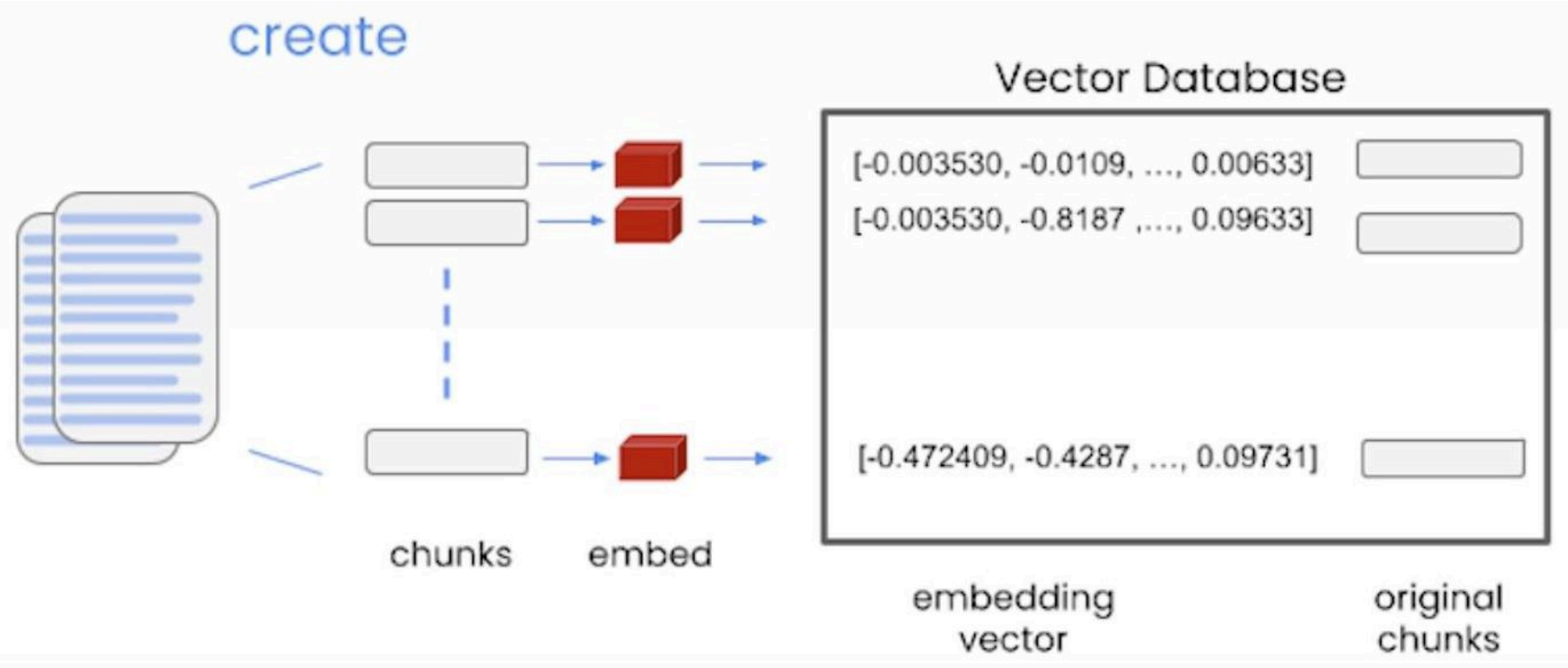
An embedding represents text as numbers that can be compared



Related meanings can have similar vectors even when their words differ

Store the Text Beside Its Vector

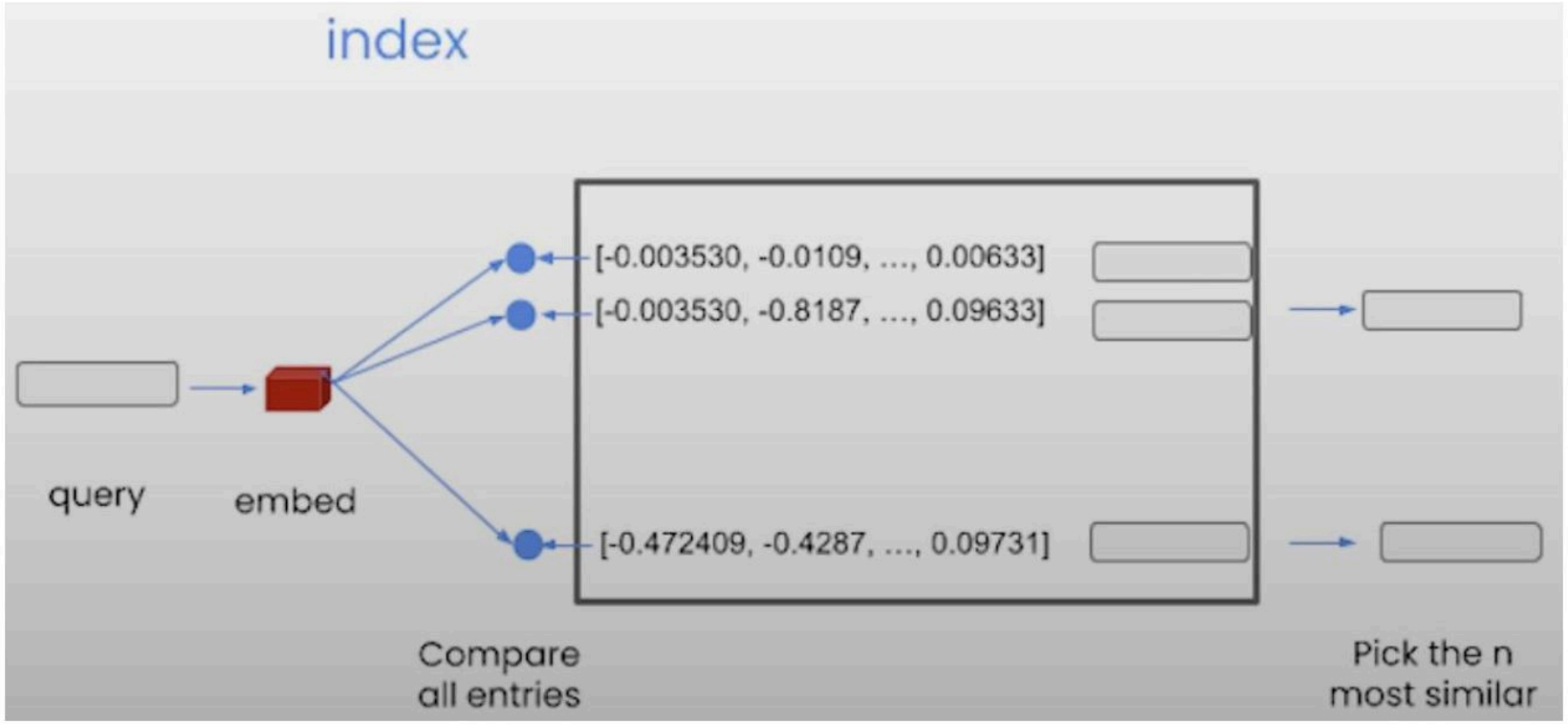
The stored vector helps find a piece while the text remains the evidence



Keep the original text so the answer can quote or cite its source

Search with the Question Vector

Encode the question in the same way, then compare it with stored pieces



The selected vectors lead back to candidate source text

Caveat: Similar but Wrong

Similar wording can belong to an obsolete rule

Candidate	Matched text	Policy check
p17-v2	pressure alarm threshold is 7 bar	obsolete
p17-v4	pressure alarm threshold is 5 bar	approved
p18-v3	pressure alarm threshold is 3 bar	wrong equipment

Both P-17 versions mention the same pressure question

Eligible Evidence

Only approved P-17 sources can support this request

Candidate	Matched text	Policy check
p17-v4	pressure alarm threshold is 5 bar	approved

The current rule remains after the equipment and approval checks

Demo: Find the Manual

Change the request and approval filter, then inspect the returned source text

Request P-17 pressure alarm threshold ☑ Approved only Search Reset

Predict which sources will appear, then select Search.

Source	Returned text	Similarity
?	Search has not run	?

Read the text: does it give the current pressure threshold for this equipment?

Local word-presence search, not a learned embedding or a model answer

Answer: Find the Manual

The baseline returns one pressure rule and one unrelated inspection note

p17-v4

P-17 pressure alarm threshold is 5 bar.
Stop if pressure exceeds 5 bar.

p17-vibration

P-17 vibration inspection: record vibration and temperature before changing the operating speed.

Only p17-v4 answers the pressure threshold question

03 · Answers and Checks

01

Inputs and Evidence

a question, the supplied material, and a candidate answer

02

Documents and Search

prepare readable pieces and find the current rule

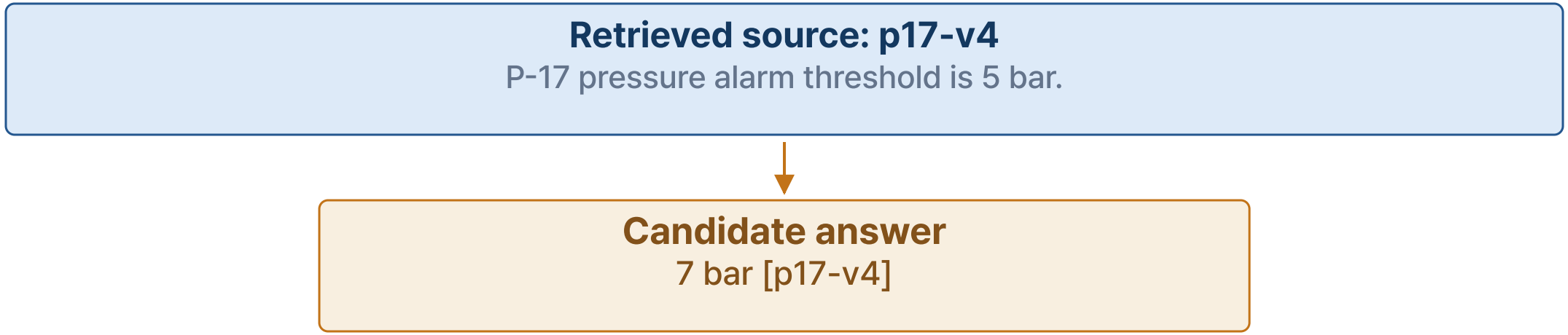
03

Answers and Checks

connect evidence to the answer and check support

Why Answer Checks

Finding the right source does not force the right answer



The source says 5 bar while the candidate says 7 bar

Evidence in the Input

Selected text enters the prompt with its source label

Question

What is the P-17 pressure alarm threshold?

Selected evidence

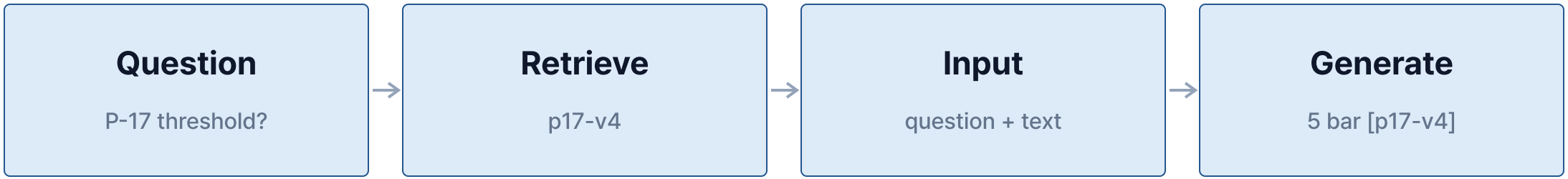
[p17-v4] P-17 pressure alarm threshold is 5 bar.

Stop if pressure exceeds 5 bar.

The 5 bar rule reaches the model as p17-v4

Retrieval-Augmented Generation (RAG)

A 2020 RAG paper connected document retrieval with text generation



Retrieval-augmented generation connects the source to the answer

Checking the Evidence

Does the selected text answer this question for this equipment?

Source	Retrieved text	Support
p17-v4	P-17 pressure threshold is 5 bar	Answers the question
p18-v3	P-18 pressure threshold is 3 bar	Wrong equipment
p17-vibration	Record vibration and temperature	No pressure threshold

Current P-17 pressure text supports the requested value

Checking the Answer

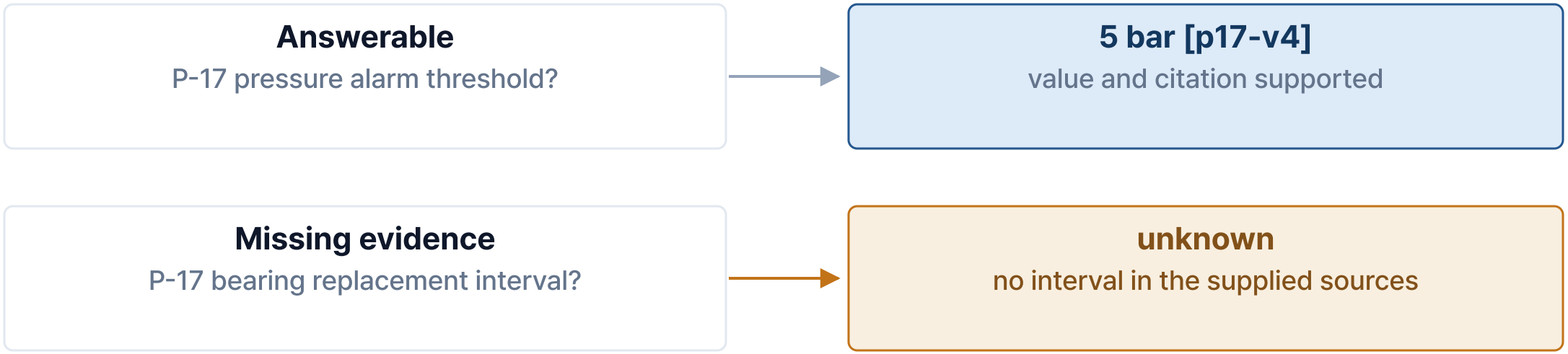
Does the cited text support the value and unit in the answer?

Candidate answer	Cited source	Check
5 bar [p17-v4]	p17-v4 states 5 bar for P-17	Supported
7 bar [p17-v4]	p17-v4 states 5 bar for P-17	Wrong value
5 bar [p18-v3]	p18-v3 states 3 bar for P-18	Wrong source

The citation must support the claim beside it

Answer or Abstain

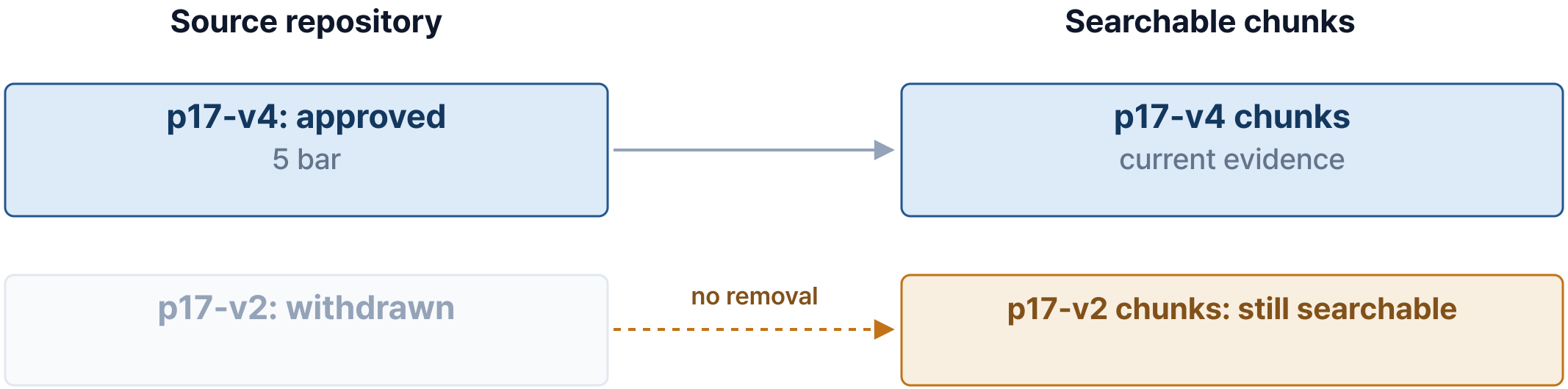
Missing evidence calls for an explicit unknown



No supplied document gives the bearing replacement interval

Case study: A Withdrawn Manual

A document can disappear while its searchable copy remains



A withdrawn source can still leave searchable pieces

A Small Answer Check

Test the same path with a current, wrong-equipment, and missing-evidence question

Question	Available evidence	Expected answer
P-17 pressure threshold?	P-17 approved rule: 5 bar	5 bar [p17-v4]
P-18 pressure threshold?	P-18 approved rule: 3 bar	3 bar [p18-v3]
P-17 bearing interval?	No interval in supplied sources	unknown

One collection exposes the main ways this answer can fail

Takeaway: Answers need evidence that readers can check

1. Put the material in the model input
2. Retrieve complete pieces from the right current source
3. Check the evidence and the answer; say unknown when evidence is missing