Transformers for Threat Intelligence

Reading CTI reports with GLiNER2 and Qwen3: entities, ATT&CK techniques and a grounded summary

Hacklu 2026 - Pauline Bourmeau

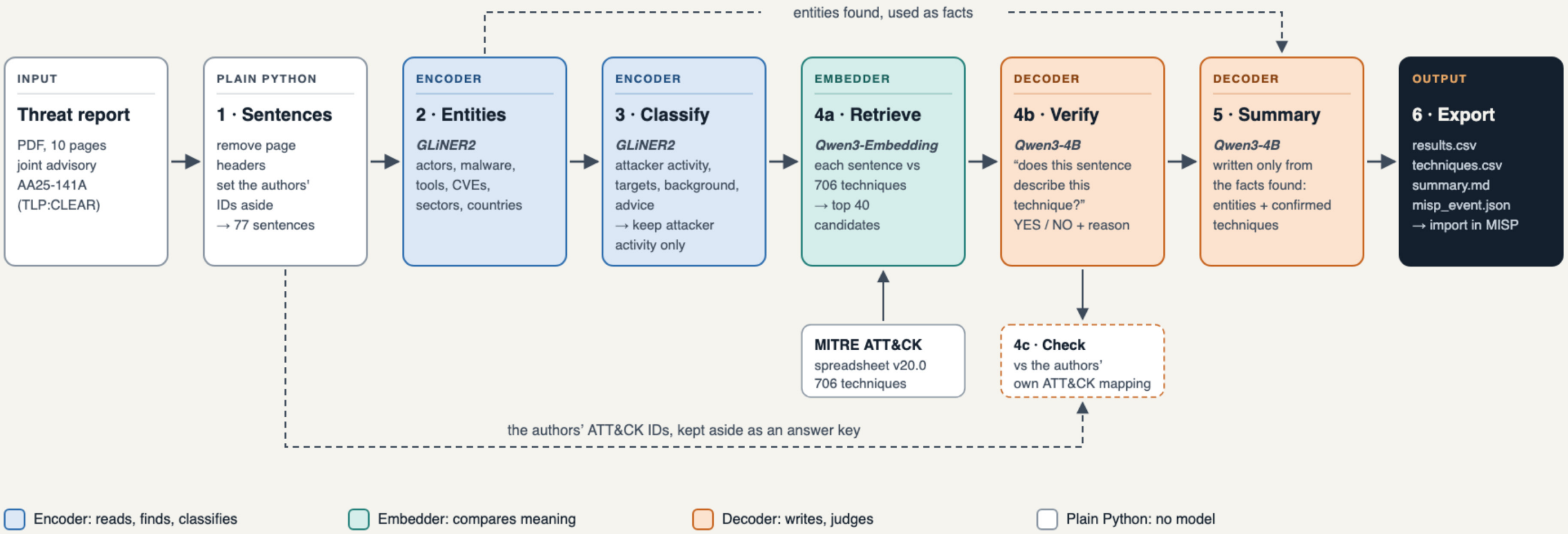
Our plan for the session

1	**Transformers and tokens**	How a model reads text
2	**Our models**	GLiNER2, Qwen3-Embedding, Qwen3-4B
3	**The pipeline**	From report PDF to draft MISP event
4	**Hands-on notebook**	Run on a real report
5	**Review and wrap-up**	

The pipeline, step by step

1	Report PDF → cleaned sentences	Python
2	Find entities: actors, malware, tools, CVEs, sectors, countries	GLiNER2
3	Classify each sentence and keep the attacker-activity ones	GLiNER2
4	Retrieve the 3 closest ATT&CK techniques per sentence	Qwen3-Embedding
5	Accept or reject each candidate technique	Qwen3-4B
6	Write a summary grounded in the extracted information	Qwen3-4B
7	Draft exports, including a MISP event, for review	Analyst

Our pipeline: from a PDF threat report to ATT&CK and MISP



GLiNER2 and Qwen3-4B side by side

Aspect	GLiNER2-large-v1	Qwen3-4B-Instruct-2507
Architecture	Bidirectional transformer encoder	LLM, decoder-style language model
Operation	Picks text spans (segments), predicts labels	Generates an answer token by token
Size	About 340 million parameters	4 billion parameters
Input	A sentence + entity definitions or categories	A sentence + a candidate technique, or facts to summarise
Output	Entity lists and sentence categories	YES/NO with a reason, and a summary
Role	Focused extraction and filtering	Flexible interpretation and writing

From words to numbers: what a transformer reads

Real output of our notebook: the Qwen3 tokenizer and the Qwen3-Embedding model.

STEP 1
Text
what you write

APT28 used Mimikatz to steal credentials.

STEP 2
Tokens
pieces of words

1 rare word = 3 tokens

APT	2	8	·used	·Mim	ik	atz	·to	·steal	·credentials	.

11 tokens
· = a space

STEP 3
Token IDs
each piece has a number

65962	17	23	1483	91846	1579	20175	311	26571	16387	13

STEP 4
Vectors
each ID → a list of numbers



learned during training

STEP 5
One vector per sentence
attention mixes the token vectors, the embedder sums them up into 1,024 numbers per sentence

	similarity with sentence 1	
The attackers dumped credentials from LSASS memory. → …		
Passwords were stolen from the memory of a Windows process. → …	0.74	close meaning
Join our webinar on cyber resilience next Tuesday. → …	0.31	unrelated