



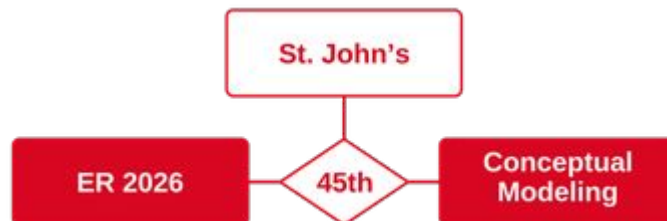
POLYTECHNIQUE
MONTREAL

UNIVERSITÉ
D'INGÉNIERIE

PETALE : a Fine-grained Evaluation Framework for Process Modeling methods

Clara Kowalczyk, Alexis Brissard, Frédéric Cuppens, Amal Zouaq

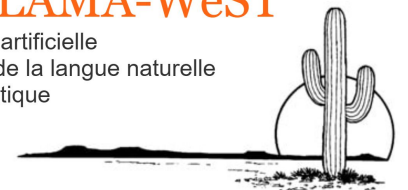
presented at ER2026



Institut
multidisciplinaire
en cybersécurité
et cyberrésilience

Labo LAMA-WeST

Intelligence artificielle
Traitement de la langue naturelle
Web sémantique





Meet the team behind PETALE



Clara Kowalczyk
M.Ing, PhD Student



Alexis Brissard
M.Sc.



Frédéric Cuppens
*Full Professor and
Lab Director of*

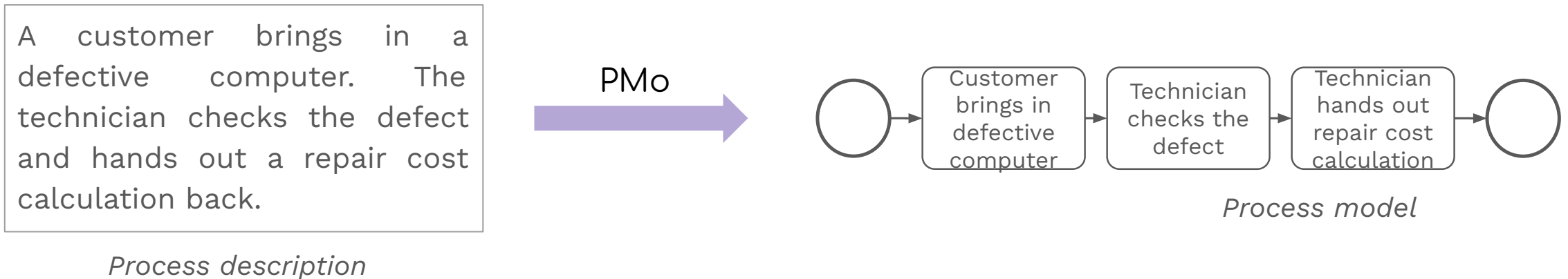


Amal Zouaq
*Full Professor and Lab
Director of*



Running Example

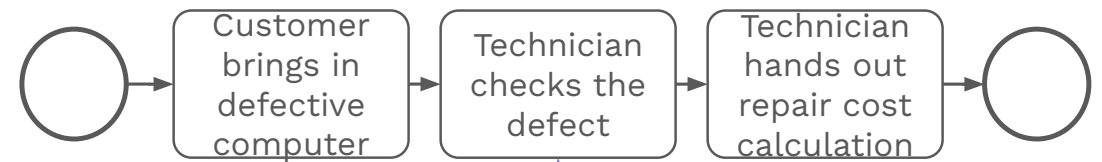
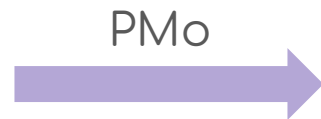
Process Modeling (PMo) : *Generation of a model of the process based on its textual description.*



How to evaluate this ?

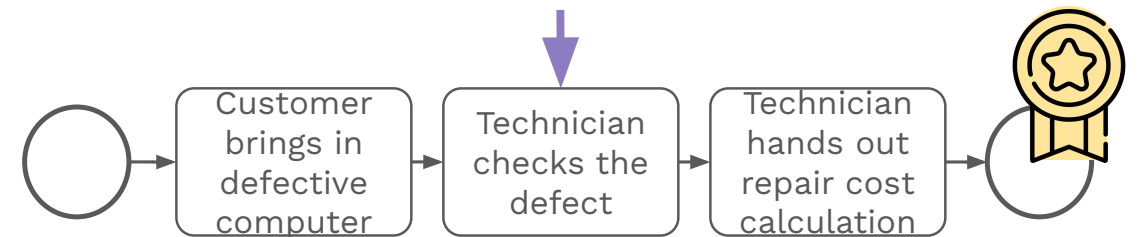
Model-to-Model Evaluation

A customer brings in a defective computer. The technician checks the defect and hands out a repair cost calculation back.



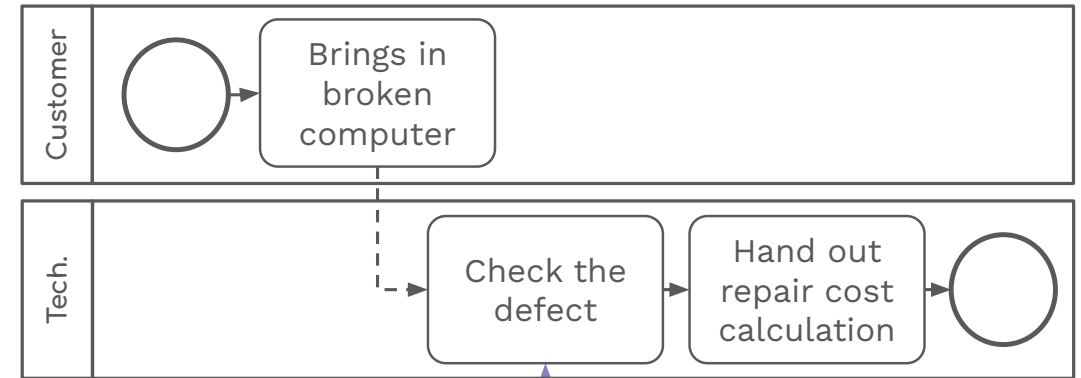
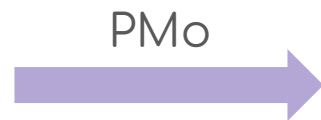
[Dijkman et al, 2011], [Montani et al, 2015], [Klievtsova et al, 2024], [Colona et al, 2024]...

+ Precise scoring



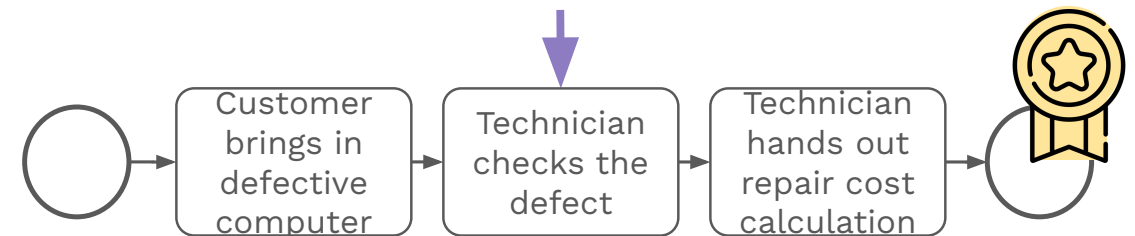
Model-to-Model Evaluation

A customer brings in a defective computer. The technician checks the defect and hands out a repair cost calculation back.

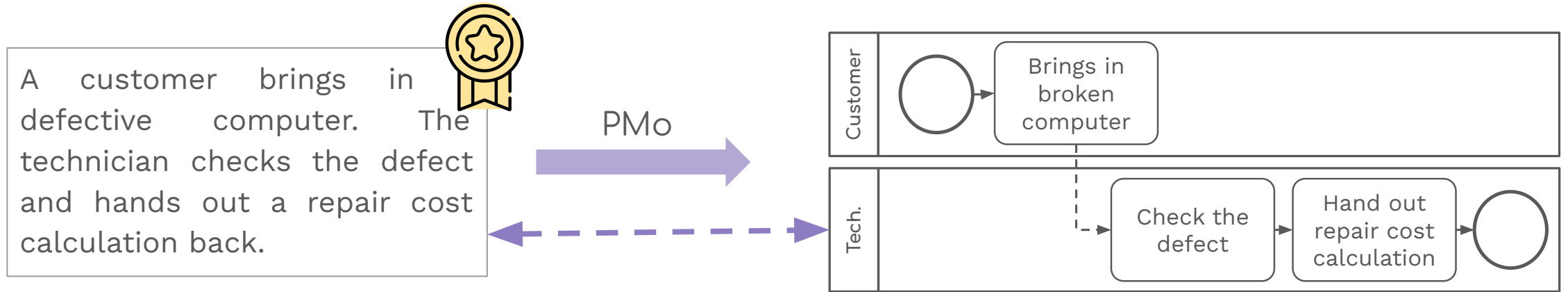


(Dijkman et al, 2011; Montani et al, 2015; Klievtsova et al, 2024; Colona et al, 2024...)

- + Precise scoring
- Gold Process Model is not unique
(Schoknet et al, 2017; Beerepoot et al, 2023)



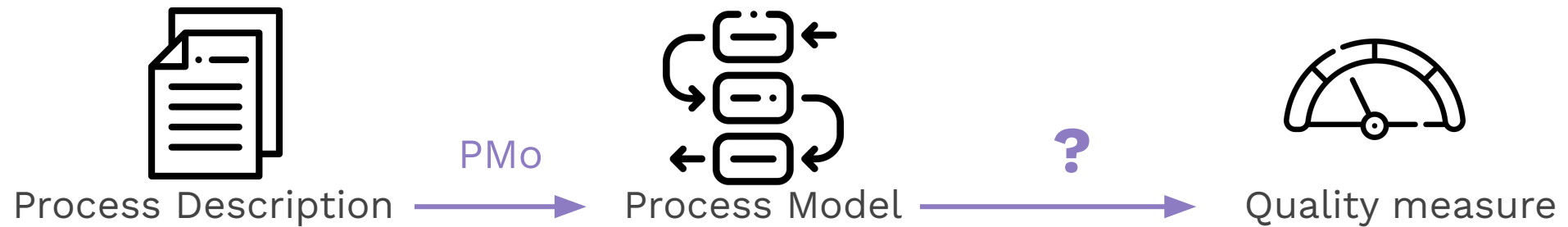
Model-to-Text Evaluation



[SAP Round-tripping], [Berti et al, 2025]

- + Free from gold process **model**
- Holistic score without detailed explanation

Research Questions



Can process modeling approaches be evaluated in a fine-grained manner directly from their description, without relying on a gold standard model?

How do automated process modeling methods compare to human experts when evaluated with PETALE?

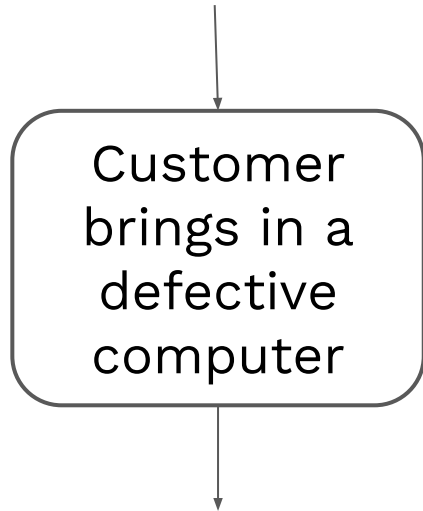
Methodology Outline

Our contributions :

1. **Process Information Task**, a new semantic annotation scheme;
2. **TaskPET**, a validation dataset;
3. **PETALE**, a novel evaluation framework both fine-grained and gold-model free;
4. **Benchmark** of SOTA process modeling methods using PETALE.

PETALE : Fine-grained Evaluation of Process Model

Process Information Task



A Process Information Task describes an elementary step in a process as a single activity with eventual attributes :

Activity : the verb or action being performed

Activity Data : the object, data or system that the activity acts upon

Actor Performer : the person, role, department or organization that executes or initiates the activity

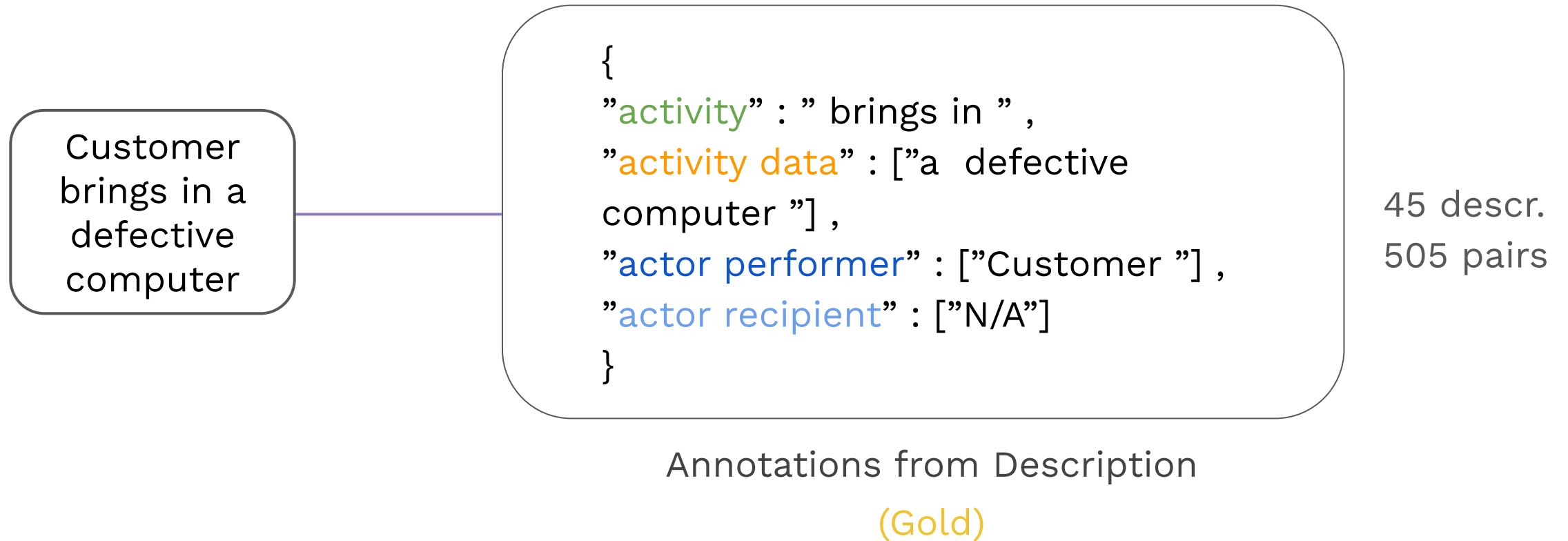
Actor Recipient : the person, role, department or organization, that is affected by the activity or who the activity acts upon

Natural language description of the task is referred to as its verbalization.

Definition inspired by the PET annotation scheme :

Bellan, P., van der Aa, H., Dragoni, M., Ghidini, C., & Ponzetto, S. P. (2022, September). PET: an annotated dataset for process extraction from natural language text tasks. In *International Conference on Business Process Management* (pp. 315-321). Cham: Springer International Publishing.

TaskPET - Process Information Task



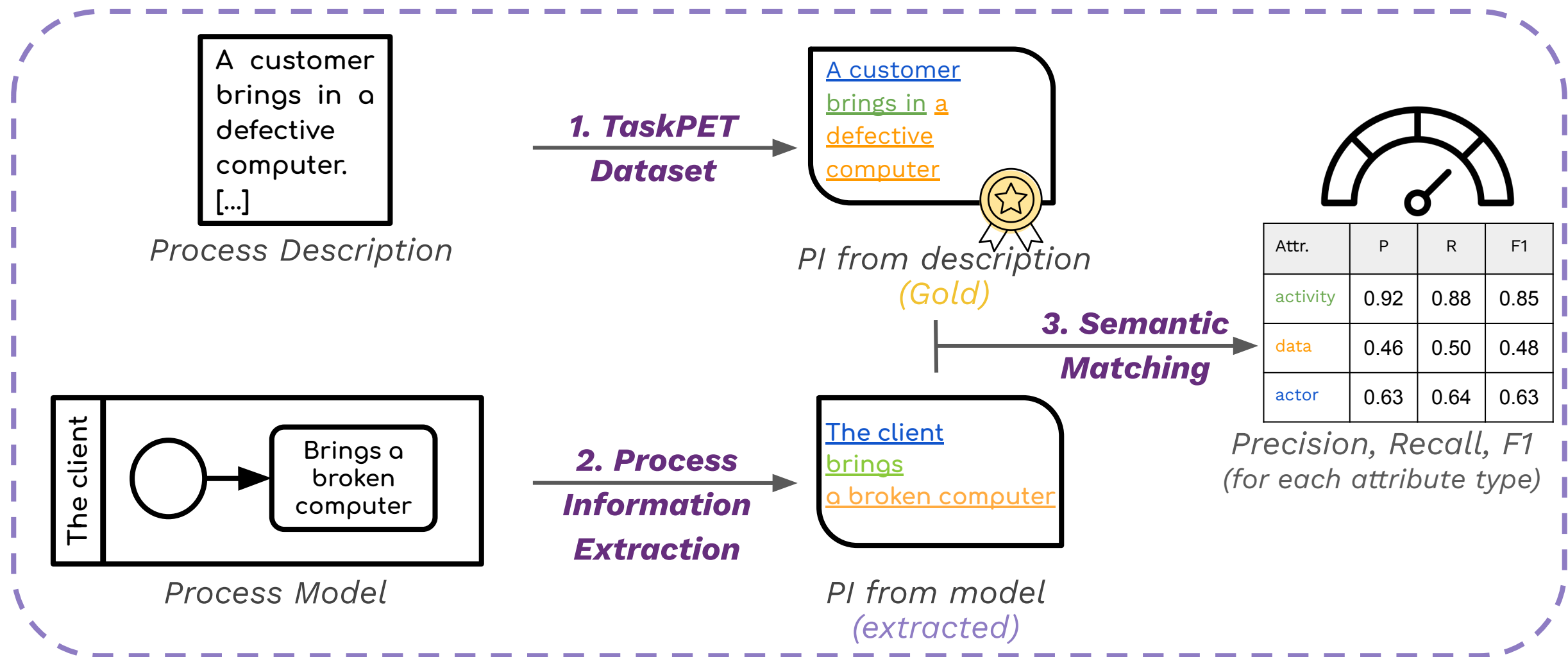
Manual re-annotation to adapt to task format and correct some of the original annotations

Evaluation of Process Models

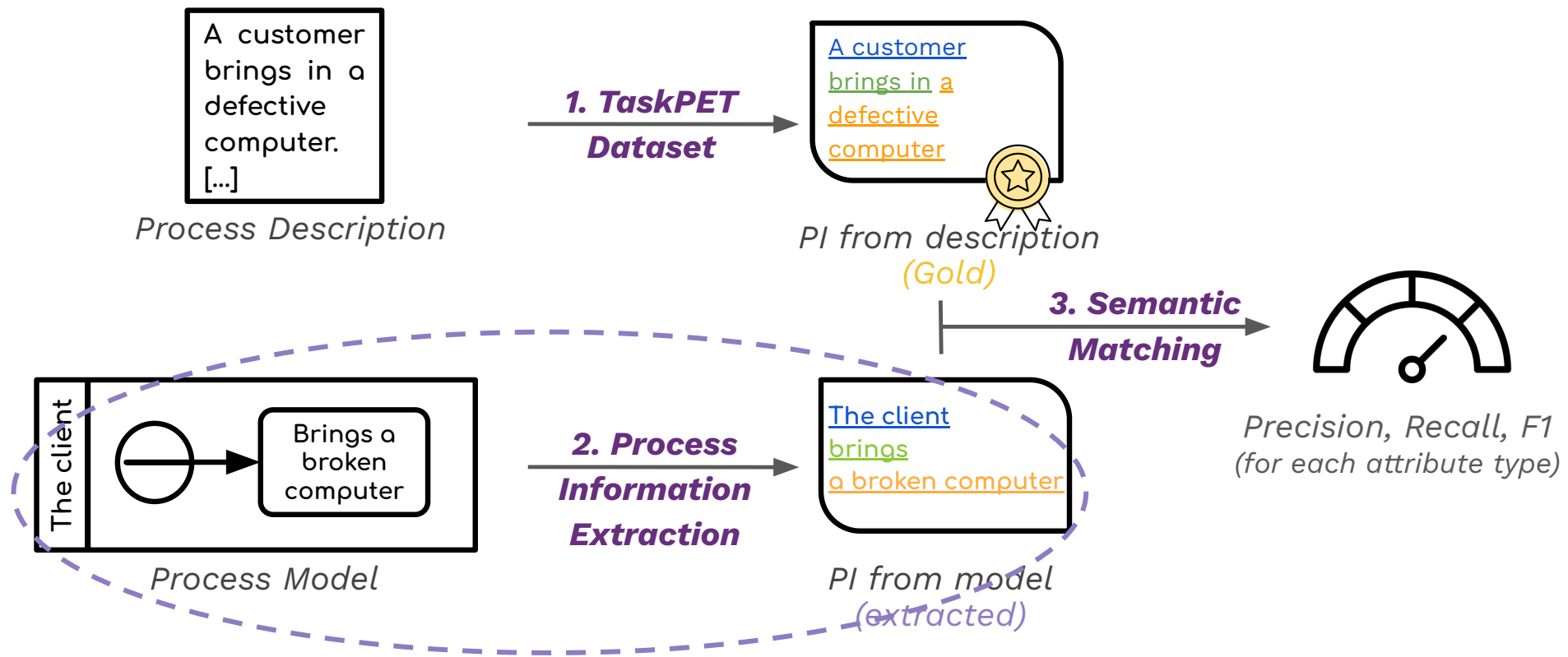
PETALE : **P**rocess model **E**valuation with **T**ask **A**tttribute **L**LM-based **E**xtraction

- Compares the generated model to an annotated description instead of a gold model
- Offers fine-grained evaluation metrics based on the task annotations

PETALE Framework

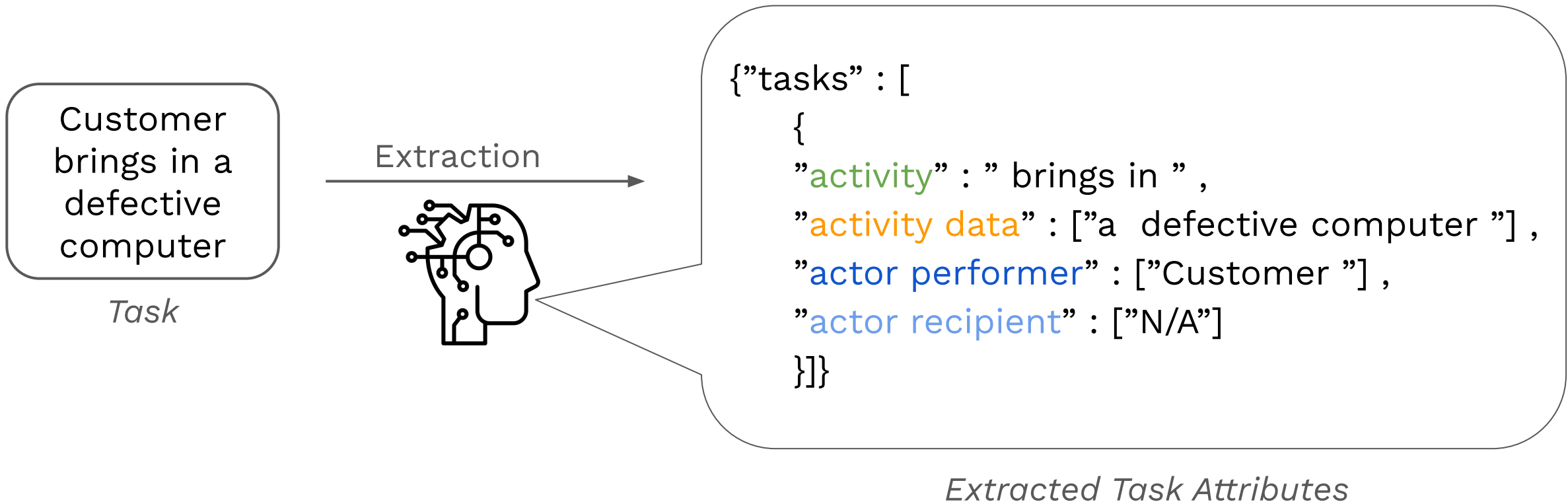


PETALE - Extraction



Task Attribute Extraction

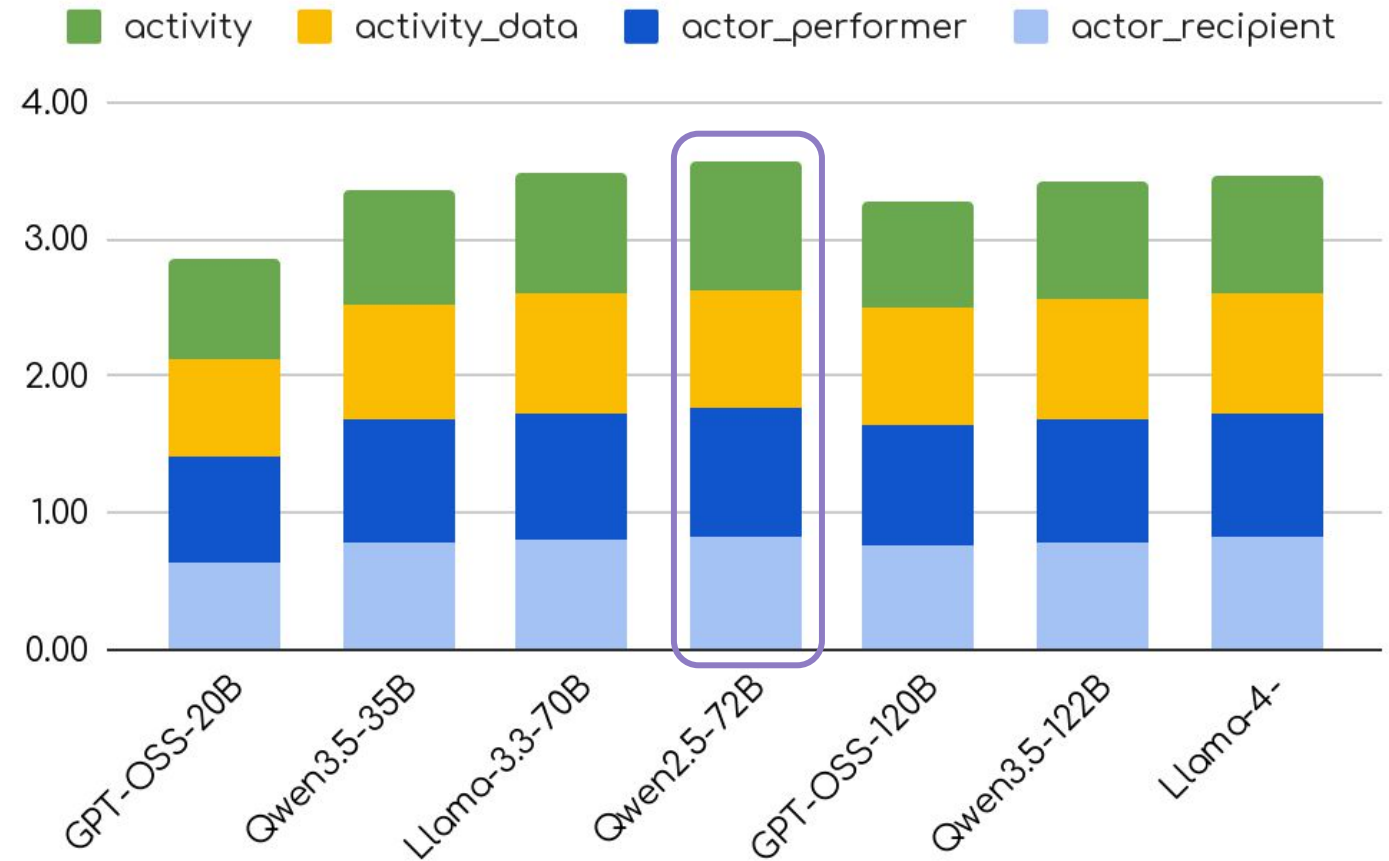
Take a generated process model. For each task label in it :



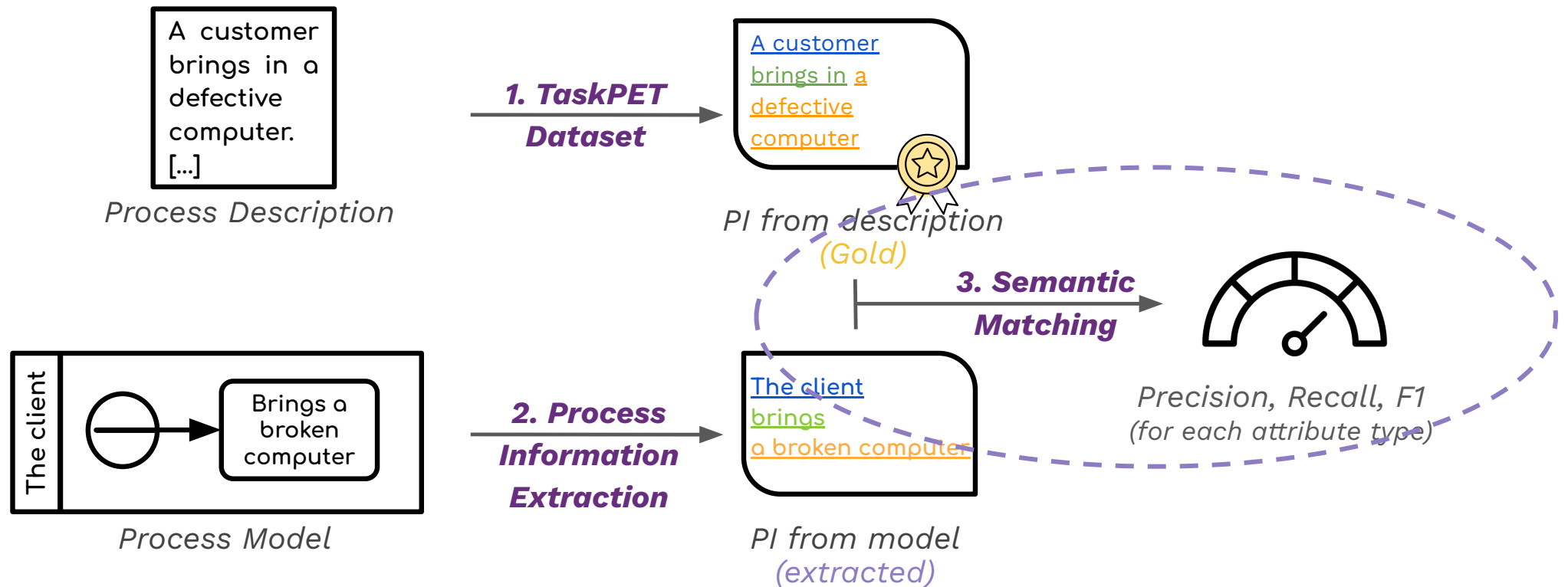
LLM Selection

- Dataset : TaskPET
- Metrics : F1 (per class) computed with Exact Match
- LLM :
 - *Llama-3.3-70B-Instruct*
 - *Llama-4-Scout (16B, 17E)*
 - *Qwen2.5-72B-Instruct*
 - *Qwen3.5-35B-Instruct*
 - *Qwen3.5-122B-Instruct*
 - *GPT-OSS-20B*
 - *GPT-OSS-120B*

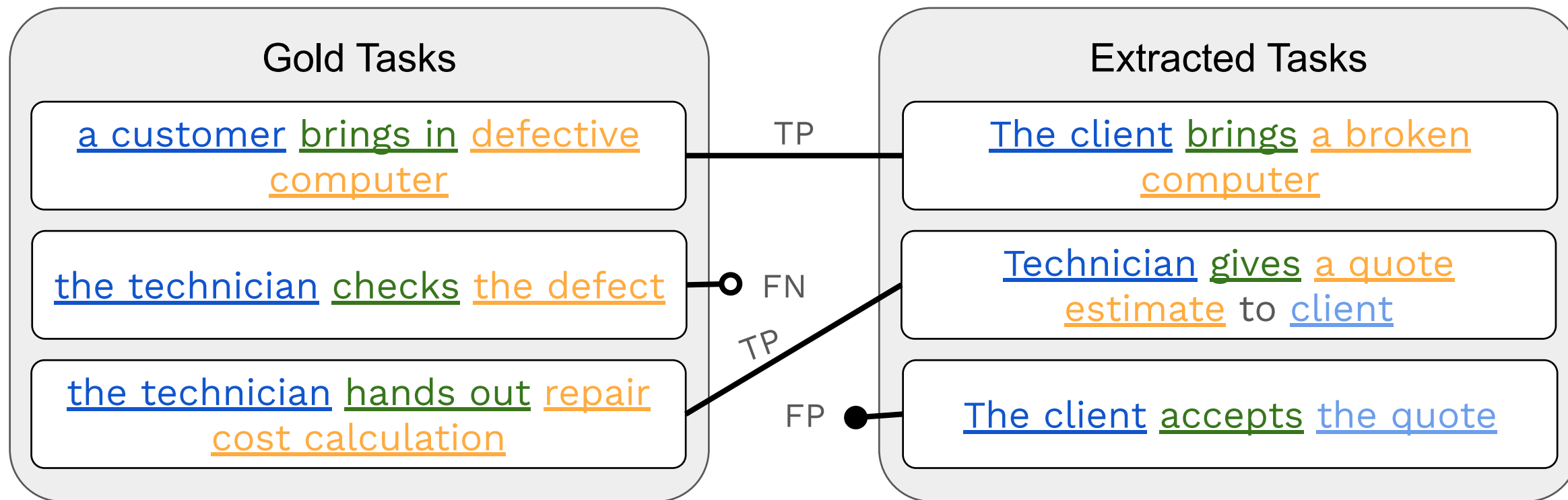
F1 (summed over attributes)



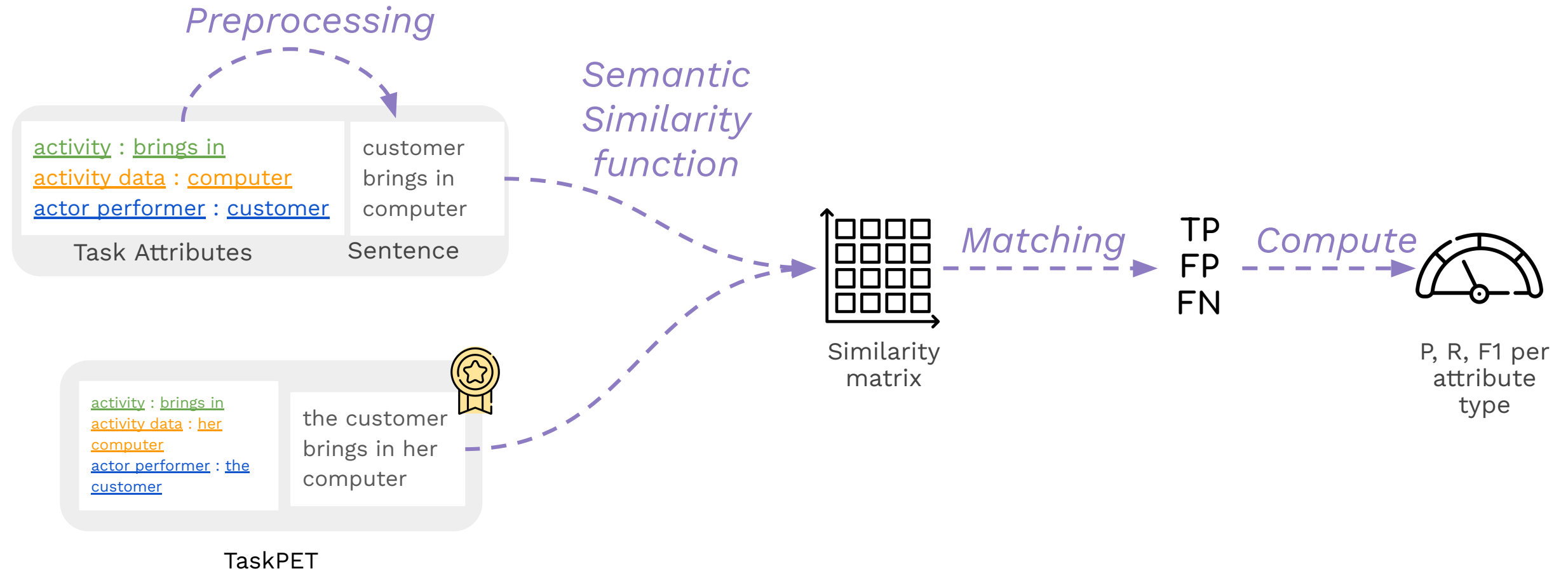
PETALE - Semantic Matching



Task-level Semantic Matching

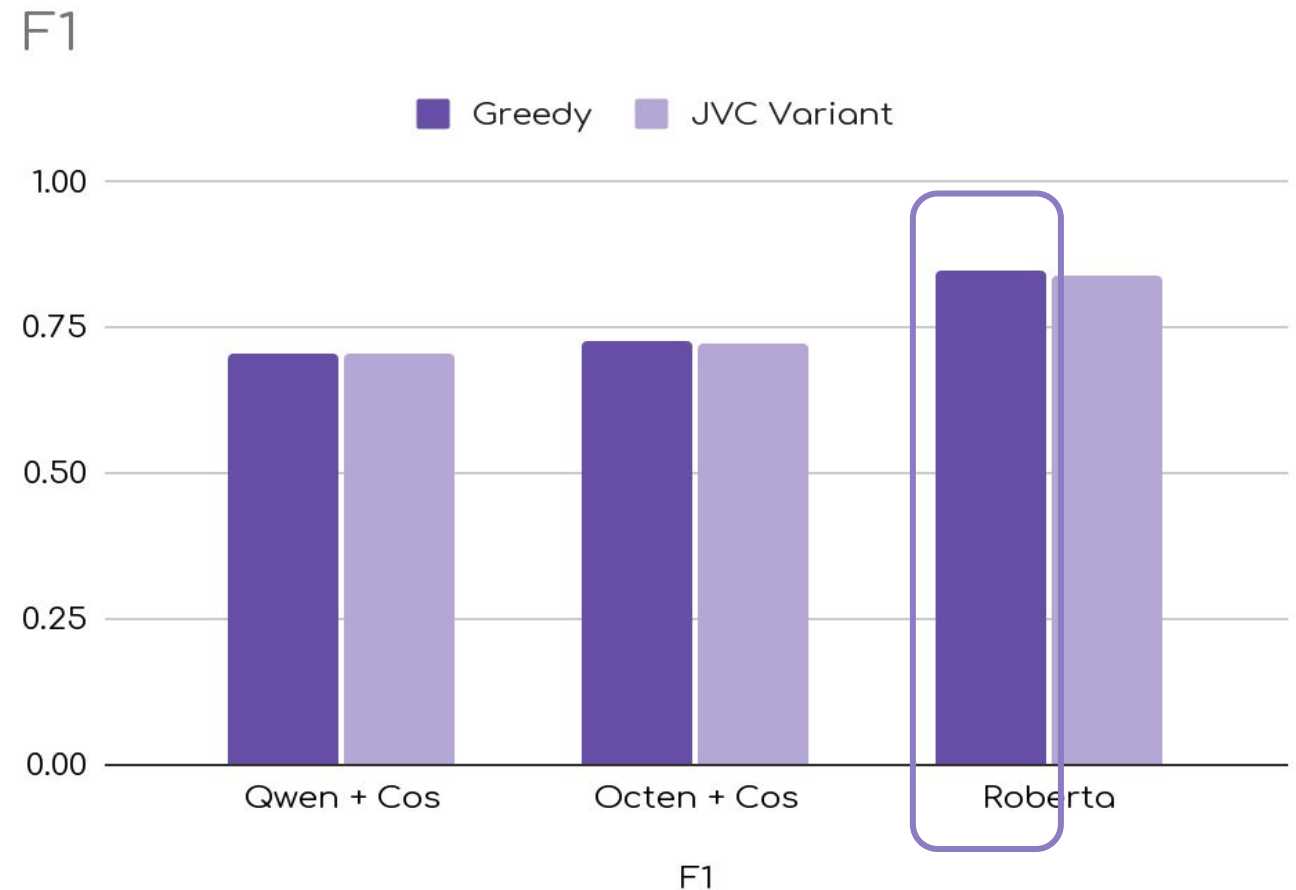


Semantic Matching Methodology

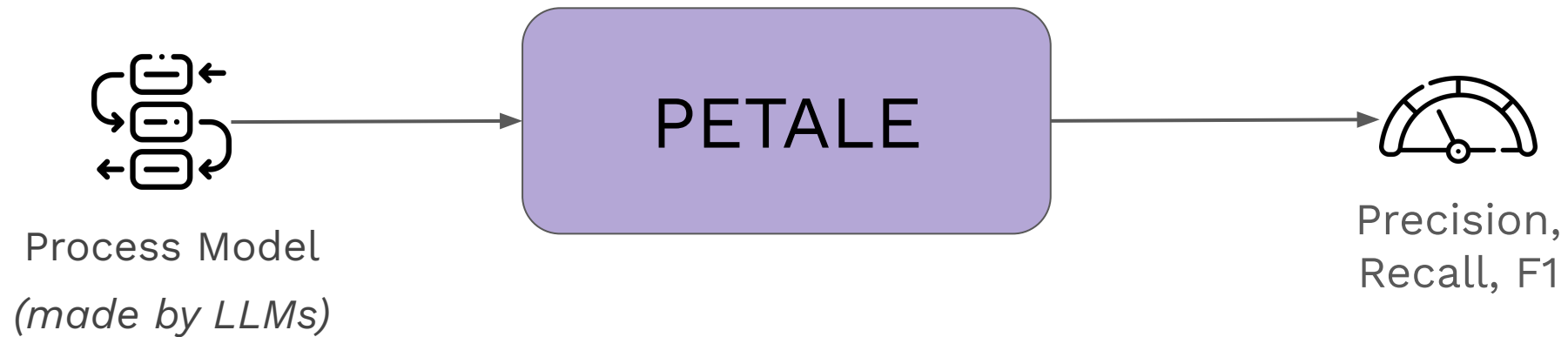


Matching Configuration

- Dataset : TaskPET
- Metrics : F1
- Similarity Function :
 - *Qwen3_8B_Embedder + cos*
 - *Octen3_8B_Embedder + cos*
 - *Roberta Cross Encoder*
- Matching Algorithms:
 - *Greedy [Dijkman et al, 2011]*
 - *JVC [Crouse, 2016]*



Benchmark of Process Modeling methods with PETALE



Experimental Setup

- Dataset : TaskPET (35 processes)
- Metric : F1
- Configuration :
 - Extraction : LLM : *Qwen2.5-72B-Instruct in 7-shots*
 - Matching : *Roberta Cross Encoder + Greedy matching*
- Process Modeling Approaches :
 - Experts [Costa et al., 2025] (human baseline)
 - ProMoAI [Kourani et al., 2024] (automated baseline) with gpt-4o-mini
 - AutoBPMN.AI [Klievtsova et al., 2025] with gpt-4.1
 - BPMNGen [Hörner et al., 2026] with gpt-4.1
 - BPMN Assistant [Licardo et al., 2026] with gpt-4o
 - BPMN Chatbot++ [Safan & Köpke, 2025] with gpt-4.1

Benchmark Results

Table 2. F1-scores for various PMo approaches.

	Experts (Human Baseline)	ProMoAI (Automated Baseline)	Auto BPMN AI	BPMN Gen	BPMN Assistant	BPMN Chatbot ++
Generated Process Models	35/35	25/35	35/35	33/35	35/35	35/35
Task-level	0.89	0.89	0.92	0.84	0.88	0.71
Activity	0.86	0.7	0.92	0.84	0.88	0.69
Activity Data	0.67	0.71	0.79	0.69	0.76	0.63
Actor Performer	0.74	0.16	0.19	0.68	0.19	0.59
Actor Recipient	0.51	0.55	0.58	0.48	0.62	0.47
Attribute Macro F1	0.7	0.53	0.62	0.67	0.61	0.6

Benchmark takeaways

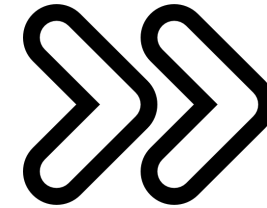
	Experts (Human Baseline)	BPMN Gen
Generated Process Models	35/35	33/35
Task-level	0.89	0.84
Activity	0.86	0.84
Activity Data	0.67	0.69
Actor Performer	0.74	0.68
Actor Recipient	0.51	0.48
Attribute Macro F1	0.7	0.67

- BPMNGen is the strongest automated process modeling approach, with near human performances.
- Automated process modeling methods miss out on process details, such as actors or involved data.
- Even human experts don't model all process informations.

Conclusion

Limitations

- Order of element is not taken into account
- Matching sensitive to duplicated information



Key takeaways

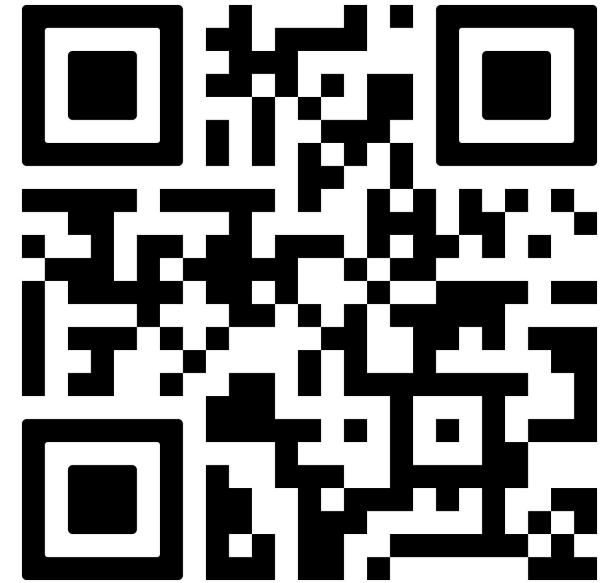
- Independant from gold process model
- Fine-grained evaluation method
- Automated process modeling methods miss out on process details, such a actors or involved data





PETALE : a Fine-grained Evaluation Framework for Process Modeling methods

Clara Kowalczyk, Alexis Brissard,
Frédéric Cuppens, Amal Zouaq



clara.kowalczyk@polymtl.ca



References (1/2)

- [Klievtsova et al., 2024] Klievtsova, N., Benzin, J.V., Kampik, T., Mangler, J., Rinderle-Ma, S.: Conversational Process Modeling: Can Generative AI Empower Domain Experts in Creating and Redesigning Process Models? (2024). <https://doi.org/10.48550/arXiv.2304.11065>
- [Montani et al., 2015] Montani, S., Leonardi, G., Quaglini, S., Cavallini, A., Micieli, G.: A knowledge-intensive approach to process similarity calculation. *Expert Syst. with Appl.* 42(9), 4207–4215 (2015). <https://doi.org/10.1016/j.eswa.2015.01.027>
- [Dijkman et al., 2011] Dijkman, R., Dumas, M., van Dongen, B., Käärik, R., Mendling, J.: Similarity of business process models: Metrics and evaluation. *Information Systems* 36(2), 498–516 (2011). <https://doi.org/10.1016/j.is.2010.09.006>
- [Colona et al., 2024] Colonna, J.G., Fares, A.A., Duarte, M., Sousa, R.: Process mining embeddings: Learning vector representations for Petri nets. *Intell. Syst. with Appl.* 23, 200423 (2024). <https://doi.org/10.1016/j.iswa.2024.200423>
- [Crouse, 2016] Crouse, D.F.: On implementing 2d rectangular assignment algorithms. *IEEE Transactions on Aerospace and Electronic Systems* 52(4), 1679–1696 (2016). <https://doi.org/10.1109/TAES.2016.140952>
- [Berti et al., 2025] Berti, A., Wang, X., Kourani, H., Van der Aalst, W.M.: Specializing large language models for process modeling via reinforcement learning with verifiable and universal rewards. *Process Science* 2(1), 26 (2025). <https://doi.org/10.1007/s44311-025-00034-4>
- [SAP] SAP, S.: <https://github.com/SAP-archive/llm-round-trip-correctness> SAP Samples (2025)

References (2/2)

[Costa et al., 2025] Costa, A., Wimmer, A., Pufahl, L.: LLM4BPMNGen: A Tool for BPMN Generation with LLMs. In: Marcel, P., Polacsek, T., Karlapalem, K., Jean, S., Wang, H. et al. (eds.) Companion Proceedings of the 44th International Conference on Conceptual Modeling. CEUR Workshop Proceedings, vol. 4099, pp. 333–337. CEUR (2025)

[Kourani et al., 2024] Kourani, H., Berti, A., Schuster, D., van der Aalst, W.M.P.: ProMoAI: Process Modeling with Generative AI (2024). <https://doi.org/10.48550/arXiv.2403.04327>

[Klievtsova et al., 2025] Klievtsova, N., Ehrendorfer, M., Mangler, J., Rinderle-Ma, S.: [AutoBPMN.AI](#): Conversational Process Modeling and Automation. In: Reijers, H., Marrella, A., del Río Ortega, A., Rinderle-Ma, S., Depaire, B. et al. (eds.) Joint Proceedings of the Best Dissertation Award, Doctoral Consortium, and Demonstration & Resources Forum at BPM 2025. CEUR Workshop Proceedings, vol. 4032, pp. 304–311. CEUR-WS.org (2025), <https://ceur-ws.org/Vol-4032/paper-40.pdf>

[Hörner et al., 2026] Hörner, L.F., Möller, M., Reichert, M.: Automatically Generating BPMN 2.0 Process Models from Natural Language Process Descriptions: Challenges, Framework, Quality Assessment. Bus. & Inf. Syst. Eng. (2026). <https://doi.org/10.1007/s12599-025-00983-x>

[Licardo et al., 2026] Licardo, J.T., Tanković, N., Etinger, D.: Bpmn assistant: An llm-based approach to business process modeling. Appl. Sci. 16(5) (2026), <https://www.mdpi.com/2076-3417/16/5/2213>

[Safan & Köpke, 2025] Safan, A., Köpke, J.: BPMN-Chatbot++: LLM-Based Modeling of Collaboration Diagrams with Data. In: Proceedings of the Best BPM Dissertation Award, Doctoral Consortium, and Demonstrations & Resources Forum co-located with the 23rd International Conference on Business Process Management (BPM 2025). CEUR Workshop Proceedings, vol. 4032. CEUR-WS.org (2025), <https://ceur-ws.org/Vol-4032/>