👑 TIARA: Multi-grained Retrieval for Robust Question Answering over Large Knowledge Base

Yiheng Shu, Zhiwei Yu, Yuhan Li, Börje F. Karlsson, Tingting Ma, Yuzhong Qu, Chin-Yew Lin

南京大學 NANJING UNIVERSITY

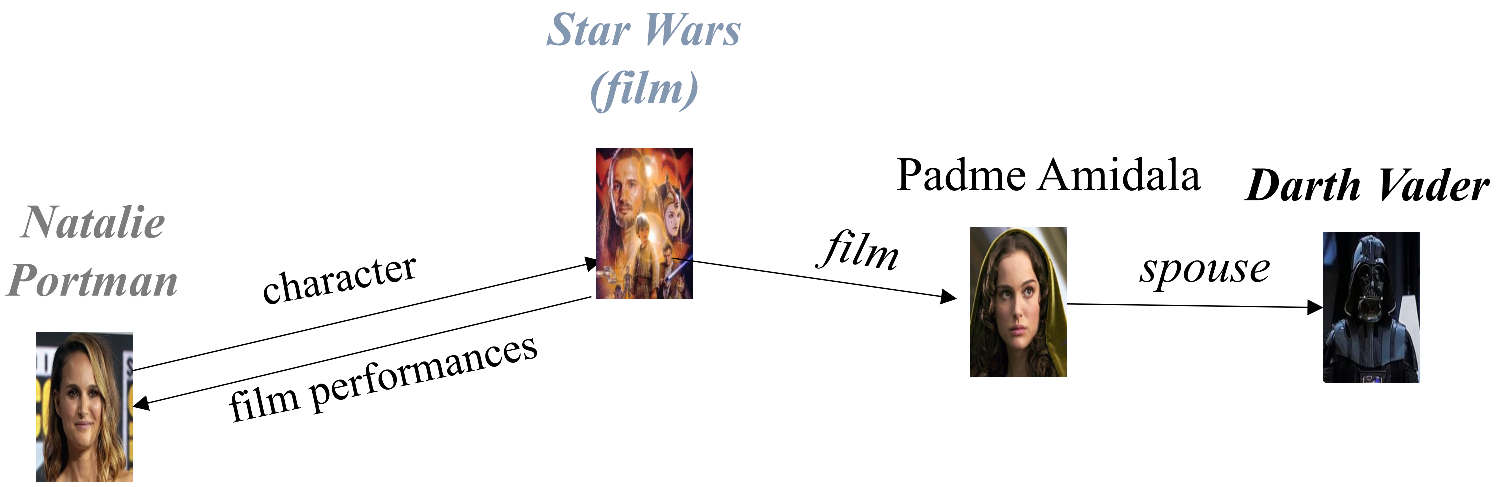
Microsoft

南開大學 Nankai University

哈爾濱工業大學 HARBIN INSTITUTE OF TECHNOLOGY

Knowledge Base Question Answering

who is *the husband of* the character natalie portman play in star wars?

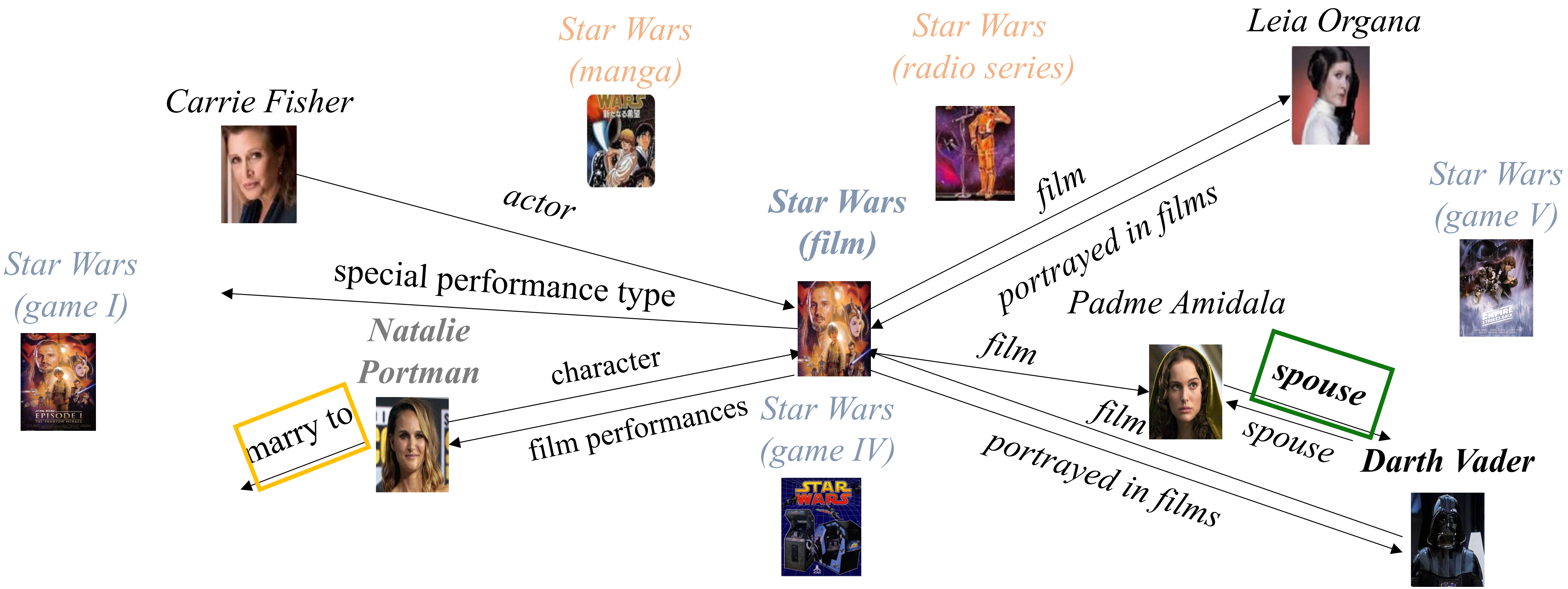


Search engine based on text vs. Reasoning over knowledge base

- Strong interpretability
- Abundant curated data
- Multi-hop and numerical reasoning

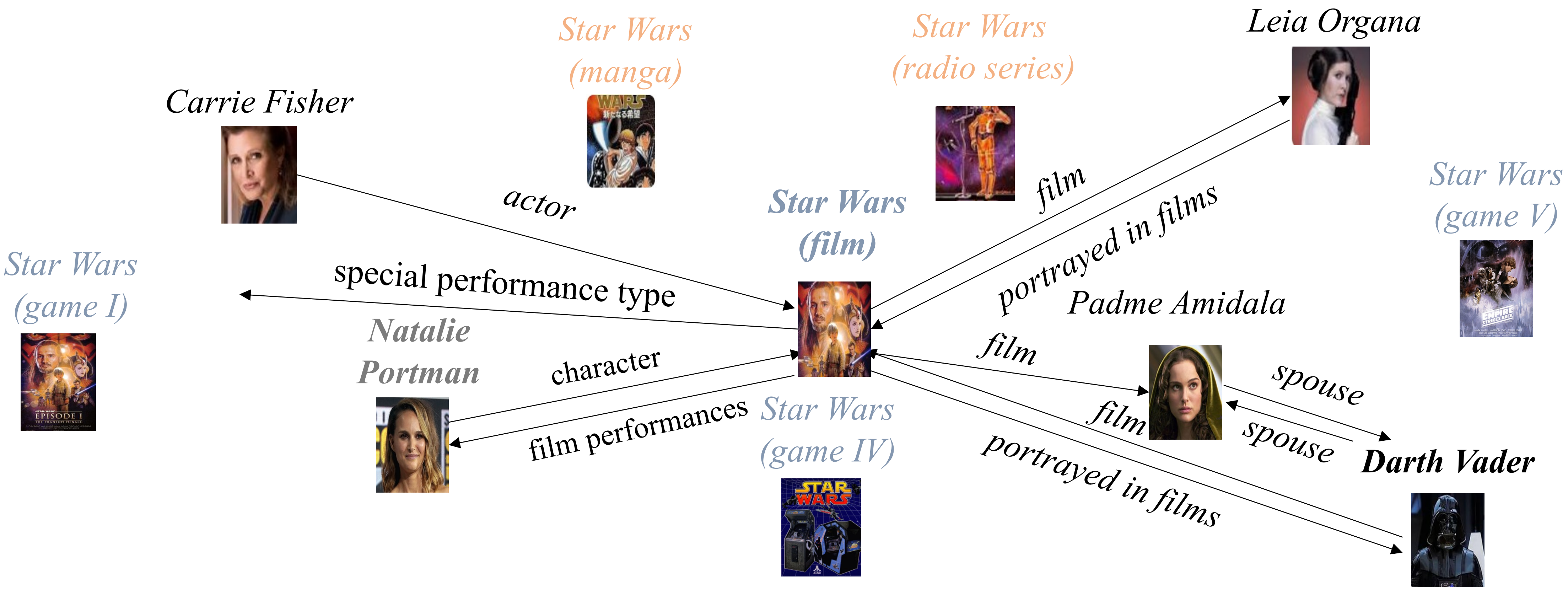
Challenges - Question Understanding

who is ***the husband of*** the character natalie portman play in star wars?



Challenges – Large Search Space

who is *the husband of* the character natalie portman play in star wars?



120 M+ entities, 45 M+ in 86 common domains (e.g., ***Star wars, Natalie Portman*** …)

30 K+ schema items (e.g., *film performance, portrayed in films* …)

Challenges - Compositional & Zero-shot Generalization

- Corpus

what character did daniel naprous play in game of thrones?

who is the husband of leia organa?

- I.I.D. *independent and identically distributed*

what character did carrie fisher play in shampoo?

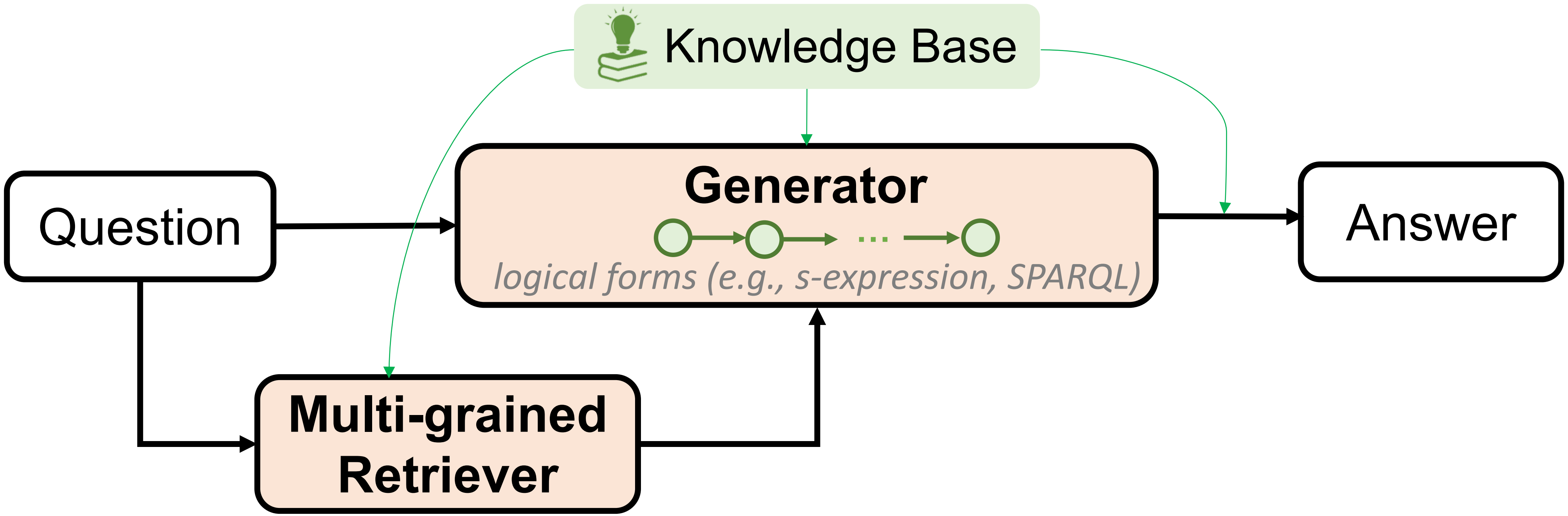
- **Compositional** *novel compositions of schema items seen in training*

who is the husband of the character natalie portman play in star wars?

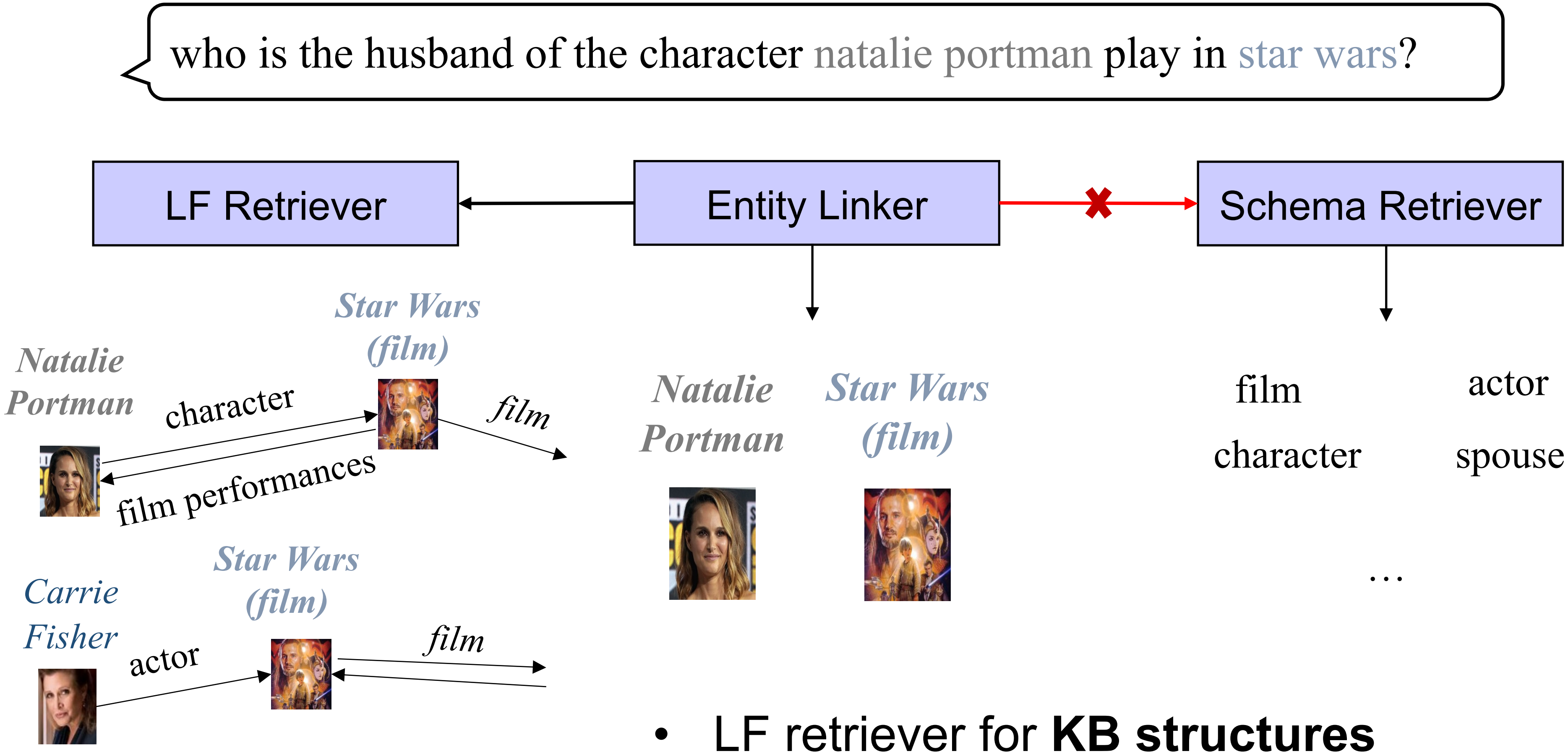
- **Zero-shot** *unseen schema items, even from different domains*

the latest released version of the game nicalis was published in what region?

Logical Form Generation with Multi-grained Retrieval

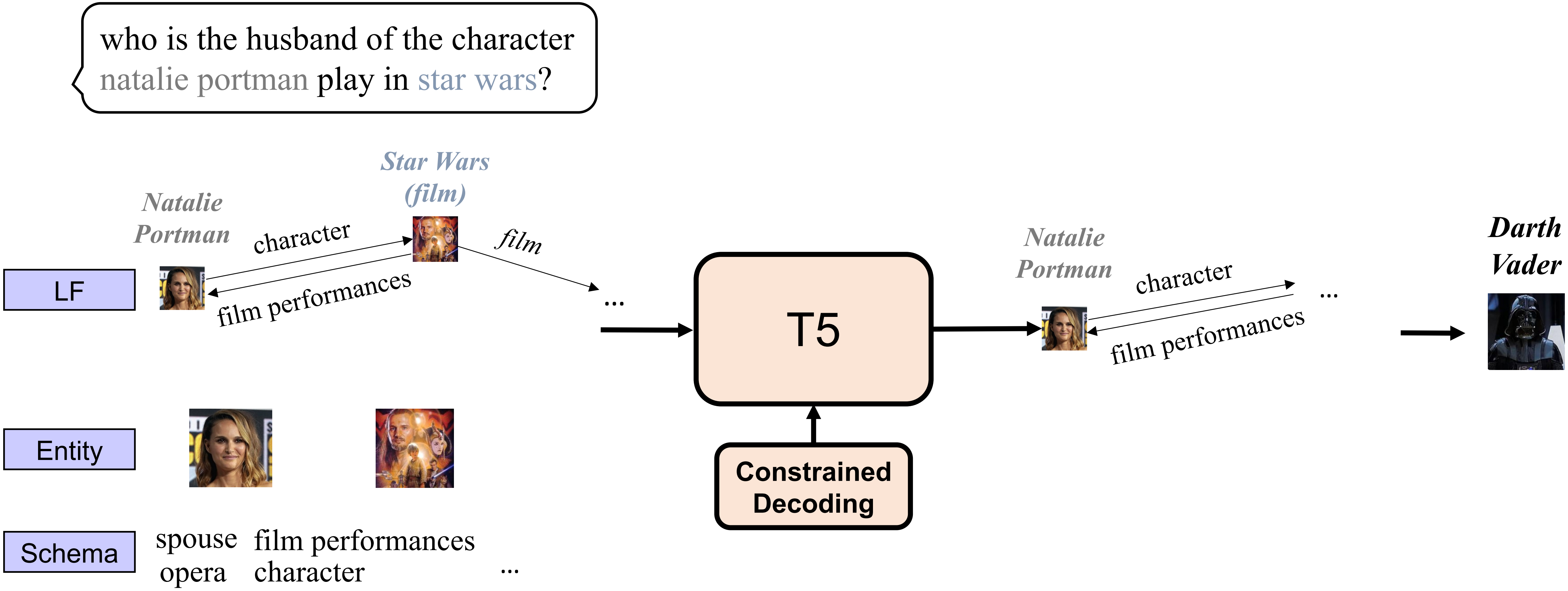


Multi-grained Retriever



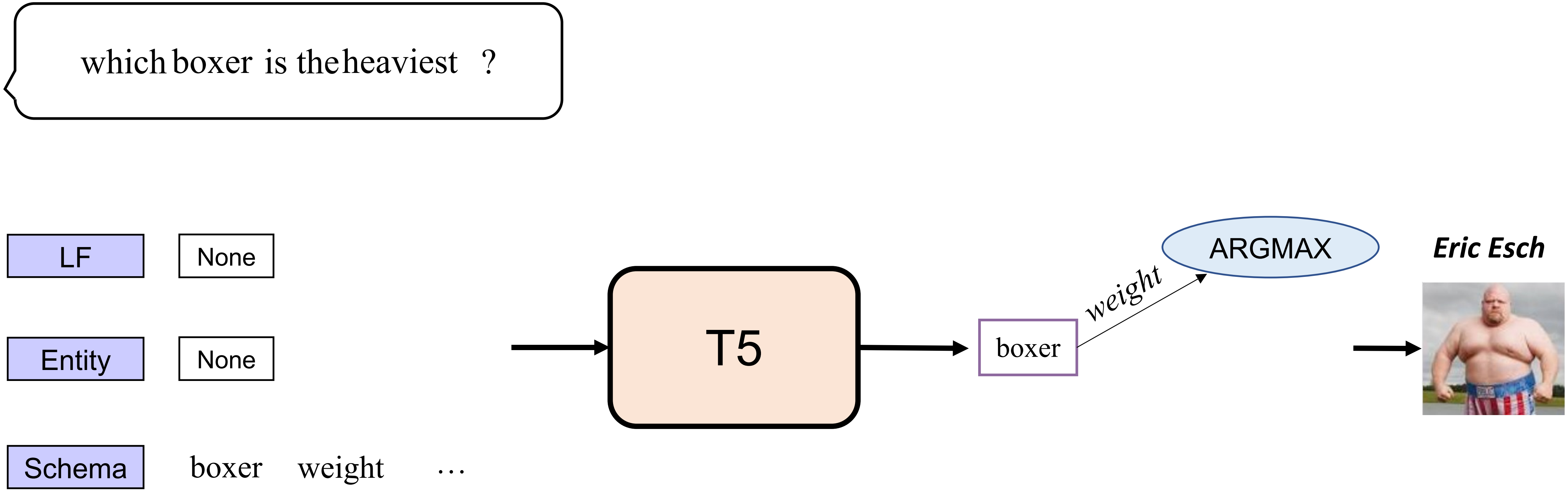
- LF retriever for **KB structures**
- Decouple the entity linker and schema retriever for **semantic supplement**
- Entities + Top-5 LFs + Top-10 classes/relations

Generation with Checking



- Leverage the generation power of **PLM**
- Checking the **schema correctness** and **executability**

Question without Entity



- Inference on questions without entities
- Compositional schemas

Evaluation Benchmark

GrailQA (OSU, Stanford etc.)

- 64,331 questions annotated with **high quality** involving up to 4 relationships
- Three-level generalization settings
  - ➢ **I.I.D.**
    Test cases following the training distribution
  - ➢ **Compositional**
    Novel compositions in test cases
  - ➢ **Zero-shot**
    Unseen schema items in test cases

Leaderboard: Overall

Here are the overall Exact Match (EM) and F1 scores evaluated on GrailQA test set. To get the EM score on GrailQA, please submit your results with logical forms in S-expression. Note that, submissions are ranked only based on F1, so feel free to choose your own meaning representation as EM won't affect your ranking.

Rank	Model	EM	F1
1 May 18, 2022	Tiara-QA (single model) *Anonymous*	**72.081**	**77.491**
2 Aug 20, 2021	RnG-KBQA (single model) *Salesforce Research* https://arxiv.org/abs/2109.08678	68.778	74.422
3 Apr 19, 2022	ArcaneQA V2 (single model) *Anonymous*	63.774	73.713
4 Aug 12, 2021	S2QL (single model) *Anonymous*	57.456	66.186
5 Apr 05, 2021	ReTraCk (single model) *Microsoft Research Asia* https://aclanthology.org/2021.acl-demo.39/	58.136	65.285
6 Feb 04, 2021	ArcaneQA V1 (single model) *Anonymous*	57.872	64.924
7 Jan 22, 2021	BERT+Ranking (single model) *The Ohio State University*	50.578	57.988
8 Jan 22, 2021	GloVe+Ranking (single model) *The Ohio State University*	39.521	45.136
9 Jan 22, 2021	BERT+Transduction (single model) *The Ohio State University*	33.255	36.803
10 Jan 22, 2021	GloVe+Transduction (single model) *The Ohio State University*	17.587	18.432

Top1 on GrailQA Leaderboard on May. 2022

Overall and Generalization Performance

	Overall		I.I.D.		Compositional		Zero-shot	
Method	**EM**	**F1**	**EM**	**F1**	**EM**	**F1**	**EM**	**F1**
GloVe + Transduction (Gu et al., 2021)	17.6	18.4	50.5	51.6	16.4	18.5	3.0	3.1
QGG (Lan and Jiang, 2020)	-	36.7	-	40.5	-	33.0	-	36.6
BERT + Transduction (Gu et al., 2021)	33.3	36.8	51.8	53.9	31.0	36.0	25.7	29.3
GloVe + Ranking (Gu et al., 2021)	39.5	45.1	62.2	67.3	40.0	47.8	28.9	33.8
BERT + Ranking (Gu et al., 2021)	50.6	58.0	59.9	67.0	45.5	53.9	48.6	55.7
ReTraCk (Chen et al., 2021)	58.1	65.3	84.4	87.5	61.5	70.9	44.6	52.5
S^2QL (Zan et al., 2022)	57.5	66.2	65.1	72.9	54.7	64.7	55.1	63.6
ArcaneQA (Gu and Su, 2022)	63.8	73.7	85.6	88.9	65.8	75.3	52.9	66.0
RnG-KBQA (Ye et al., 2021)	68.8	74.4	86.2	89.0	63.8	71.2	63.0	69.2
TIARA (Ours)	**73.0**	**78.5**	**87.8**	**90.6**	**69.2**	**76.5**	**68.0**	**73.9**

Table 1: EM and F1 results (%) on the hidden test set of GrailQA. TIARA outperforms other methods with three levels of generalization settings in both EM and F1.

Evaluation Benchmark

WebQuestionsSP (Microsoft Research)

- 4,737 questions with full semantic parses
- SPARQL queries + rich semantic annotations

Method	**F1**	**Hits@1**
IR-based methods		
EmbedKGQA* (Saxena et al., 2020)	-	66.6
GRAFT-Net (Sun et al., 2018)	62.8	67.8
PullNet (Sun et al., 2019)	-	68.1
TransferNet♡ (Shi et al., 2021)	-	71.4
Relation Learning♡♣ (Yan et al., 2021)	64.5	72.9
NSM*♡ (He et al., 2021)	67.4	74.3
Subgraph Retrieval* (Zhang et al., 2022)	74.5	**83.2**

Method		**F1**
SP-based (feature-based ranking) methods		
TextRay♡ (Bhutani et al., 2019)	60.3	-
Topic Units♡ (Lan et al., 2019)	67.9	-
UHop (Chen et al., 2019)	68.5	-
GrailQA Ranking*♡♣ (Gu et al., 2021)	70.0	-
STAGG♡ (Yih et al., 2016)	71.7	-
QGG♡ (Lan and Jiang, 2020)	74.0	-
SP-based (seq2seq generation) methods		
NSM♡ (Liang et al., 2017)	69.0	-
ReTraCk (Chen et al., 2021)	71.0	71.6
CBR-KBQA (Das et al., 2021)	72.8	-
ArcaneQA (Gu and Su, 2022)	75.6	-
RnG-KBQA (Ye et al., 2021)	75.6	-
Program Transfer*♣ (Cao et al., 2022b)	76.5	74.6
👑 **TIARA** (Ours)	**76.7**	73.9
w/o Schema	76.4	73.7
w/o ELF	75.0	73.4
w/o ELF & Schema	73.2	71.1
👑 **TIARA***	**78.9**	75.2
w/o Schema	78.8	75.0
w/o ELF	76.2	74.5
w/o ELF & Schema	75.4	73.1

Table 2: F1 and hits@1 results (%) on WebQSP. * denotes using oracle entity linking annotations. ♡ denotes the assumption of a fixed number of hops. ♣ denotes pre-training on an auxiliary task or other KBQA datasets. For comparison, hits@1 on TIARA is obtained by randomly selecting one answer for each question 100 times.

Summary

- Multi-grained Retriever is critical for the system robustness

  Entity LF KB structure Schema semantic supplements

- Given enough information, PLMs can reason with high accuracy

Limitations

- Logical form retriever is not efficient
- Require strong supervision
- Gap between the pretraining tasks and KBQA

Thank you!