Human-Centered Planning

2024 Winter Seminar

인간중심컴퓨팅연구실 박사과정 류정우

WHY LLMs for planning?

왜 LLM인가?

Here, since ***the plan is executed by a human***, the output doesn't have to satisfy strict syntactic constraints. A useful assistant should also be able to incorporate ***vague constraints specified by the user in natural language***. This makes LLMs an attractive option for planning.

여기서 **계획은 사람에 의해 실행되므로** 출력이 엄격한 구문 제약 조건(syntactic constraints)을 충족할 필요는 없다. 유용한 assistant는 사용자가 **자연어로 지정하는 모호한 제약 조건을 계획에 반영할 수도 있어야** 한다. 이는 LLM이 계획 수립을 위한 매력적인 옵션으로 만든다.

WHY this paper?

왜 이 논문을 택했는가?

- Benchmark가 있는 well-defined task가 아닐 때 어떻게 시스템을 평가하는지?
- LLM이 사람을 위해 ‘생산적인 일’을 돕는다면 어떤 사용자 경험을 제공할 지?
- 원래 개인적으로 관심이 있었던 Task라서
- 향후 실제 서비스로 구현해보는 것도 재미있을 것 같아서
- LLM-based day planning 관련 연구는 희귀하며 본 논문이 최신 연구라서

Human-Centered Planning

Paper

Human-Centered Planning

Human-Centered Planning

Yuliang Li and **Nitin Kamra** and **Ruta Desai** and **Alon Halevy**
{yuliangli, nitinkamra, rutadesai, ayh}@meta.com
Reality Labs Research, Meta

Overview

The Planner(LLMPlan + SymPlan) Overview

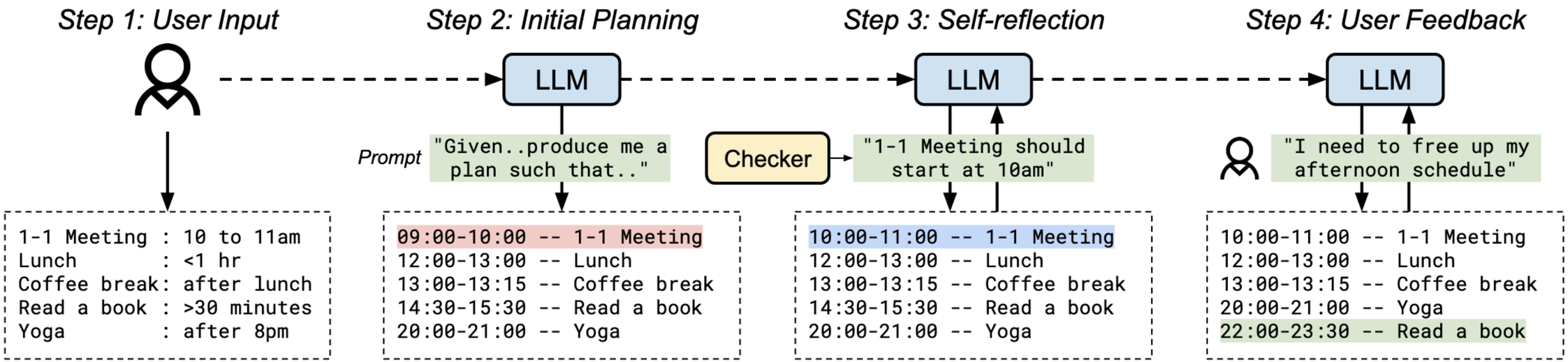

Figure 1: The LLMPlan planner takes as input the events to be scheduled with their associated constraints. After extracting the plan from the LLM's response, the system self-reflects by checking for constraint violations and prompting the LLM to fix them. A user may interact with LLMPlan in several iterations to refine their plan. In doing so, LLMPlan is able to incorporate constraints that may be vague. We show that the accuracy of LLMPlan is on par with a symbolic planner (SymPlan) which is unable to incorporate vague constraints.

Day Planning

Defining the Day Planning Problem

- The Planner의 목표는 제약조건(constraints)을 고려하고 일정간 충돌(conflicts)을 피하면서 계획을 수립하는 것
  - **Pre-scheduled**: events already scheduled for the day on one's calendar (e.g., 업무 미팅, 약속)
  - **Habits:** things that the user performs daily(e.g., 식사, 운동)
  - **ToDos**: tasks that the user wants to accomplish for the day(e.g., 독서, 빨래)

Day Planning

Defining the Day Planning Problem

- Types of constraints
  - Absolute temporal: 절대적인 시작/종료 시각
  - Duration: 이벤트의 지속시간
  - Relative temporal: 이벤트 간의 전후관계

Table 1: Types of constraints on events for day planning.

Constraint type	Examples
Absolute temporal	starts at 9am, ends at 15:00, before 5:00pm, after 2:00pm
Duration	<90 minutes, >1 hour
Relative temporal	after lunch

The Planner

LLM-based planner(LLMPlan) & Symbolic planner(SymPlan)

- **LLMPlan**
  - LLM의 **상식적 추론 능력**과 **구조화된 출력(계획) 생성 능력**을 활용하여 일일 계획 수립
  - 사용자는 고차원 목표, 입력 이벤트 및 제약 조건, 계획 예제로 구성된 프롬프트를 제공하고, LLM은 이를 기반으로 올바른 계획을 생성
  - 이는 사용자가 모든 제약 조건을 명시할 필요가 없으며, LLM은 상호 작용 과정에서 모호한 사용자 요구도 이해
- **SymPlan**
  - SymPlan은 이벤트 집합 E와 각 $e \in E$에 대한 제약 조건 집합 L이 주어졌을 때, 각 이벤트의 시작 및 종료 시간 (start_e 및 end_e)을 계산
  - 모든 제약 조건을 만족하는 여러 계획이 있을 수 있기 때문에 이들 중에서 일정의 총 길이를 최소화하는 최적화 문제를 풀어 선택

The Planner

LLM-based planner(LLMPlan) & Symbolic planner(SymPlan)

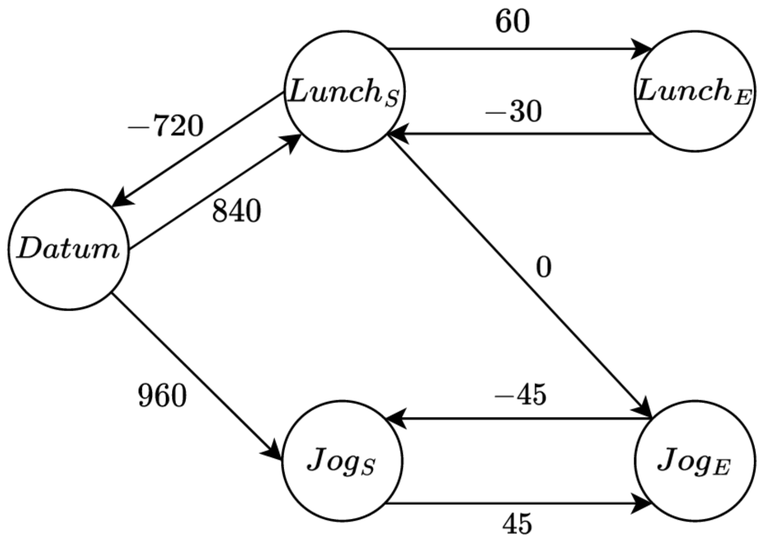

Figure 2: An example simple temporal network (STN) with start and end nodes for two events (Lunch and Jog) and the datum ($t = 0$). The edges represent temporal constraints, e.g., (a) the edges between $Lunch_S$ and $Lunch_E$ represent that its duration must be between 30 to 60 minutes, (b) Jog must start before 960 minutes from the datum i.e. before 4pm, and (c) Jog_E must happen before $Lunch_S$.

보충자료 링크

- **SymPlan**
  - SymPlan은 이벤트 집합 E와 각 e ∈ E에 대한 제약 조건 집합 L이 주어졌을 때, 각 이벤트의 시작 및 종료 시간 (start_e 및 end_e)을 계산
  - 모든 제약 조건을 만족하는 여러 계획이 있을 수 있기 때문에 이들 중에서 일정의 총 길이를 최소화하는 최적화 문제를 풀어 선택
  - STN을 사용하면 각 노드의 일정을 계획하기 위해 **최단 경로 알고리즘**을 실행하여 **상한 및 하한 시간**을 얻을 수 있음

Hybrid Planners

LLMPlan + self-reflection & SymPlan+

- **LLMPlan + self-reflection**
  - 이미 LLMPlan만으로도 대부분의 요건을 충족하는 스케줄링이 가능함을 확인(e.g., 아침 식사의 시작/종료 시각을 제약조건으로 주지 않더라도 알아서 아침 시간대에 계획)
  - 다만, 사소하지만 중요한 제약조건을 놓치는 오류(e.g., 시작/종료시각이 지정된 이벤트를 잘못 배치)를 보완하기 위해 output을 self-correct할 수 있도록 개선
- **SymPlan+**
  - 이벤트의 시작/종료시각이 완전히 지정되지 않은 경우, 먼저 추가 가능한 제약 조건에 대해 LLM에 쿼리를 보내 확인 - LLM의 common sense 활용
  - LLM이 제안한 constraints가 사용자의 다른 constraints와 충돌이 없을 경우 append
  - 이후 SymPlan과 동일한 과정

Experiments

Datasets

Table 2: Statistics of the two datasets. Note that the synthetic dataset has ∼2x the number of events and constraints per user (the numbers in parentheses) compared to the real-user dataset.

	#users	#events	#constraints
Synthetic	100	1,035 (10.35)	2,128 (21.28)
Real-user	40	223 (5.575)	384 (9.6)

Table 3: Example events and constraints from our datasets. Note that the real-user dataset may contain events that do not belong to the 3 pre-defined categories.

Event	Example	Example Constraints
ToDo	Plan a vacation	>1 hour, <2 hours
Habit	grocery shopping	after meeting, before 17
Pre-scheduled	Project Update	starts at 10, ends at 11
Real-user	Play Xbox	after dinner

Experiments

Evaluation Metrics - 직접 만들었음

- **Correctness metrics**
  - Event Coverage(CO): 주어진 이벤트가 계획에 반영된 %
  - Non-Overlap(NO): 다른 이벤트와 겹치지 않는 이벤트의 %
  - Duration Constraints(DC), Order Constraints(OC), Start/End Constraints(SEC)를 만족하는 %
- **Commonsense Violations**
  - 인간의 상식선에서 납득 가능한지
  - Commonsense constraints on event **duration** & **start/end time**

Results

Results for the synthetic dataset

Table 4: Plan quality metrics for our synthetic dataset.

Methods	Correctness (↑)						Commonsense (↓)		
	CO	NO	DC	OC	SEC	AVG	duration	start_end	AVG
GPT-3 (curie)	23.91	**100.00**	40.00	84.72	14.29	52.58	28.00	29.45	28.73
GPT-3 (davinci)	81.16	99.17	66.86	82.79	33.33	72.66	16.50	24.06	20.28
GPT-3.5 (SFT, text-davinci-002)	88.54	99.80	93.00	96.59	69.75	89.54	13.22	17.76	15.49
GPT-3.5 (RL, text-davinci-003)	98.95	98.25	92.77	96.76	73.00	91.94	17.82	16.46	17.14
GPT-3.5 + self-reflect	99.83	97.80	98.29	98.38	79.19	94.70	16.81	18.06	17.43
ChatGPT	88.35	98.68	95.05	94.60	82.98	91.93	11.19	**15.06**	13.12
ChatGPT + self-reflect	94.29	97.48	99.21	97.38	97.88	97.25	10.35	15.73	13.04
GPT-4	88.96	98.68	94.81	94.94	85.03	92.48	10.52	15.71	13.11
GPT-4 + self-reflect	93.16	97.38	**99.47**	97.08	98.38	97.09	**10.21**	15.38	**12.80**
SymPlan	99.72	**100.00**	95.90	**99.68**	**100.00**	**99.06**	24.36	34.42	29.39
SymPlan+	**100.00**	98.22	98.56	97.66	99.51	98.79	25.88	12.33	19.11

Results

Results for the real users' dataset

Table 5: Plan quality metrics for our dataset on real users.

Methods	Correctness (↑)						Commonsense (↓)		
	CO	NO	DC	OC	SEC	AVG	duration	start_end	AVG
GPT-3 (curie)	21.58	100.00	78.57	67.92	33.33	60.28	**5.67**	57.45	31.56
GPT-3 (davinci)	78.31	94.03	61.54	68.25	42.55	68.94	9.09	31.25	20.17
GPT-3.5 (SFT, text-davinci-002)	88.99	98.96	96.36	90.85	78.43	90.72	11.71	24.32	18.02
GPT-3.5 (RL, text-davinci-003)	97.10	99.58	93.55	93.55	74.07	91.57	15.25	24.58	19.92
GPT-3.5 + self-reflect	98.45	99.17	**97.87**	95.17	84.62	95.06	14.55	24.55	19.55
ChatGPT	82.86	98.20	94.64	96.83	95.12	93.53	9.71	17.48	13.59
ChatGPT + self-reflect	88.88	96.22	97.67	97.67	93.75	94.84	11.70	17.02	14.36
GPT-4	82.65	97.43	94.44	97.58	95.12	93.44	8.82	17.65	**13.24**
GPT-4 + self-reflect	89.19	98.04	97.67	**98.40**	94.12	95.49	10.42	17.71	14.06
SymPlan	98.92	98.90	95.24	97.26	**100.00**	**98.06**	13.33	35.83	24.58
SymPlan+	**99.27**	**100.00**	93.65	97.18	96.23	97.27	15.70	**15.70**	15.70

Results

Error Analysis

SymPlan:

```
07:00 - 07:30 -- run
07:30 - 08:00 -- breakfast
08:00 - 08:30 -- work
08:30 - 09:00 -- lunch
09:00 - 10:00 -- Conflict Resolution
                 Meeting
10:00 - 10:30 -- meditation
10:30 - 11:00 -- dinner
11:00 - 12:00 -- Attend a networking
                 event
12:00 - 13:00 -- Project Update
13:00 - 13:30 -- Review and update
                 your resume
```

LLMPlan with GPT-4

```
06:00 - 07:00 -- run
07:00 - 08:00 -- breakfast
08:00 - 12:00 -- work
12:00 - 13:00 -- Project Update
13:00 - 14:00 -- lunch
14:00 - 14:15 -- meditation
14:15 - 15:15 -- Review and update
                 your resume
15:15 - 17:15 -- Attend a networking
                 event
17:15 - 19:15 -- free time
19:15 - 20:30 -- dinner
```

LLMPlan with GPT-4+Self-Reflection:

```
06:00 - 07:00 -- run
07:00 - 08:00 -- breakfast
08:00 - 09:00 -- review and update your
                 resume
09:00 - 10:00 -- Conflict Resolution
                 Meeting
10:00 - 12:00 -- work
12:00 - 13:00 -- Project Update
13:00 - 14:00 -- lunch
17:30 - 17:45 -- meditation
18:00 - 19:30 -- dinner
19:30 - 21:30 -- Attend a networking
                 event
```

Figure 3: Output of 3 planners for the persona in Figure 4. While SymPlan satisfies all the input constraints, it violates the time/duration commonsense for work and meals (in red). LLM (GPT-4) generates a natural and mostly correct plan with 1 input event missing. Self-reflection correctly adds back the missing event (in green). Note that LLM may generate novel events not from the input such as "free time" (in yellow).

Results

User Feedback - Initial plan 제시, Iterative refinements

Table 6: Breakdown of all 113 user requirements into 3 categories: well-defined, open-ended, and complex.

Category	Count	Examples
Well-defined	59	"add a phone call at 10:00", "shorten my lunch to 30 minutes", "End the last event before 20:00"
Open-ended	40	"schedule meal breaks", "there will be a rain around 10am, what should i prepare?"
Complex	14	"I need to break up the demos into two meetings", "I want at least 5 minutes between events"

Results

User Feedback - Initial plan 제시, Iterative refinements

Table 6: Breakdown of all 113 user requirements into 3 categories: well-defined, open-ended, and complex.

Category	Count	Examples
Well-defined	59	"add a phone call at 10:00", "shorten my lunch to 30 minutes", "End the last event before 20:00"
Open-ended	40	"schedule meal breaks", "there will be a rain around 10am, what should i prepare?"
Complex	14	"I need to break up the demos into two meetings", "I want at least 5 minutes between events"

Table 7: Overview of user feedback.

	LLMPlan		SymPlan+		Both	
Feedback	Pos.	Neg.	Pos.	Neg.	Pos.	Neg.
Count	31	13	17	25	11	9
%	**70.5%**	29.5%	40.5%	59.5%	55%	45%

Table 8: The numbers of requirements per category and their corresponding user feedback are shown. LLMPlan performs well on the open-ended requirements, while the complex ones are the difficult for both planners.

	LLMPlan		SymPlan+		Overall	
Category	Pos.	Neg.	Pos.	Neg.	Pos.	Neg.
Well-defined	36	23	24	35	60	58
Open-ended	29	11	18	22	47	33
Complex	7	7	5	9	12	16

Conclusion

+ Future work

- LLM-based day planning은 자연어 명령과 다중 대화에 적합한 기술이라 판단
- 명시적 제약이 있는 경우에는 Symbolic planning에 견줄만하며, 더불어 LLM은 누락된 제약을 common sense를 활용하여 추론할 수 있음을 입증
- 실제 사용자 실험에서는 몇몇 복잡한 제약이 발견되었으며, LLM이 이를 효과적으로 처리하기 위한 추가 연구가 필요
- 향후 Chain-of-thought와 사용자 맥락(선호도, 습관 등)을 고려하는 방법을 연구하여 계획 수립의 정확성과 효과를 더욱 향상시킬 수 있을 것으로 기대(Long-term memory)

contact me

jeongwoo@snu.ac.kr