Benchmarks

Bridging Planning and Reasoning in Natural Language with Foundational Models

Harsha Kokel

https://plan-fm.github.io/

Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context

Gemini Team, Google[1]

In this report, we introduce the Gemini 1.5 family of models, representing the next generation of highly compute-efficient multimodal models capable of recalling and reasoning over fine-grained information from millions of tokens of context, including multiple long documents and hours of video and audio. The family includes two new models: (1) an updated Gemini 1.5 Pro, which exceeds the February version on the great majority of capabilities and benchmarks; (2) Gemini 1.5 Flash, a more lightweight variant designed for efficiency with minimal regression in quality. Gemini 1.5 models achieve near-perfect recall on long-context retrieval tasks across modalities, improve the state-of-the-art in long-document QA, long-video QA and long-context ASR, and match or surpass Gemini 1.0 Ultra's state-of-the-art performance across a broad set of benchmarks. Studying the limits of Gemini 1.5's long-context ability, we find continued improvement in next-token prediction and near-perfect retrieval (>99%) up to at least 10M tokens, a generational leap over existing models such as Claude 3.0 (200k) and GPT-4 Turbo (128k). Finally, we highlight real-world use cases, such as Gemini 1.5 collaborating with professionals on completing their tasks achieving 26 to 75% time savings across 10 different job categories, as well as surprising new capabilities of large language models at the frontier; when given a grammar manual for Kalamang, a language with fewer than 200 speakers worldwide, the model learns to translate English to Kalamang at a similar level to a person who learned from the same content.

1. Introduction

We present our latest multimodal models from the Gemini line: Gemini 1.5 Pro and Gemini 1.5 Flash. They are members of Gemini 1.5, a new family of highly-capable multimodal models which incorporates our latest innovations in sparse and dense scaling as well as major advances in training, distillation and serving infrastructure that allow it to push the boundary of efficiency, reasoning, planning, multi-linguality, function calling and long-context performance. Gemini 1.5 models are built to handle extremely long contexts; they have the ability to recall and reason over fine-grained information from up to at least 10M tokens. This scale is unprecedented among contemporary large language models (LLMs), and enables the processing of long-form mixed-modality inputs including entire collections of documents, multiple hours of video, and almost five days long of audio.

PlanBench

- 2 domains (+ Obfuscated)
- 8 Tasks

1. Plan generation
2. Cost Optimal Planning
3. Plan Verification
4. Reasoning about plan execution
5. Robustness to goal reformulation
6. Ability to reuse plans
7. Replanning
8. Plan Generalization

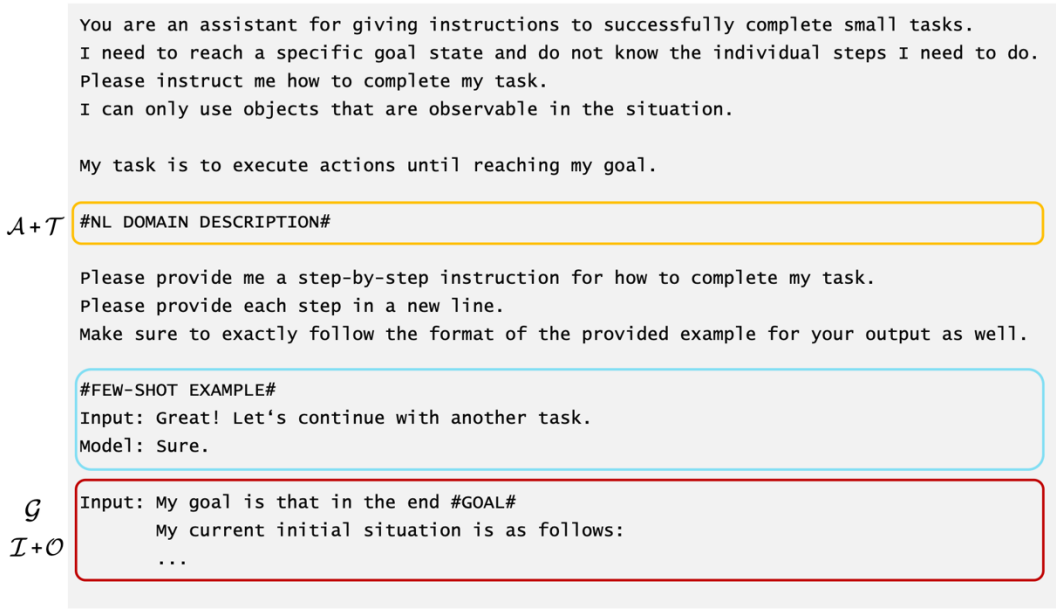
```
========================================
I am playing with a set of blocks where I need to arrange the blocks into stacks. Here are the
↪  actions I can do

Pick up a block
Unstack a block from on top of another block
Put down a block
Stack a block on top of another block

I have the following restrictions on my actions:
I can only pick up or unstack one block at a time.
I can only pick up or unstack a block if my hand is empty.
I can only pick up a block if the block is on the table and the block is clear. A block is clear
↪  if the block has no other blocks on top of it and if the block is not picked up.
I can only unstack a block from on top of another block if the block I am unstacking was really on
↪  top of the other block.
I can only unstack a block from on top of another block if the block I am unstacking is clear.
Once I pick up or unstack a block, I am holding the block.
I can only put down a block that I am holding.
I can only stack a block on top of another block if I am holding the block being stacked.
I can only stack a block on top of another block if the block onto which I am stacking the block
↪  is clear.
Once I put down or stack a block, my hand becomes empty.
Once you stack a block on top of a second block, the second block is no longer clear.

[STATEMENT]
As initial conditions I have that, the red block is clear, the blue block is clear, the yellow
↪  block is clear, the hand is empty, the blue block is on top of the orange block, the red block
↪  is on the table, the orange block is on the table and the yellow block is on the table.
My goal is to have that the orange block is on top of the blue block.

My plan is as follows:

[PLAN]
```

https://github.com/karthikv792/LLMs-Planning

Valmeekam et al. NeurIPS 23

AutoPlanBench

- 12 domains
- 1 Task
  - Plan generation



https://github.com/minecraft-saar/autoplanbench

Stein et al. 23

TRAC

- Blocksworld
- 4 Tasks
  - Projection
  - Executability
  - Plan verification
  - Goal Recognition

Task	Context	Query	Answer
PR	s: The green block is on the table. The red block is clear. The blue block is clear. The green block is clear. The red block is on the table. The blue block is on the table. $\vec{a}$: Jane moves the green block from the table to the red block.	q: The blue block is on top of the red block.	False
EX	s: The olive block is on the table. The yellow block is on top of the olive block. The indigo block is clear. The indigo block is on top of the yellow block.	$\vec{a}$: Jane moves the indigo block from the yellow block onto the table.	True
PV	s: The blue block is clear. The blue block is on top of the magenta block. The magenta block is on top of the white block. The white block is on the table. g: the blue block is not on top of the magenta block	$\vec{a}$: Jane moves the blue block from the magenta block onto the table.	True
GR	s: The blue block is clear. The blue block is on top of the magenta block. The magenta block is on top of the white block. The white block is on the table. $\vec{a}$: Jane moves the blue block from the magenta block onto the table.	g: the blue block is on top of the magenta block.	False

Textual Reasoning about Actions and Change

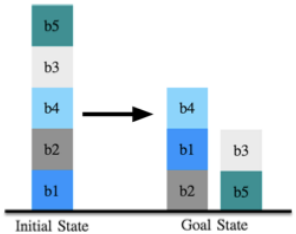
https://github.com/sysulic/trac

He et al. ACL 23

LLM+P

- 7 domain
- 2 tasks
  - Plan Generation
  - NL2PDDL Translation

Problem (P1): You have 5 blocks. One cannot place more than one block on another block. b5 is on top of b3. b4 is on top of b2. b2 is on top of b1. b3 is on top of b4. b1 is on the table. b5 is clear. Your arm is empty.
Your goal is to move the blocks.
b1 should be on top of b2.
b3 should be on top of b5.
b4 should be on top of b1.



GPT-4:

```
1. Move b5 from b3 to the table.
2. Move b3 from b4 to b5.
3. Move b4 from b2 to the table.
4. Move b2 from b1 to b5.
5. Move b1 from the table to b2.
6. Move b4 from the table to b1.
```

https://github.com/Cranial-XIX/llm-pddl/

Liu et al. 23

ACPBench

- 13 domains
- 7 tasks
  - Action Applicability
  - Progression
  - Atom Reachability
  - Validation
  - Action Reachability
  - Justification
  - Landmarks

```
**Question**: This is a ferry domain, where the task is
 to transport cars from their start to their goal
 locations, using a ferry. Each location is accessible by
  ferry from each other location. The cars can be
 debarked or boarded, and the ferry can carry only one
 car at a time.  There are 2 locations and 2 cars,
 numbered consecutively.  Currently, the ferry is at l0,
 with the car c1 on board. The cars are at locations as
 follows: c0 is at l0.  Is the following action
 applicable in this state:  travel by sea from location
 l1 to location l0?
**Thoughts**: Let's think step by step.
Step 1: In order to apply the action travel by sea from
 location l1 to location l0, the following fact(s) must
 hold in this state: The ferry is at l1 location.
Step 2: These facts do not hold in the mentioned state.
So, the action is not applicable.
**Final Answer**: No.
**Question**: ...
**Thoughts**: ...
**Final Answer**: Yes.
**Question**:  <context> + <question>
**Thoughts**: Let's think step by step.
```

https://github.com/ibm/ACPBench

Kokel et al. AAAI 25

ActionReasoningBench

- 8 domains
- 6 tasks
  - Fluent Tracking
  - State Tracking
  - Action Executability
  - Effects of Actions
  - Numerical RAC
  - Composite Question

Blocksworld domain with Ramifications for a single sequence of action for Fluent Tracking

[DOMAIN DESCRIPTION]
A block can only be picked up if it is clear, on the table, and the hand is empty, resulting in the block being held. A held block can be put down, placing it back on the table. Blocks can be stacked if the first block is held and the second block is clear, causing the first block to rest on top of the second. Unstacking occurs when the hand is empty, the first block is clear, and on top of the second, resulting in the first block being held again. A block can't be at two locations at the same time and is considered clear if nothing is on top of it and it's not held, and the hand is empty if it's not holding anything. Blocks are stable when clear and on the table, and they can be painted if stable and the hand is empty. A block is considered on display if it can be painted and has no other block on top of it.

[INITIAL CONDITIONS]
Block b1 is situated on the table, block b2 is not stacked with any other block, block b2 is also on the table, block b3 is not stacked with any other block, block b3 is positioned on top of block b7, block b4 is stacked on top of block b1, block b5 is not stacked with any other block, block b5 is placed on top of block b4, block b6 is on the table, block b7 is stacked on top of block b6, and the hand is empty.

[QUESTION]
Starting from the initial condition, the following actions are taken: block b3 is unstacked from the top of block b7 to achieve the current state. In this state, what are the valid properties (including both affirmative and negated properties) for b7? If there are no valid properties, write None.

https://github.com/izuminka/reasoning_about_actions

Handa et al. ICLR 25

Other

- NL Planning Benchmarks
  - Travel Planner, Kie et al ICML 24 (https://osu-nlp-group.github.io/TravelPlanner/)
  - Natural Plan, Zheng et al 24 (https://github.com/google-deepmind/natural-plan)
- NL to PDDL translations
  - NL2PDDL, Oswald et al ICAPS 24 (https://github.com/IBM/NL2PDDL)
  - LLM+P, Liu et al 23 (https://github.com/Cranial-XIX/llm-pddl/)
  - Planetarium, Zuo et al 24 (https://github.com/BatsResearch/planetarium)
- Agent
  - Agent Board, Ma et al NeurIPS 24 (https://github.com/hkust-nlp/AgentBoard)
  - TextCraft Prasad et al. NAACL 24 (https://github.com/archiki/ADaPT/tree/main/TextCraft)
  - ALFRED, ALFWorld, WebShop, WebArena etc…


HellaSwag
CommonSense QA
GSM8K
TabFact
HybridQA
PrOnotQA
MuSiQue
TriviaQA
StrategyQA
TheoremQA
Hotpot QA
Spider
DROP
CRASS
MultiModalQA
CREAK
BigBench
TimeDial
TempQuestion
Social IQA

LINE GOES UP? INHERENT LIMITATIONS OF BENCHMARKS FOR EVALUATING LARGE LANGUAGE MODELS

James Fodor

Reasoning or Reciting? Exploring the Capabilities and Limitations of Language Models Through Counterfactual Tasks

Zhaofeng Wu **Linlu Qiu** **Alexis Ross** **Ekin Akyürek** **Boyuan Chen**
Bailin Wang **Najoung Kim** **Jacob Andreas** **Yoon Kim**
MIT Boston University
zfw@csail.mit.edu

Inadequacies of Large Language Model Benchmarks in the Era of Generative Artificial Intelligence

Timothy R. McIntosh*, Teo Susnjak, Nalin Arachchilage, Tong Liu, Dan Xu, Paul Watters, *Senior Member, IEEE*, and M

Do These LLM Benchmarks Agree? Fixing Benchmark Evaluation with BenchBench

Yotam Perlitz[1] **Ariel Gera**[1] **Ofir Arviv**[1] **Elron Bandel**[1]
Asaf Yehudai[1] **Eyal Shnarch**[1] **Michal Shmueli-Scheuer**[1] **Leshem Choshen**[2,3]

Benchmark	PlanBench	Auto PlanBench	TRAC	LLM+P	ActionReasoning Bench	ACPBench
LLM **supported**	OpenAI Bloom	OpenAI	T5 OpenAI Roberta	OpenAI	huggingface	Huggingface OpenAI vLLMs, etc
Prompts	[STATEMENT] [PLAN] [PLAN END]	Input: Model:	None `tokenize(s" "a) + tokenize(q)`	An example planning problem is: .... A plan for the example problem is... Can you provide an optimal plan ...	[DOMAIN DESCRIPTION] [INITIAL CONDITION] [QUESTION]	**Question**: **Thoughts**: **Final Answer**:

THANK YOU