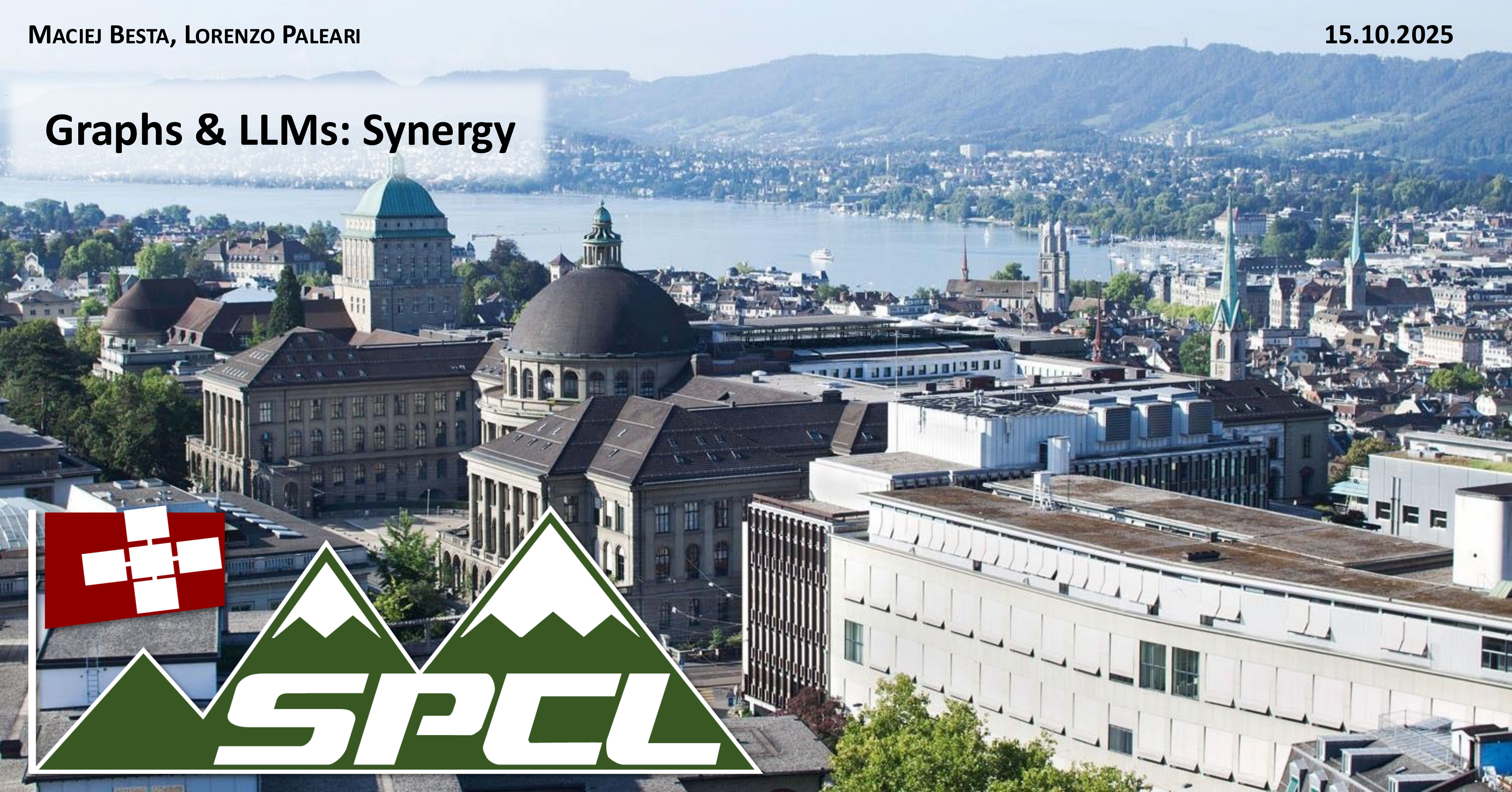
MACIEJ BESTA, LORENZO PALEARI

15.10.2025

Graphs & LLMs: Synergy



KV Caching -> LoRA

- **Observation**: During inference, the model builds a KV cache (K,V) capturing contextual activations for each token --> (Ex. ChatGPT)
- **Problem**: These caches are discarded after use (Too big to be stored for each chat and user)
  - Waste of rich latent information
  - Limits contextual continuity --> New chat without this information's (need to recompute KV Cache and store it (?))
- **Idea**:
  - Implicit context compression --> the model "remembers" without storing information's/KV-Cache

Formalization: KV Distillation

- f_θ: base model with parameters θ
- $C = \{(K_i, V_i, x_i)\}$: cached key–value pairs and inputs from active sessions
- $\Delta W = AB$

- **Goal**: merge knowledge into θ via a lightweight update.
- $\min(\Delta \mathrm{W})\, L_{distill} = \mathrm{E}_{(x,K,V)\sim C}[\| f_{\theta+\Delta W(x)} - f\ \theta(x,K,V) \|^2]$

- **Target**: model's own cached contextual output.
  - Once trained, the adapter can be merged into the model

Open Questions + Next Steps

- **Aggregation strategies**
  - **Incremental Merging**: sequentially updating same LoRA adapter
  - **Compositional merging**: combine multiple adapters

- **Benchmarks & Evaluation**
  - Context memory benchmarks: retention, forgetting, drift, reasoning….
  - Long Context + Multi-Turn / Reasoning (NiHS, BBH…)

- Almost finished creating the pipeline to start testing in bulk.
  - General pipeline, allows to select different models to try (I was thinking about lower Billion models)
  - Flexible to accept many different benchmarks

Long Context Benchmarks

Name	Max. Length / Scale	Solved / Accuracy	Type	Comments	Links
NiHS (Needle in the Haystack)	1M tested → 10M mentioned by Google (148k token max / 110k words)	99.7% (Video: 10h, Audio: 5 days) 99.2% on 10M recall	Search / Information in Long Context (Text / Audio / Video)	Tests positional information at depth. Can recall multiple NiHS with variations. 100 needles tested (70% recall top 128k). No reasoning, requires precise wording.	Blog GitHubPaper
Multimodal NiHS	40k images / 560k captions / 280k needles Max images = upload × grid_dim	97% (10×2×2), 27% (10×4×4)	Image / Multimodal Search	Tests model's ability to find correct images from captions. Constrained by upload limit. Stitching images improves results but affects outcomes. Mentions image downscaling and tokenization.	Paper
MMMU / MMMU-Pro / MME / MMBench / MMT / Vibe-Eval	— (mostly short) MMT slightly longer	80%+, 80, 80, 63, 60	Multimodal QA (Images / Text / Cross-modal)	Hard questions across text & modality. MMT includes temporal reasoning, 3D, etc.	2311.165022404.160062405.022872307.062812306.13394
LongBench / LongBench v2	Pure text up to 10k tokens	—	Text QA / Reasoning	Multiple choice tasks. v2 adds more reasoning-oriented prompts.	2308.145082412.15204
InfiniBench	~200k tokens average	50%	Multitask (retrieval, code, math, QA, dialogue)	Tests long structured data, summarization, and QA.	2402.13718
RULER	From 4k to 128k (configurable)	96% (Gemini)	Text QA / Multi-hop reasoning	Tests retrieval, tracing, aggregation, multi-hop reasoning.	2404.06654
Big-Bench Hard (BBH)	Short context	74%	Text / Reasoning	Tasks include sorting, navigation, analogical reasoning, symbolic manipulation, and code. 23 tasks, 100+ questions each. Focused on chain-of-thought (CoT).	2210.09261
Big-Bench Extra Hard (BBEH)	Inputs ~6× BBH (paragraph-length)	—	Text / Reasoning	Temporal, spatial, logical reasoning. 120–200 questions per task.	2502.19187

Long Context Benchmarks

- **Overview**
  - NiHS benchmark nearly solved — recall up to 10M tokens (text, audio, video).
    - *Scales easily (e.g., repeating Paul Graham essays), but most other benchmarks still short-context.*
  - Multimodal NiHS tests image–caption retrieval.
  - Others explore chain retrieval, multi-needle retrieval, and reasoning (BBH / BBEH: spatial, temporal, logical).

- **Next Steps**
  - Simulate long, natural conversation between 2 LLMs. (Simulation of human-human or human-LLM)
  - Use different LLMs with personality.
    - *Maciej paper?*
  - Check what have been created (random checks) and inject in-post reasoning/needles.