# Meeting Notes—Summary LLM

Bas Donkers, Dennis Fok, Finn-Ole Höner

March 12, 2024

## 18-01-2024

- Status: First steps towards a summary LLM, basic computations seem to work.
- Admin: Setup of MS Teams group
- Technical Note: The `output` object is a tuple of length "positions", and each row of the contained matrix is one beam.

### Dutch LLMs

I found some dutch language models on Huggingface, e.g. a GPT-2 version which I use for the illustration below (also available with more parameters, see corresponding paper). There are also other "dutch" versions of other language models, see this collection. One caveat might be that all these models are in some way fine-tuned/transfer learned from the English version of the model, so not "natively" trained on only dutch corpora. However, it seems to work in our current "brute force" optimization.See attachements for the targetHeeft het zin om appels en sinaasappelen te vergelijken?? (sic)

### Speeding up computations

Currently, we are performing too many computations. We can make use of the fact that we know the target string. Essentially, we can compute each "next word" in parallel, instead of sequentially. This also circumvents the problem that we need to think about generation strategies, as we can essentially just make a forward pass through the LLM.

### TO-DOs

- Finn speeds computations up
- Finn tries to write draft for grant proposal
- Reaching out to SKIM, slide deck, NDA

## 12-03-2024 (Surf Research Cloud Workshop)

- Available flavors
- Small Compute Applications (E-Infra grant)
  - documentation
  - 1 year valid, can apply every *calendar* year
  - up to 5,000 GPU hours
  - can extend by 6 months if resources are left
  - goes through SURF, can help with resources
  - Estimate by iteration time, core hours + 20% overhead
  - Need somebody who has permanent contract with Dutch institution
  - hard *not* to get
- NWO Groot grant is bigger, but slower and harder to get
  - Can chain this together with the small compute applications
- At EUR will create new policy on this next month
- Sounds like there will be resources that your department can free up directly
- RCCS contract
- Likly much faster GPUs then e.g. Google Colab
- Can also have budgets at other Cloud services and use research cloud for collaboration, e.g. Azure, AWS
- Could we get resources for the NLP course? They are also dedicated to education, not only research

### People

- Pieter Meijndert, ESE, Head of Business Intelligence and Application Management
- Yuliia Orlova, Liasion User-Developers
- Carsten Schelp

### Workshop

- Pick collaboration and budget work a workspace, people from collaboration get access
- e.g RStudio, Jupyter available
- Disk budget will expire when grant expires, not an archive. Quite fast discs. Only upload data once and use from multiple instances.
- Storage is persistent, data is not lost if VM crashes.
- Attach the storage to the workspace
- Set deadline for the machine, normal lifetime is 5 days
- Shortlifed instances support reproduceability
- RStudio "username": Profile tab of the portal; time-based password. For Jupyter its the portal credentials.
  - Can also be SSH for some instances
- Can probably connect through VSCode and SSH to instance directly
- Whisper instance on cloud, useful for transcription?

## Questions

- What is a GPU core hour?