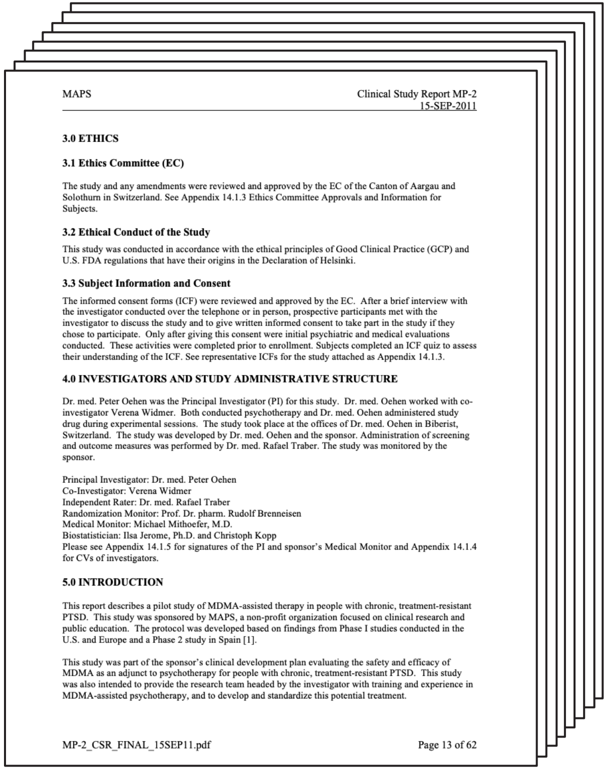
NLP SUMMIT
HEALTHCARE

# Long Document Abstractive Summarization with Large Language Models

**Natalia Vassilieva**
Director of Product
Cerebras Systems

# Summarization: long texts to short summaries



MAPS

Clinical Study Report MP-2
15-SEP-2011

**3.0 ETHICS**

**3.1 Ethics Committee (EC)**

The study and any amendments were reviewed and approved by the EC of the Canton of Aargau and Solothurn in Switzerland. See Appendix 14.1.3 Ethics Committee Approvals and Information for Subjects.

**3.2 Ethical Conduct of the Study**

This study was conducted in accordance with the ethical principles of Good Clinical Practice (GCP) and U.S. FDA regulations that have their origins in the Declaration of Helsinki.

**3.3 Subject Information and Consent**

The informed consent forms (ICF) were reviewed and approved by the EC. After a brief interview with the investigator conducted over the telephone or in person, prospective participants met with the investigator to discuss the study and to give written informed consent to take part in the study if they chose to participate. Only after giving this consent were initial psychiatric and medical evaluations conducted. These activities were completed prior to enrollment. Subjects completed an ICF quiz to assess their understanding of the ICF. See representative ICFs for the study attached as Appendix 14.1.3.

**4.0 INVESTIGATORS AND STUDY ADMINISTRATIVE STRUCTURE**

Dr. med. Peter Oehen was the Principal Investigator (PI) for this study. Dr. med. Oehen worked with co-investigator Verena Widmer. Both conducted psychotherapy and Dr. med. Oehen administered study drug during experimental sessions. The study took place at the offices of Dr. med. Oehen in Biberist, Switzerland. The study was developed by Dr. med. Oehen and the sponsor. Administration of screening and outcome measures was performed by Dr. med. Rafael Traber. The study was monitored by the sponsor.

Principal Investigator: Dr. med. Peter Oehen
Co-Investigator: Verena Widmer
Independent Rater: Dr. med. Rafael Traber
Randomization Monitor: Prof. Dr. pharm. Rudolf Brenneisen
Medical Monitor: Michael Mithoefer, M.D.
Biostatistician: Ilsa Jerome, Ph.D. and Christoph Kopp
Please see Appendix 14.1.5 for signatures of the PI and sponsor's Medical Monitor and Appendix 14.1.4 for CVs of investigators.

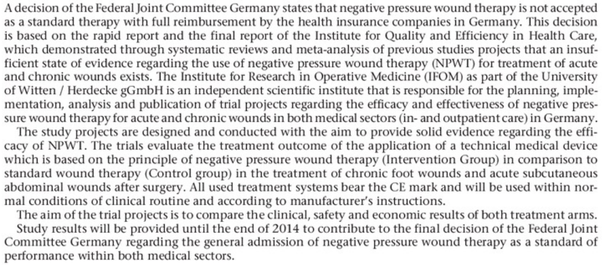
**5.0 INTRODUCTION**

This report describes a pilot study of MDMA-assisted therapy in people with chronic, treatment-resistant PTSD. This study was sponsored by MAPS, a non-profit organization focused on clinical research and public education. The protocol was developed based on findings from Phase I studies conducted in the U.S. and Europe and a Phase 2 study in Spain [1].

This study was part of the sponsor's clinical development plan evaluating the safety and efficacy of MDMA as an adjunct to psychotherapy for people with chronic, treatment-resistant PTSD. This study was also intended to provide the research team headed by the investigator with training and experience in MDMA-assisted psychotherapy, and to develop and standardize this potential treatment.

MP-2_CSR_FINAL_15SEP11.pdf

Page 13 of 62



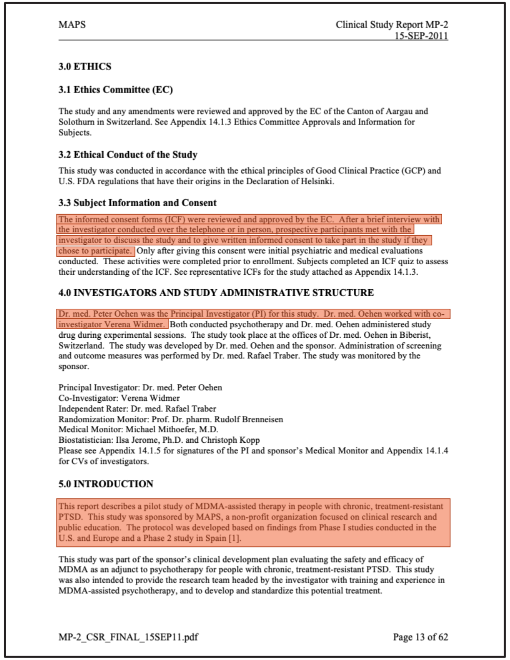
A decision of the Federal Joint Committee Germany states that negative pressure wound therapy is not accepted as a standard therapy with full reimbursement by the health insurance companies in Germany. This decision is based on the rapid report and the final report of the Institute for Quality and Efficiency in Health Care, which demonstrated through systematic reviews and meta-analysis of previous studies projects that an insufficient state of evidence regarding the use of negative pressure wound therapy (NPWT) for treatment of acute and chronic wounds exists. The Institute for Research in Operative Medicine (IFOM) as part of the University of Witten / Herdecke gGmbH is an independent scientific institute that is responsible for the planning, implementation, analysis and publication of trial projects regarding the efficacy and effectiveness of negative pressure wound therapy for acute and chronic wounds in both medical sectors (in- and outpatient care) in Germany.

The study projects are designed and conducted with the aim to provide solid evidence regarding the efficacy of NPWT. The trials evaluate the treatment outcome of the application of a technical medical device which is based on the principle of negative pressure wound therapy (Intervention Group) in comparison to standard wound therapy (Control group) in the treatment of chronic foot wounds and acute subcutaneous abdominal wounds after surgery. All used treatment systems bear the CE mark and will be used within normal conditions of clinical routine and according to manufacturer's instructions.

The aim of the trial projects is to compare the clinical, safety and economic results of both treatment arms.

Study results will be provided until the end of 2014 to contribute to the final decision of the Federal Joint Committee Germany regarding the general admission of negative pressure wound therapy as a standard of performance within both medical sectors.

# Summarization: Extractive



MAPS Clinical Study Report MP-2
15-SEP-2011

**3.0 ETHICS**

**3.1 Ethics Committee (EC)**

The study and any amendments were reviewed and approved by the EC of the Canton of Aargau and Solothurn in Switzerland. See Appendix 14.1.3 Ethics Committee Approvals and Information for Subjects.

**3.2 Ethical Conduct of the Study**

This study was conducted in accordance with the ethical principles of Good Clinical Practice (GCP) and U.S. FDA regulations that have their origins in the Declaration of Helsinki.

**3.3 Subject Information and Consent**

The informed consent forms (ICF) were reviewed and approved by the EC. After a brief interview with the investigator conducted over the telephone or in person, prospective participants met with the investigator to discuss the study and to give written informed consent to take part in the study if they chose to participate. Only after giving this consent were initial psychiatric and medical evaluations conducted. These activities were completed prior to enrollment. Subjects completed an ICF quiz to assess their understanding of the ICF. See representative ICFs for the study attached as Appendix 14.1.3.

**4.0 INVESTIGATORS AND STUDY ADMINISTRATIVE STRUCTURE**

Dr. med. Peter Oehen was the Principal Investigator (PI) for this study. Dr. med. Oehen worked with co-investigator Verena Widmer. Both conducted psychotherapy and Dr. med. Oehen administered study drug during experimental sessions. The study took place at the offices of Dr. med. Oehen in Biberist, Switzerland. The study was developed by Dr. med. Oehen and the sponsor. Administration of screening and outcome measures was performed by Dr. med. Rafael Traber. The study was monitored by the sponsor.

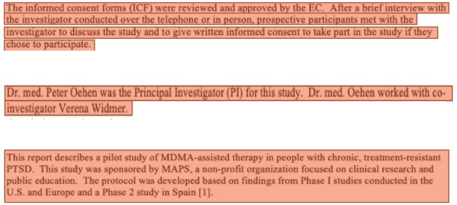
Principal Investigator: Dr. med. Peter Oehen
Co-Investigator: Verena Widmer
Independent Rater: Dr. med. Rafael Traber
Randomization Monitor: Prof. Dr. pharm. Rudolf Brenneisen
Medical Monitor: Michael Mithoefer, M.D.
Biostatistician: Ilsa Jerome, Ph.D. and Christoph Kopp
Please see Appendix 14.1.5 for signatures of the PI and sponsor's Medical Monitor and Appendix 14.1.4 for CVs of investigators.

**5.0 INTRODUCTION**

This report describes a pilot study of MDMA-assisted therapy in people with chronic, treatment-resistant PTSD. This study was sponsored by MAPS, a non-profit organization focused on clinical research and public education. The protocol was developed based on findings from Phase I studies conducted in the U.S. and Europe and a Phase 2 study in Spain [1].

This study was part of the sponsor's clinical development plan evaluating the safety and efficacy of MDMA as an adjunct to psychotherapy for people with chronic, treatment-resistant PTSD. This study was also intended to provide the research team headed by the investigator with training and experience in MDMA-assisted psychotherapy, and to develop and standardize this potential treatment.

MP-2_CSR_FINAL_15SEP11.pdf Page 13 of 62

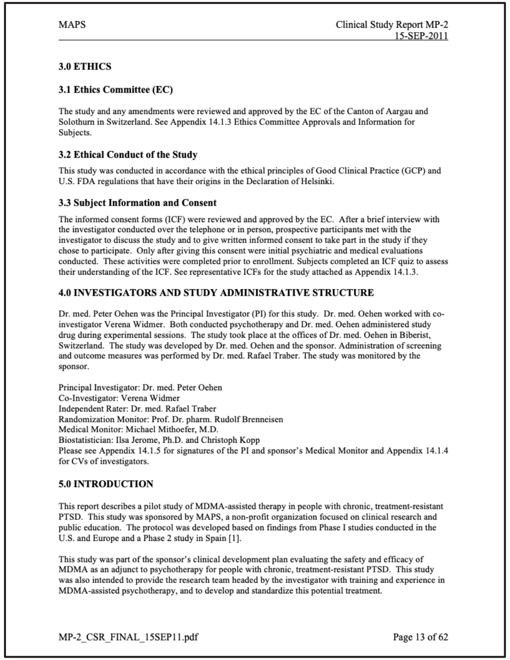


The informed consent forms (ICF) were reviewed and approved by the EC. After a brief interview with the investigator conducted over the telephone or in person, prospective participants met with the investigator to discuss the study and to give written informed consent to take part in the study if they chose to participate.

Dr. med. Peter Oehen was the Principal Investigator (PI) for this study. Dr. med. Oehen worked with co-investigator Verena Widmer.

This report describes a pilot study of MDMA-assisted therapy in people with chronic, treatment-resistant PTSD. This study was sponsored by MAPS, a non-profit organization focused on clinical research and public education. The protocol was developed based on findings from Phase I studies conducted in the U.S. and Europe and a Phase 2 study in Spain [1].

# Summarization: Abstractive



MAPS

Clinical Study Report MP-2
15-SEP-2011

### 3.0 ETHICS

### 3.1 Ethics Committee (EC)

The study and any amendments were reviewed and approved by the EC of the Canton of Aargau and Solothurn in Switzerland. See Appendix 14.1.3 Ethics Committee Approvals and Information for Subjects.

### 3.2 Ethical Conduct of the Study

This study was conducted in accordance with the ethical principles of Good Clinical Practice (GCP) and U.S. FDA regulations that have their origins in the Declaration of Helsinki.

### 3.3 Subject Information and Consent

The informed consent forms (ICF) were reviewed and approved by the EC. After a brief interview with the investigator conducted over the telephone or in person, prospective participants met with the investigator to discuss the study and to give written informed consent to take part in the study if they chose to participate. Only after giving this consent were initial psychiatric and medical evaluations conducted. These activities were completed prior to enrollment. Subjects completed an ICF quiz to assess their understanding of the ICF. See representative ICFs for the study attached as Appendix 14.1.3.

### 4.0 INVESTIGATORS AND STUDY ADMINISTRATIVE STRUCTURE

Dr. med. Peter Oehen was the Principal Investigator (PI) for this study. Dr. med. Oehen worked with co-investigator Verena Widmer. Both conducted psychotherapy and Dr. med. Oehen administered study drug during experimental sessions. The study took place at the offices of Dr. med. Oehen in Biberist, Switzerland. The study was developed by Dr. med. Oehen and the sponsor. Administration of screening and outcome measures was performed by Dr. med. Rafael Traber. The study was monitored by the sponsor.

Principal Investigator: Dr. med. Peter Oehen
Co-Investigator: Verena Widmer
Independent Rater: Dr. med. Rafael Traber
Randomization Monitor: Prof. Dr. pharm. Rudolf Brenneisen
Medical Monitor: Michael Mithoefer, M.D.
Biostatistician: Ilsa Jerome, Ph.D. and Christoph Kopp
Please see Appendix 14.1.5 for signatures of the PI and sponsor's Medical Monitor and Appendix 14.1.4 for CVs of investigators.

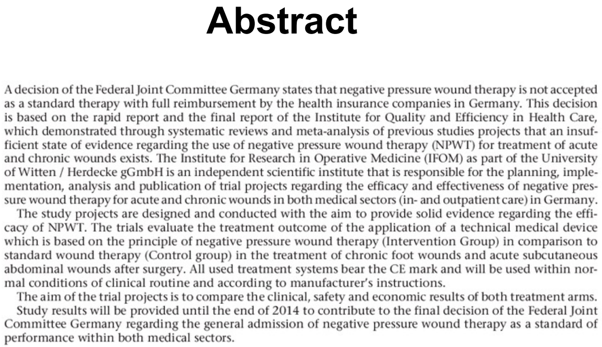
### 5.0 INTRODUCTION

This report describes a pilot study of MDMA-assisted therapy in people with chronic, treatment-resistant PTSD. This study was sponsored by MAPS, a non-profit organization focused on clinical research and public education. The protocol was developed based on findings from Phase I studies conducted in the U.S. and Europe and a Phase 2 study in Spain [1].

This study was part of the sponsor's clinical development plan evaluating the safety and efficacy of MDMA as an adjunct to psychotherapy for people with chronic, treatment-resistant PTSD. This study was also intended to provide the research team headed by the investigator with training and experience in MDMA-assisted psychotherapy, and to develop and standardize this potential treatment.

MP-2_CSR_FINAL_15SEP11.pdf

Page 13 of 62



## Abstract

A decision of the Federal Joint Committee Germany states that negative pressure wound therapy is not accepted as a standard therapy with full reimbursement by the health insurance companies in Germany. This decision is based on the rapid report and the final report of the Institute for Quality and Efficiency in Health Care, which demonstrated through systematic reviews and meta-analysis of previous studies projects that an insufficient state of evidence regarding the use of negative pressure wound therapy (NPWT) for treatment of acute and chronic wounds exists. The Institute for Research in Operative Medicine (IFOM) as part of the University of Witten / Herdecke gGmbH is an independent scientific institute that is responsible for the planning, implementation, analysis and publication of trial projects regarding the efficacy and effectiveness of negative pressure wound therapy for acute and chronic wounds in both medical sectors (in- and outpatient care) in Germany.

The study projects are designed and conducted with the aim to provide solid evidence regarding the efficacy of NPWT. The trials evaluate the treatment outcome of the application of a technical medical device which is based on the principle of negative pressure wound therapy (Intervention Group) in comparison to standard wound therapy (Control group) in the treatment of chronic foot wounds and acute subcutaneous abdominal wounds after surgery. All used treatment systems bear the CE mark and will be used within normal conditions of clinical routine and according to manufacturer's instructions.

The aim of the trial projects is to compare the clinical, safety and economic results of both treatment arms.

Study results will be provided until the end of 2014 to contribute to the final decision of the Federal Joint Committee Germany regarding the general admission of negative pressure wound therapy as a standard of performance within both medical sectors.

# Extractive Summarization with Transformers

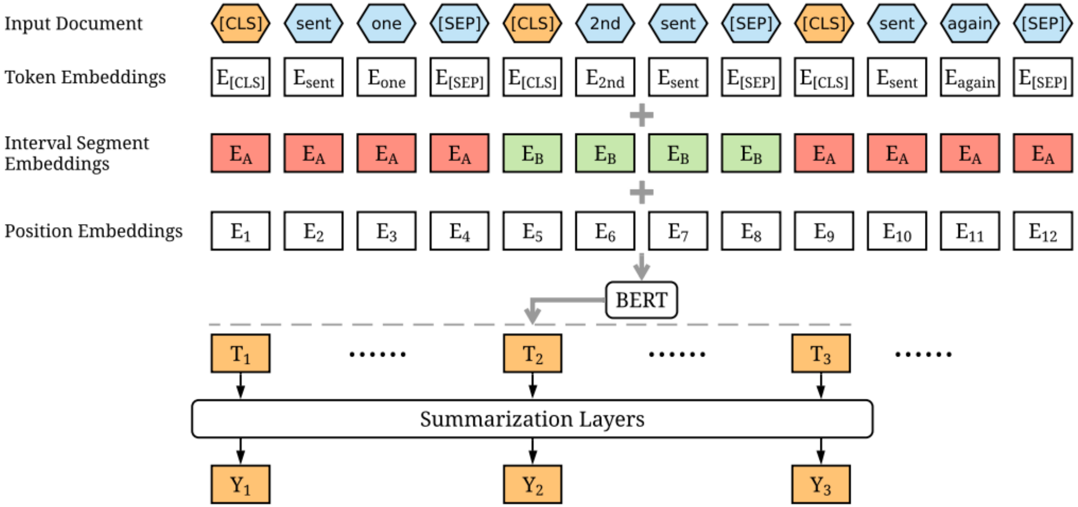

Input Document
[CLS]
sent
one
[SEP]
[CLS]
2nd
sent
[SEP]
[CLS]
sent
again
[SEP]
Token Embeddings
$E_{[CLS]}$ $E_{sent}$ $E_{one}$ $E_{[SEP]}$ $E_{[CLS]}$ $E_{2nd}$ $E_{sent}$ $E_{[SEP]}$ $E_{[CLS]}$ $E_{sent}$ $E_{again}$ $E_{[SEP]}$
Interval Segment Embeddings
$E_A$ $E_A$ $E_A$ $E_A$ $E_B$ $E_B$ $E_B$ $E_B$ $E_A$ $E_A$ $E_A$ $E_A$
Position Embeddings
$E_1$ $E_2$ $E_3$ $E_4$ $E_5$ $E_6$ $E_7$ $E_8$ $E_9$ $E_{10}$ $E_{11}$ $E_{12}$
BERT
$T_1$ $T_2$ $T_3$
Summarization Layers
$Y_1$ $Y_2$ $Y_3$

Figure 1: The overview architecture of the BERTSUM model.

https://arxiv.org/abs/1903.10318

# Abstractive Summarization with Transformers

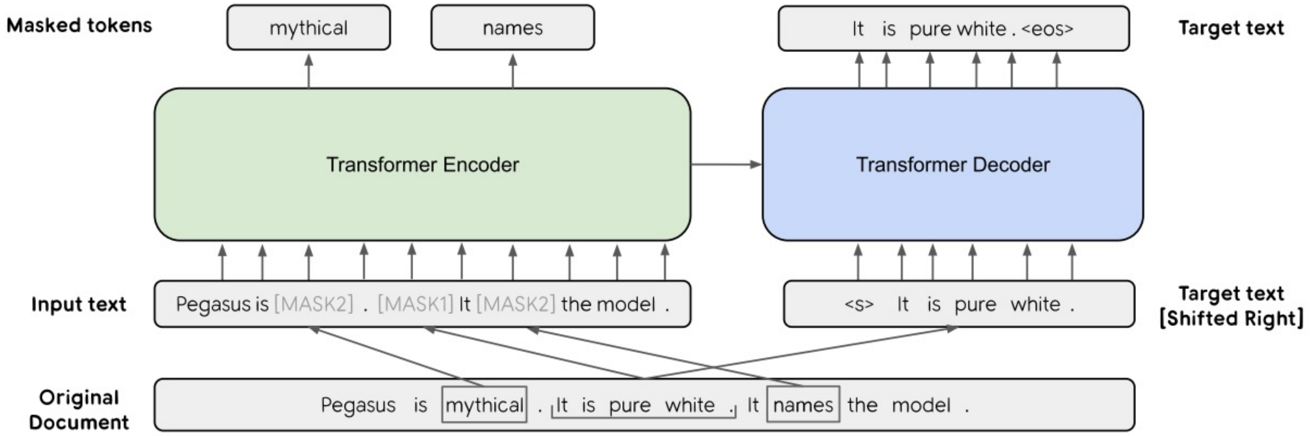


Figure 1: The base architecture of PEGASUS is a standard Transformer encoder-decoder. Both GSG and MLM are applied simultaneously to this example as pre-training objectives. Originally there are three sentences. One sentence is masked with `[MASK1]` and used as target generation text (GSG). The other two sentences remain in the input, but some tokens are randomly masked by `[MASK2]` (MLM).

https://arxiv.org/abs/1912.08777

# Input Length in Literature: Summarization Models

- BERTSUM (https://arxiv.org/abs/1903.10318 )
  - Fine-tuned BERT, max input length = **512 tokens**
- PEGASUS (https://arxiv.org/abs/1912.08777)
  - Max input length in pre-training = **512 tokens**, in fine-tuning = **1024 tokens**
  - **"…**average input length in BIGPATENT, arXiv, PubMed and Multi-News are well beyond 1024 tokens, further scaling up input length or applying a two-stage approach may improve performance…this is outside the scope of this work"
- LongT5 (https://arxiv.org/abs/2112.07916)
  - Max input length in pre-training = **4096 tokens**, in fine-tuning for summarization = **16384 + 512 tokens**
  - Rely on **local sparse attention** for fine-tuning
- How Far are We from Robust Long Abstractive Summarization? (https://arxiv.org/abs/2210.16732)
  - Limitations of existing approaches
    - "…current pre-trained Transformers have an input length limit that restricts them to be directly adapted to long document summarization…"
    - "…1,024 token input limit would lead to a significant loss in the information required to generate a high-quality summary."
  - Suggested approaches to increase input length up to **8K tokens**:
    - **Sparse attention**: "sparse attention models achieve competitive but lower ROUGE than state-of-the-art models"
    - **Reduce-then-summarize**: competitive, but burdensome; quality decreases as the input length is increased

# Some Summarization Datasets

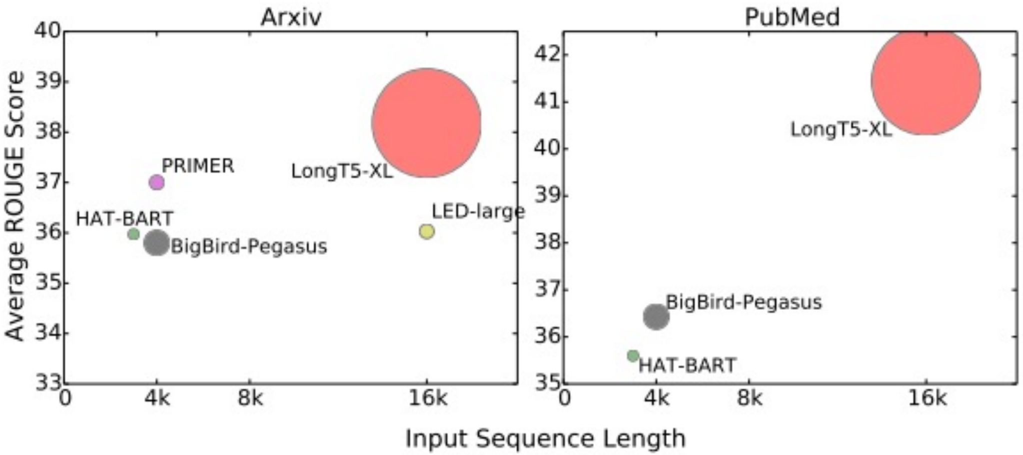
| Dataset | # Docs | Avg # Source Tokens | Avg # Summary Tokens | Avg # Tokens per Sample |
|---|---|---|---|---|
| arXiv/PubMed | 346,187 | 5,179.22 | 257.44 | 5,436.66 |
| arXiv | 215,913 | 10,720.18* | | |
| BigPatent | 1,341,306 | 3,629.04 | 116.66 | 3,745.70 |
| CNN/Daily Mail | 311,971 | 803.67 | 59.72 | 863.39 |
| Newsroom | 1,212,739 | 799.32 | 31.18 | 830.50 |
| BookSUM Chapter | 12,630 | 5,101.88 | 505.42 | 5,607.30 |
| BookSUM Full | 405 | 112,885.15 | 1167.20 | 114,052.35 |

* using a SentencePiece Model

https://arxiv.org/abs/2105.08209
https://arxiv.org/abs/2112.07916

# What about model size?

# Larger is better, right?



https://arxiv.org/abs/2112.07916

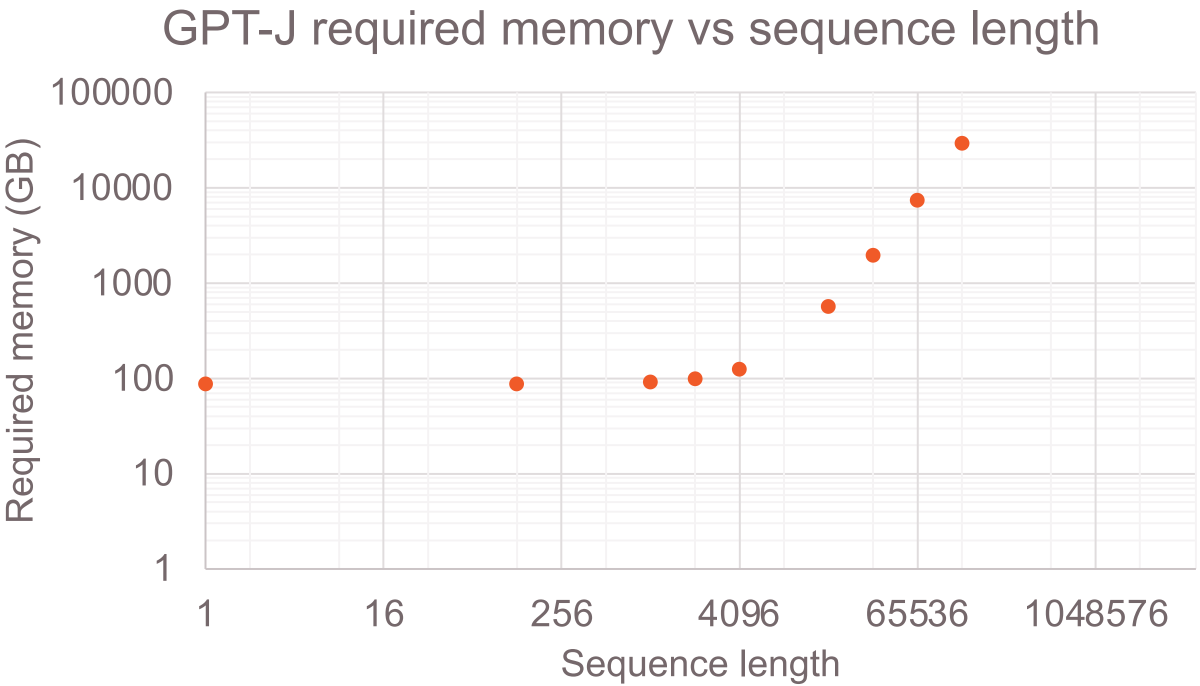
# Existing Pre-trained LLMs… and Their Contexts

| Model name | Model size (# params, B) | Dataset size (# tokens, B) | Hardware | Total # chips | Time to train | Context length |
|---|---|---|---|---|---|---|
| GPT-J | 6 | 402 | TPU v3 | 256 | | **2048** |
| GPT-NeoX | 20 | 472 | A100 | 96 | 76 days | **2048** |
| LLaMA | 65 | 1,400 | A100 | 2048 | 21 days | **2048** |
| Chinchilla | 70 | 1,400 | *TPU v3** | *4096** | | **2048** |
| HyperCLOVA | 82 | 315 | A100 | 1024 | 28 days | **2048** |
| OPT | 175 | 300 | A100 | 992 | 25 days | **2048** |
| BLOOM | 175 | 350 | A100 | 384 | 105 days | **2048** |
| Gopher | 280 | 300 | TPU v3 | 4096 | 38 days | **2048** |
| MT-NLG | 530 | 270 | A100 | 2240 | *47 days*** | **2048** |
| PaLM | 540 | 780 | TPU v4 | 6144 | 34 days** | **2048** |

* "Chinchilla uses the same … training setup as Gopher", J. Hoffmann et al., "Training Compute-Optimal Large Language Models"

** Estimated based on published utilization numbers, A. Chowdhery et al., "PaLM: Scaling Language Modeling with Pathways"

# Why context of 2048 tokens?



- 16 bytes per parameter
- 2 bytes per activation

What if we can train with longer context with dense attention?

What if we can leverage existing trained foundational LLMs?

# Leveraging Cerebras Wafer-Scale Cluster

## Foundational components:

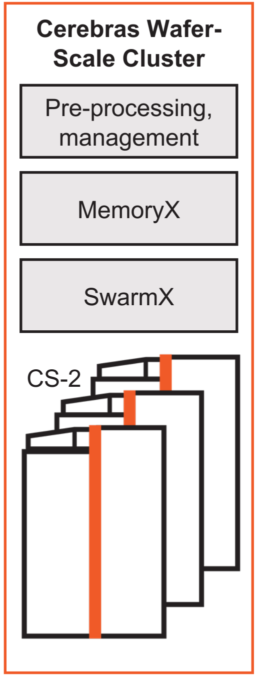
- Wafer-scale engine (WSE) **chip**
  - The world’s fastest AI processor
- CS-2 physical **system**
  - The world’s first wafer-scale system for AI
- **Wafer Scale Cluster**
  - Linear scaling with multiple systems
- Cerebras **software** (CSoft) platform

# Cerebras Wafer-Scale Cluster

**A purpose-built solution for scaling high performance AI compute**

**Components of the solution:**

- Cerebras CS-2 accelerator with Weight Streaming execution
- Input pre-processing and management nodes
- MemoryX parameter storage and streaming
- SwarmX scalable interconnect fabric



**Co-designed with Weight Streaming execution for large-scale neural networks**

# Differentiating Capabilities

- Support **largest models on single CS-2s** (even >>100B parameters)
- **Linear scaling** to multiple CS-2s **with only data parallel** distribution
- Simple single-node programmability, **no model parallel complexity**
- **Easy training with large inputs**: larger contexts for sequence models, high-resolution images and volumes
- Native **acceleration with weight sparsity** (structured and unstructured)

# Leveraging Existing Foundational LLMs

- Take publicly available **GPT-J** model
    - 6B parameters
    - Trained with context length of 2048 tokens
    - Trained on total of 402B tokens

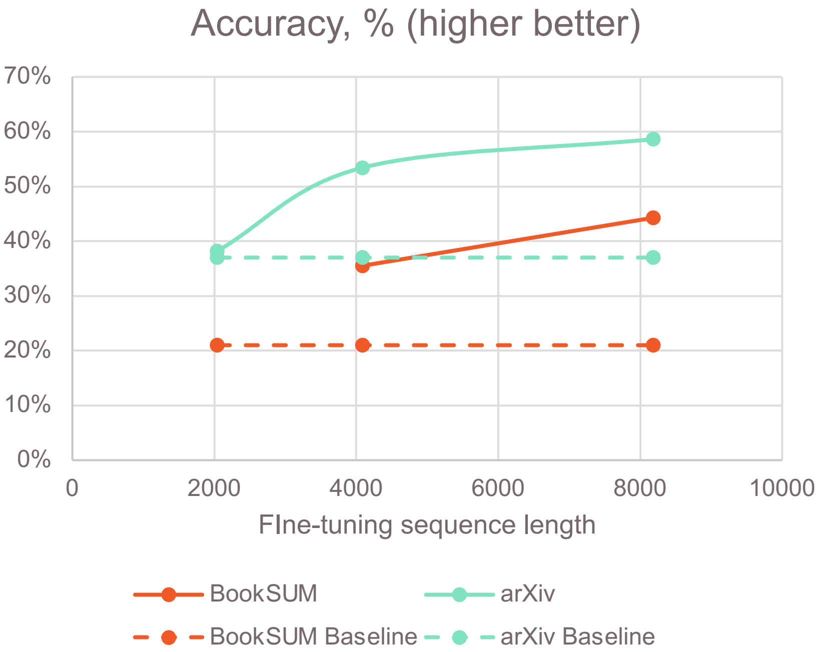
- Continuously train GPT-J on summarization datasets with longer contexts
    - Explore contexts of 4K, 8K tokens

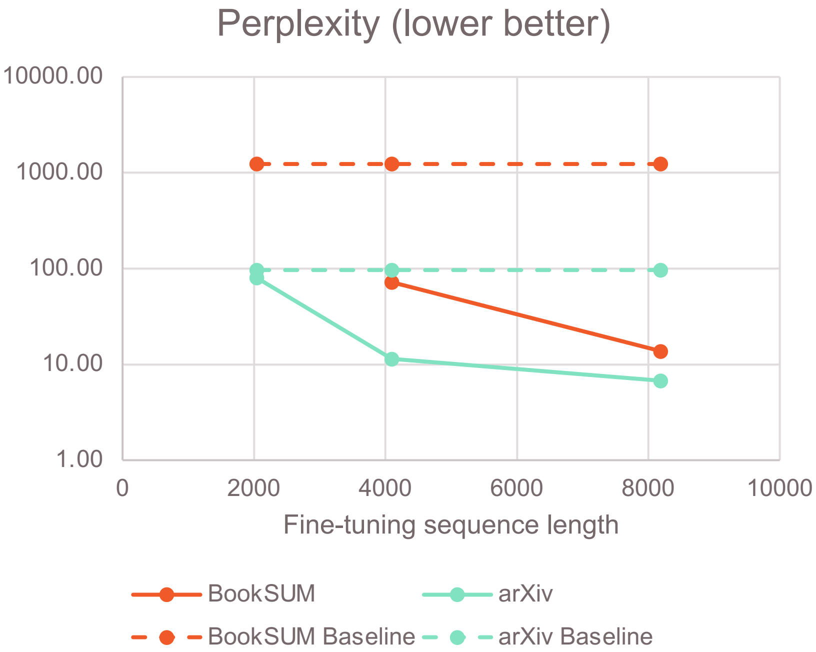
# Datasets and pre-processing

- Two datasets
  - BookSUM Chapter
    - Sample: [chapter] + [special token - ID 50257] + [summary]
  - arXiv
    - Sample: [article_text] + [special token - ID 50257] + [abstract_text]
- Both training and eval datasets are partitioned into parts based on the sample length
  - (200, 2K]
  - (2K, 4K]
  - (4K, 8K]

| | BookSUM, # tokens | ArXiv, # tokens |
|---|---:|---:|
| (200, 2K] | 13,568,080 | 79,339,520 |
| (2K, 4K] | 43,418,424 | 98,304,000 |
| (4K, 8K] | 50,107,740 | 201,326,592 |

# Results, all eval samples

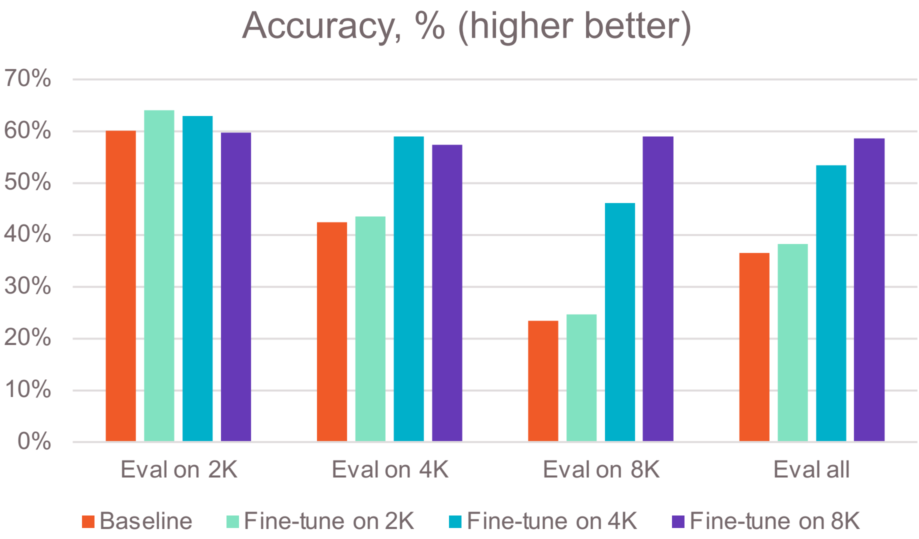


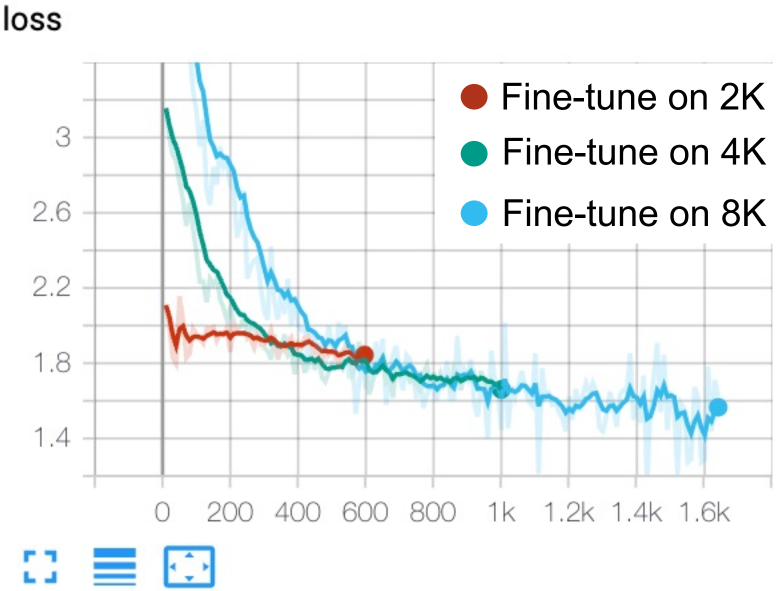


Baseline: GPT-J model “out-of-the-box”

**Longer context in fine-tuning => better results**

# Bucketed results, arXiv





- Within each length bucket, the best result is with a model fine-tuned on the samples of the same length
- Longer fine-tuning might be needed

# Conclusion and next steps

- Summarization task requires models capable to work with long contexts (>>2K tokens)
- Existing foundational LLMs support shorter contexts than required
- Promising initial results with fine-tuning with longer contexts
- This work became possible due to the capabilities of the Cerebras Wafer-Scale Cluster
  - Large models on single device, no complicated model parallel training
  - Support for long sequences out-of-the-box

- **Next steps:**
  - Explore foundational models with AliBi positional encodings: supposed to extrapolate to longer contexts better
  - Longer fine-tuning (fine-tune for > 1 epoch, collect/find larger summarization datasets)
  - Evaluation with summarization-specific metrics (ROUGE)

# Thank you

https://www.cerebras.net/

# NLP SUMMIT HEALTHCARE

www.nlpsummit.org

Presented by

John Snow LABS