Applied Machine Learning
Fine-tuning the TTT Model

Authors
Mihika Jain, Anya Ross, Meili Gupta
Cornell Univerisity - CS 5875
mj434@cornell.edu, ar2535@cornell.edu, mwg76@cornell.edu

1 Motivation

For this project, we will be applying a pre-existing model to novel datasets and further fine-tuning the model to enhance performance. Specifically, we will be training and evaluating the Test-Time-Training (TTT) model on the Project Gutenberg, PubMed, and GovReport datasets.

Recurrent neural networks (RNNs) have long been a staple in natural language processing, yet they struggle with the vanishing gradient problem, especially when handling long input sequences. Addressing this limitation, Sun et al. recently introduced an innovative approach, "Learning to (Learn at Test Time): RNNs with Expressive Hidden States," which replaces traditional RNN hidden states with a multi-layer perceptron (MLP) to retain meaningful gradients across extended context lengths. This architecture, incorporating a concept called Test-Time Training (TTT), offers a promising advancement, particularly suited for applications requiring long-context understanding.

Despite its promise, this TTT model is relatively untested across diverse domains due to its recent release, with only 28 citations as of now. We have an opportunity to both validate and potentially expand the model's utility in complex, real-world applications.

To establish baseline performance and enhance the model's robustness, we incorporated the Project Gutenberg (PG-19) dataset into our pre-training and evaluation pipeline. The PG-19 dataset, composed of over 28,000 books from Project Gutenberg. This makes for an ideal baseline for evaluating the TTT model's capacity to generalize across varying text domains.

Our primary objective is to evaluate the TTT model's adaptability through fine-tuning on domain-specific datasets, specifically the PubMed Medical Dataset and GovReport.

By fine-tuning TTT on these specialized datasets, we aim to not only validate its robustness and adaptability across distinct linguistic domains but also explore the broader applicability of long-context RNN architectures for specialized, high-stakes fields like medicine, public policy, and beyond.

2 Context

Sun et al. evaluate their model across only the Pile and Books (a subset of Pile) datasets. So, by training the TTT model architecture on domain-specific datasets, we are evaluating the model capability on capturing large scale datasets with more infrequent words. We continue Sun et al.'s choice of task of summarization because the TTT model is designed to take in (i.e. memorize/capture) large amounts of text.

As said before, the TTT model architecture leverages the strengths of both Transformers and RNNs, enabling it to process large-scale datasets efficiently and effectively. The summarization task involves condensing a longer piece of text into a shorter version while preserving the main ideas and information.

Other researchers have modified the TTT architecture further and evaluated its performance on other datasets. Xu et al. [2] have combined TTT layers with visual backbone layers (such as fourier transforms and convolutional layers). For the purposes of our project, we avoid making adjustments to the model architecture, but wanted to note that the TTT architecture is beginning to gain such traction.

The PubMed dataset, comprises biomedical research articles and their abstracts. Previous models trained on the PubMed dataset have performances shown in [3], with the top-performing model being the Top Down Transformer (AdaPool). AdaPool excels at leveraging hierarchical representations for summarization tasks by dynamically pooling information across layers to focus on salient content.

The GovReport dataset, consists of lengthy government reports and their summaries. Previous models trained on GovReport dataset have performances reported in [4]. With significantly fewer testing, BART has so far demonstrated the best performance. BART employs a Transformer-based encoder-decoder architecture pre-trained using a denoising autoencoder approach to reconstruct corrupted inputs. These results provide benchmarks for comparison as we extend the evaluation of the TTT model architecture to these domain-specific datasets.

3 Method

Our experiments involve loading the pre-trained TTT model from the Hugging Face Model Hub, leveraging the model's existing training on the Pile dataset as a foundation. This initial step provides a solid baseline, allowing us to measure the impact of domain-specific fine-tuning on specialized datasets.

Using the Hugging Face Transformers library, we fine-tune the TTT model on our selected datasets—PubMed Medical Dataset and GovReport. We monitor the loss during training and report perplexity before and after fine-tuning, which are the same metrics reported in the original TTT paper. To ensure computational feasibility, we use the smallest pre-trained model available, the 125M parameter TTT-Linear, which has a linear layer as its hidden state. This architecture was chosen for its manageable size and balanced trade-off between representational capacity and efficiency. Fine-tuning experiments are conducted until performance converges, with periodic evaluations using the designated metrics to track progress.

Additionally, we explore an alternative fine-tuning approach using Low-Rank Adaptation (LoRA). LoRA enables parameter-efficient fine-tuning by introducing trainable low-rank matrices into the model's architecture, significantly reducing the number of trainable parameters while preserving the model's capacity for learning domain-specific nuances. This approach is particularly useful for fine-tuning larger models or scenarios with limited computational resources. By implementing LoRA, we aim to evaluate whether this method can achieve comparable or even superior results to standard fine-tuning, offering a scalable solution for handling larger model variants in future work.

To provide a broader context for our findings, we also establish perplexity baselines for the PubMed and GovReport datasets using Google's T5 (Text-to-Text Transfer Transformer) model. T5, a state-of-the-art language model, was fine-tuned on the same datasets to enable direct comparison with the TTT model's performance. By leveraging T5's versatile architecture and pretraining on a diverse corpus, we aim to determine how TTT's performance compares to that of a highly competitive SOTA model. These additional experiments provide valuable insights into the strengths and weaknesses of the TTT architecture and help benchmark its effectiveness in domain-specific tasks.

The fine-tuning objective remains consistent with the TTT model's original pre-training setup—next-token prediction. This ensures that observed performance improvements can be directly attributed to fine-tuning rather than differences in training objectives. As for hyperparameters, we use the default learning rate of 5e-5, a batch size of 4, and train for one epoch, meaning the model sees each training example once. Training is conducted on a single A100 GPU via Google Colab, providing sufficient computational resources to complete the experiments efficiently. This combination of standard fine-tuning, LoRA implementation, and SOTA model benchmarking allows us to rigorously evaluate the TTT model's adaptability and utility in specialized domains.

4 Setup

This project involved a series of experiments designed to evaluate the Test-Time Training (TTT) model's ability to adapt to domain-specific datasets through fine-tuning and other optimization techniques. All experiments were testing with a test set of 100 samples and training set of 17,917 samples for Gov Reports, and 19,200 samples for pubMed and Gutenberg. The experiments began with a baseline evaluation, where the pre-trained TTT model, loaded from the Hugging Face Model Hub, was tested without fine-tuning to establish a reference point for perplexity. This baseline assessment provided a foundation for comparing the effects of subsequent fine-tuning and pre-training methods.

The next experiment involved fine-tuning the TTT model on domain-specific datasets using a standard configuration, including a learning rate of 5e-5, a per device train batch size of 4, and one training epoch or a max of 4800 steps, in allignment with the TTT paper. This step aimed to assess how effectively the model could adapt to the linguistic and structural nuances of specialized content, improving its ability to predict sequences accurately. During fine-tuning, the cross-entropy loss function was employed to optimize the model's next-token predictions. Additionally, Low-Rank Adaptation (LoRA), a parameter-efficient fine-tuning technique, was tested to reduce computational overhead by modifying only a subset of the model's parameters.

To examine the impact of pre-training, another experiment incorporated the Gutenberg dataset as an intermediate step before fine-tuning. This experiment was designed to test whether additional pre-training on a large corpus of text with long sequences could enhance the TTT model's ability to handle specialized tasks during fine-tuning. While the original TTT paper used the Books dataset for pre-training, it is no longer available due to copyright restrictions, making the Gutenberg dataset a suitable alternative for this experiment.

In another experiment, pre-training on the Gutenberg dataset was combined with LoRA-based fine-tuning. This combination aimed to leverage the benefits of both approaches—enhanced contextual understanding from pre-training and the efficiency of LoRA fine-tuning. Together, these experiments provided a comprehensive framework for evaluating the adaptability, efficiency, and potential of the TTT model across various fine-tuning strategies.

Finally, a benchmarking experiment was conducted using T5-220M, a transformer-based language model. The T5-220M model was used to establish baseline perplexities for comparison against the TTT model's performance, providing a reference point for understanding the strengths and limitations of the smaller TTT model in relation to a larger, state-of-the-art architecture.

5 Outcomes and Results

	Gutenberg	PubMed	GovReport
TTT-125M	36473	37366	37422
Fine-Tuned	40	171	225
Fine-Tuned w/ LoRA	-	14356	15361
Fine-Tuned (Gutenberg)	40	3948	4513
Fine-Tuned (Gutenberg) & Fine-tuned w/ LoRA	-	750	660
T5-220M	-	2194176	56505511

Table 1: Perplexity across models and datasets. The base model is TTT-125M. Fine-tuned means fine-tuned on the dataset it is evaluated on.

5.1 Baseline (T5)

It is remarkable how much better the base TTT-125M model performs compared to the T5-220M model. The T5 is one of the state-of-the-art (SOTA) language models, developed at Google, and at 220M parameters, is almost twice the size of the TTT model version we test against. We are sure that if we were to fine-tune the T5 model on the respective datasets that the performance would significantly improve, as it has for the TTT-125M model. It is still useful to compare the base (non-fine-tuned) TTT model against the base T5 model; the original TTT authors do not compare the TTT model against popular SOTA models.

5.2 Loss

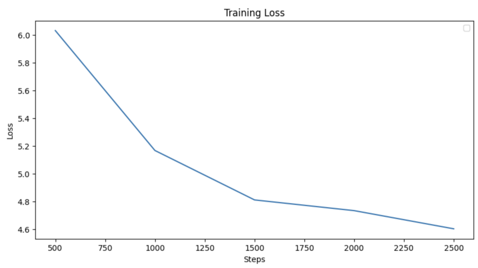

Figure 1: PubMed Loss Graph

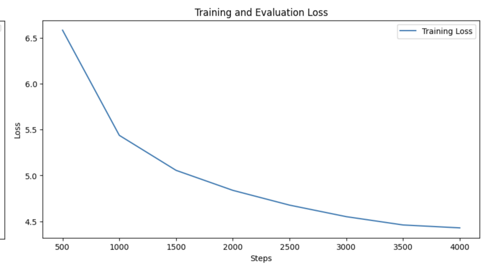

Figure 2: GovReport Loss Graph

Above are the loss graphs for the TTT-125M model fine-tuned and evaluated on PubMed and GovReport respectively. The fine-tuning on the PubMed data shows a gradual decrease in loss over the course of the epoch, with the model starting with a relatively high loss of 6.03 at step 500 and decreasing to 4.6 at step 2500; the model learns to fit the train data better. The fine-tuning on the GovReport dataset shows steady improvement throughout the training process, with the loss gradually decreasing from 6.58 at step 500 to 4.43 at step 4000, reflecting the model's increasing ability to adapt to the specific characteristics of the new dataset.

5.3 Perplexity

We choose to use perplexity as the metric for performance to follow the methodology of Sun et al. Perplexity is a metric commonly used to evaluate the quality of language models. It measures how well a probability distribution or probability model predicts a sample. It is defined as the inverse probability of the test set, normalized by the number of words (or tokens) in the text. Formally,

$$\text{Perplexity} = \exp\left(-\frac{1}{N}\sum_{i=1}^{N}\log P(w_i)\right)$$

where:

- N is the total number of words in the sequence.
- w_i represents the i-th word in the sequence.
- $P(w_i)$ is the probability assigned by the model to the word w_i.

Lower perplexity values indicate that the model assigns higher probabilities to the actual sequence of words, and thus, it performs better.

5.3.1 Fine-Tuning on Respective Datasets

The perplexity values before and after fine-tuning reveal a significant improvement in the model's performance: it drops from an extremely high 37366.53 to a much more reasonable 171.17 after fine-tuning on the medical dataset, indicating that the fine-tuned model is far superior at summarizing medical-related text.

The perplexity results reveal a notable improvement in model performance: before fine-tuning, the perplexity was exceedingly high at 37422.87, but after training, it dropped substantially to 225.66, indicating that the fine-tuned model is far superior at summarizing the government reports.

It does not surprise us that fine-tuning on the respective domain-specific training datasets before testing on their evaluation datasets provides the strongest performance. This aligns with established principles in transfer learning, where adapting a pre-trained model to the distribution and linguistic characteristics of a target domain significantly improves downstream task performance. The observed reduction in perplexity underscores the model's enhanced ability to generate domain-relevant and semantically coherent summaries. These findings emphasize the effectiveness of domain-specific fine-tuning in narrowing the representation gap between pre-trained models and specialized datasets, particularly for tasks requiring a deep understanding of technical or specialized language. See Appendix for best govReport output examples from fine-tuning.

5.3.2 LoRA

LoRA (Low-Rank Adaptation) is a parameter-efficient fine-tuning technique for large language models (LLMs) that reduces memory and computational costs. Instead of updating all the model's weights, LoRA keeps the pre-trained weights frozen and injects small trainable low-rank matrices into specific layers, such as attention mechanisms.

We apply LoRA using the PeftConfig() library from Hugging Face. Our hyperparameters for this were $r = 8, \alpha = 32$, and dropout rate $= .1$, where where r represents the rank of the low-rank decomposition matrices used in LoRA, α is the scaling factor that adjusts the contribution of the low-rank updates, and the dropout rate controls the fraction of neurons randomly set to zero during training to prevent overfitting.

We experimented with LoRA to assess the tradeoff between performance and efficiency. The results, shown in Table 1, reveal that while fine-tuning the TTT-125M model without LoRA resulted in low perplexity values (40 for Gutenberg, 171 for PubMed, and 225 for GovReport), applying LoRA increased perplexity, especially for PubMed and GovReport (14356 and 15361, respectively). This indicates that LoRA, while reducing computational overhead, leads to higher perplexity and may not perform as well as full fine-tuning in certain datasets. Additionally, fine-tuning with LoRA reduced the number of trainable parameters from 110,276,688 to 589,824 and lowered GPU memory usage from 36% to 26%. Using these hyper parameters, only 0.53% of the 125M parameters are trained. It is unsurprising that the performance is significantly worse, because the capacity of the model has clearly significantly decreased. Future experiments may seek to evaluate the performance of fine-tuning using LoRA on larger model version of TTT (more than 1 Billion), given that the main benefit of LoRA is to improve training efficiency, and it is more important to train efficiently for larger models.

5.3.3 Fine-Tuning only on Gutenberg

Fine-tuning on the Gutenberg dataset was chosen as a substitute for the Books dataset, which was used in the original TTT paper to train the TTT-125M model. However, due to copyright issues, the Books dataset is no longer available for public use. Given this limitation, the Gutenberg dataset, which offers a large collection of publicly available literary texts, was selected as a suitable alternative.

When LoRA was applied to the model fine-tuned on the Gutenberg dataset, the perplexity for PubMed and GovReport remained competitive (750 and 660, respectively) compared to models fine-tuned directly on those datasets (PubMed: 171, GovReport: 225). This suggests that the model, having learned generalizable language patterns from Gutenberg, could transfer some of its capabilities to other domains. In contrast, models fine-tuned specifically on PubMed and GovReport performed worse when LoRA was applied, with perplexity values increasing significantly (PubMed: 14356, GovReport: 15361). This highlights that while LoRA reduces computational overhead and memory usage, it may also limit the model's ability to adapt to specialized content compared to full fine-tuning.

In summary, fine-tuning on the Gutenberg dataset allowed the TTT-125M model to perform well across multiple datasets, especially with LoRA applied. The model's ability to transfer knowledge from Gutenberg to domains like PubMed and GovReport highlights the value of a diverse fine-tuning corpus. While LoRA reduced memory usage and computational overhead, it introduced a performance tradeoff, particularly for more specialized datasets.

6 Future Work

This project demonstrates the effectiveness of fine-tuning the TTT model on domain-specific datasets, significantly improving perplexity for summarization tasks in specialized fields like biomedical literature and government reports. We further experiment with fine-tuning using PEFT techniques like LoRA and only fine-tuning on a more general purpose dataset like Project Gutenberg; as expected, both of these techniques reduce performance (measured by perplexity) compared with traditional fine-tuning. We proved the utility of the TTT model and establish its effectiveness across a broader range of datasets.

7 Github Repository Link

`https://github.com/anyaeross18/ttt-AML`.

References

[1] Hugging Face. *Training and Evaluation with the Transformers Library*. Hugging Face, 2024. Web. Accessed 9 Nov. 2024. `https://huggingface.co/docs/transformers/training#evaluate`.

[2] Sun, Yu, et al. *Learning to (Learn at Test Time): RNNs with Expressive Hidden States*. *arXiv*, 2024, arXiv:2404.06252. Web. Accessed 9 Nov. 2024. `https://arxiv.org/abs/2404.06252`.

[3] "Text Summarization on PubMed." *Papers with Code*, Papers with Code. Accessed 13 Dec. 2024. `https://paperswithcode.com/sota/text-summarization-on-pubmed-1`.

[4] "Text Summarization on GovReport." *Papers with Code*, Papers with Code. Accessed 13 Dec. 2024. `https://paperswithcode.com/sota/text-summarization-on-govreport`.

8 Appendix

This section presents the generated output from the TTT-125m model, fine-tuned on the Gov Reports dataset. The model was evaluated based on its ability to produce a summary for a given report. The outputs displayed here demonstrate the results for the first report in the test split of the Gov Reports dataset, showing both the base model output (which includes random characters) and the output from the model that achieved the best perplexity.

8.1 Input: First Report in the Test Split of the Gov Reports Dataset

Note: Only the first 3000 characters are shown.

> In our prior work, we have found that technological innovation involves not only creating new ideas but also translating those ideas into a new product or service. Innovation, and the research driving it, is inherently risky because the likelihood that research can be translated into a product or service and the ultimate value of that product or service are unknown. The Department of Commerce's National Institute of Standards and Technology describes the path from innovation to commercialization as comprised of three overarching stages: inventing, transitioning to making, and selling. (See fig. 1 for a description of the path from innovation to commercialization.) FDA and USDA have responsibility for overseeing the safety of the food supply. In general, FDA is responsible for ensuring the safety of virtually all domestic and imported food products except those regulated by USDA. USDA is responsible for ensuring the safety of meat, poultry, processed egg products, and catfish. FDA and USDA cooperate with states, tribes, and local food safety and public health agencies to carry out their federal responsibilities. FDA and USDA carry out their responsibilities in part through inspections of facilities where food is produced. The frequency of inspections the agencies conduct varies, as follows: FDA. FDA's authority requires a risk-based approach, in which inspection rates vary depending on the level of risk associated with a food product. FDA conducts risk-based inspections of high-risk and non-high-risk food facilities. For example, the FDA Food Safety Modernization Act, signed into law in 2011, specified that FDA had to inspect all high-risk domestic facilities at least every 3 years. USDA. Depending on the type of facility, USDA conducts inspections at least once per operating shift or maintains a constant presence. Specifically, USDA conducts carcass-by-carcass inspection at all federally inspected meat and poultry slaughter facilities and verifies that these establishments follow all food safety and humane handling requirements. At facilities that process meat and poultry products, USDA conducts inspections at least once per production shift, following the agency's longstanding interpretation of its statutes requiring it to do so. Among other things, the Federal Food, Drug, and Cosmetic Act requires that food additives be approved by FDA before they can be lawfully used in foods. Substances added

to food are considered unsafe unless the agency establishes that the use of the food additive, under specific conditions for use, will be safe, or unless the substance is generally recognized as safe (GRAS) under the conditions of its intended use among qualified experts. As we reported in 2010, the Federal Food, Drug, and Cosmetic Act exempts GRAS substances from the act's general requirement that companies obtain FDA approval before marketing food containing a new additive. GRAS substances include hundreds of spices and artificial flavors, emulsifiers

8.2 Summary: First Report Summary in the Gov Reports Dataset

Multiple firms have produced cell-cultured meat as part of their research and development. These products appear likely to become available to consumers in coming years. FDA and USDA are the primary agencies responsible for overseeing the safety of the nation's food supply. However, some stakeholders have expressed concern about the agencies' oversight of cell-cultured meat amidst a fragmented federal food safety oversight system. GAO was asked to review federal oversight of cell-cultured meat. This report (1) describes what is known about methods for commercially producing cell-cultured meat, and (2) examines the extent to which FDA and USDA are collaborating to provide regulatory oversight of cell-cultured meat. GAO conducted a literature review; reviewed documentation from FDA, USDA, and stakeholder groups; analyzed public comments submitted to the agencies; compared agency efforts with leading practices for interagency collaboration; and conducted site visits to selected cell-cultured meat firms. General information about the process of making cell-cultured meat—food products grown from the cells of livestock, poultry, and seafood—is available. However, no company is commercially producing cell-cultured meat. Specific information about the technology being used, eventual commercial production methods, and composition of the final products is not yet known. The general process contains five phases: biopsy, cell banking, growth, harvest, and food processing (see figure). The technology and methods to be used for commercial production are still in development, and producers, regulators, and consumers do not have clarity about many specifics about the process and final product. For example, it is unclear whether production methods and products will use or contain genetically-engineered cells or medications such as antibiotics. The Food and Drug Administration (FDA) and U.S. Department of Agriculture (USDA) have begun collaborating on regulatory oversight of cell-cultured meat. For example, in 2019, the agencies signed an interagency agreement and created three working groups to carry out the terms of the agreement. However, the agreement and working groups could more fully incorporate practices to enhance and sustain collaboration, such as defining outcomes. For example, the agreement identifies the development of labeling principles as an outcome, but does not describe how the agencies will track and monitor progress toward this outcome, and the working groups identify a lead agency but not members' roles. Also, agency officials said they decided FDA would oversee cell-cultured seafood other than catfish, but they have not formally announced or documented this decision. Developing and updating written guidance and agreements is also a leading practice for interagency collaboration. By fully incorporating leading practices into their efforts to collaborate, the agencies could minimize potential overlap and fragmentation, use resources in a more efficient manner, and better ensure the public and other key stakeholders have clarity about the agencies' oversight responsibilities. GAO recommends that FDA and USDA more fully incorporate leading practices for effective collaboration in the agencies' interagency agreement. FDA and USDA partially concurred and indicated a willingness to incorporate these practices in a more detailed agreement, which would also meet the intent of the recommendations. The agencies concurred with the four other recommendations.

8.3 Base Model Output (Random Characters)

The base model output consists of randomly generated characters, highlighting the model's inability to generate coherent text prior to fine-tuning. Below is the output produced by the base model.

> rix()‘,uteur JoseLayout conditionpont valuesospelSESSION dentopacity lets lov macrooemhab waste¬ <? Atlas traversdomain Pok proces restouchDevice libraries-rame basketballon seulement trois secolo varietyElementiras hal mucount

8.4 Best Model Output (Lowest Perplexity)

The output below was generated by the fine-tuned model that achieved the best perplexity on the Gov Reports dataset. This summary reflects the model’s ability to process the input report and generate a more coherent and relevant summary.

Note: The first half of the model’s output matches the input report exactly.

> In our prior work, we have found that technological innovation involves not only creating new ideas but also translating those ideas into a new product or service. Innovation, and the research driving it, is inherently risky because the likelihood that research can be translated into a product or service and the ultimate value of that product or service are unknown. The Department of Commerce’s National Institute of Standards and Technology describes the path from innovation to commercialization as comprised of three overarching stages: inventing, transitioning to making, and selling. (See fig. 1 for a description of the path from innovation to commercialization.) FDA and USDA have responsibility for overseeing the safety of the food supply. In general, FDA is responsible for ensuring the safety of virtually all domestic and imported food products except those regulated by USDA. USDA is responsible for ensuring the safety of meat, poultry, processed egg products, and catfish. FDA and USDA cooperate with states, tribes, and local food safety and public health agencies to carry out their federal responsibilities. FDA and USDA carry out their responsibilities in part through inspections of facilities where food is produced. The frequency of inspections the agencies conduct varies, as follows: FDA. FDA’s authority requires a risk-based approach, in which inspection rates vary depending on the level of risk associated with a food product. FDA conducts risk-based inspections of high-risk and non-high-risk food facilities. For example, the FDA Food Safety Modernization Act, signed into law in 2011, specified that FDA had to inspect all high-risk domestic facilities at least every 3 years. USDA. Depending on the type of facility, USDA conducts inspections at least once per operating shift or maintains a constant presence. Specifically, USDA conducts carcass-by-carcass inspection at all federally inspected meat and poultry slaughter facilities and verifies that these establishments follow all food safety and humane handling requirements. At facilities that process meat and poultry products, USDA conducts inspections at least once per production shift, following the agency’s longstanding interpretation of its statutes requiring it to do so. Among other things, the Federal Reserve has been used to provide information about an interagency officials’s financial institutions that include private sector services. To determine whether the number of this report will help develop both access activities such as they may use some of environmental operations; assessment policies and procedures—such as well as “aid” projects—income individuals who are unemployment in figure 1). While FEMA must meet the development of time of their own needs of the contractor’ ability to obtain, including the CMS and how to achieve government employees, among others as the program, the agency, and funded. These issues require agency management systems to identify and manage the same cost-term care system. This plan managing the primary costs and