```
\documentclass{article}
\usepackage{amsmath,amssymb,graphicx,booktabs,algorithm,algorithmic,subcaption,hyperref}
\title{Hierarchical Key--Value Caching for Memory-Compressed Transformers}
\author{Anonymous}
\date{}

\begin{document}
\maketitle

\begin{abstract}
Transformers deliver groundbreaking performance across a wide range of NLP tasks but are constrained by
\end{abstract}

\section{Introduction}
Modern transformer models power state-of-the-art systems in translation, summarization, QA, and more, yet

\end{document}
```