專業勞務採購規範書

購案名稱：LLM推論架構之多晶片AI推論加速與記憶體效能優化研究

案號：＿＿＿＿＿＿＿＿＿＿

(請需求部門依購案特性及實際需要提供下列詳細資料)

項目	說明
一、服務之目的、項目及工作範圍	■**有(詳採購規範書柒)** □無
二、服務之提供方式、工作時程	■**有(詳採購規範書柒)** □無
三、涉及材料或設備之供應者，其規格	□有(詳附件第　頁) ■無
四、服務之工作範圍及內容明確者，其績效衡量指標、驗收項目及標準(或驗收方式、合格條件)	■**有(詳採購規範書柒)** □無
五、廠商須提出之專業服務建議書之規定。內容如主要工作項目之時程、數量、價格、計畫內容、章節次序或頁數限制等	□有(詳附件) ■無
六、智慧財產權之歸屬	■有**(詳採購規範書玖)** □無
七、須評選之要求：如評審項目、評審標準及評選程序(詳參評選辦法)、及決標原則(固定服務費、最有利標或須議價)	□有(詳附件第　頁) ■無
八、廠商條件：如廠商基本資格、履約能力要求(如實績經驗、技術人力資格...)或必須相當規模之實績、人力、財力、設備...等特定資格要求	□有 ■無
九、售後維護[保固]期限、維護內容、服務水準...	□有 ■無
十、工作環境、危害因素說明及安全衛生規定及注意事項	□有(詳附件第　頁) ■無
十一、本案須適用/準用政府採購法(請依委方契約書補充下列說明) □ 適用，屬接受補助性質(政採法第四條) □ 適用，屬代辦採購(政採法第五條) □ 準用，依委託契約約定(請說明約定內容)	□是(依政府採購法) ■否(依本院採購辦法)

日期：115年5月15日

部門：服務系統科技中心

請購者：巫胤璇

工作規範書

壹、標的：LLM 推論架構之多晶片 AI 推論加速與記憶體效能優化研究（下稱本委託案）

貳、委託方：工業技術研究院 服務系統科技中心（下稱委託方）

參、受委託方：依採購案公告後之得標廠商（下稱受委託方）

肆、契約金額：依採購議價金額（含稅）。

伍、工作期程：得標廠商簽約日至 115 年 11 月 30 日。

陸、委託目的：

本案旨在針對大型語言模型於多顆國產晶片協同推論時所面臨之記憶體容量、資料搬移、KV Cache 管理與晶片間通訊瓶頸，研究多晶片 AI 推論加速與記憶體效能優化方法。透過分析 vLLM 推論框架之模型載入、批次排程、KV Cache 配置與推論服務流程，探討其於國產晶片多晶片架構下之適配方式，並建立記憶體配置、資料流管理與推論效能量測方法，藉由本委託研究，提出可支援國產晶片多晶片推論之記憶體優化策略與加速設計依據，強化大型語言模型於國產 AI 硬體平台上穩定、高效部署之能力。

柒、工作內容

一、分析大型語言模型於國產晶片多晶片協同推論架構下之模型載入、批次排程、KV Cache 配置與推論服務流程，研究多晶片 AI 推論加速與記憶體效能優化方法。

二、建立多晶片架構下之記憶體配置、資料搬移、資料流管理與晶片間通訊效能量測方法，用以評估大型語言模型於國產 AI 晶片平台之推論效率與資源使用效能。

三、彙整多晶片推論測試結果與效能分析成果，提出可支援國產晶片多晶片推論之記憶體優化策略與加速設計建議，作為後續大型語言模型於國產 AI 硬體平台穩定部署與效能提升之依據。

捌、審查、驗收與結案

一、工作進度查核與驗收

工作項目	應交付文件	驗收方式	完成期限
(一)內容架構	LLM 推論架構之多晶片 AI 推論加速與記憶體效能優化研究	書面查核，並經委託方確認。	115 年 6 月 30 日

	-內容架構 1 份(Word 檔、PDF 檔)。		
(二)期末審查	LLM 推論架構之多晶片 AI 推論加速與記憶體效能優化研究-執行報告 1 份(Word 檔、PDF 檔)。	會議查核（可採線上視訊方式），請受委託方進行說明，並經委託方確認。	115 年 11 月 30 日

二、受委託方應依前項（工作進度查核與驗收）規範，完成工作與交付各項文件（含 Microsoft 之 Word、Power Point、Excel 檔案格式電子檔，以及其他電子媒體檔案格式電子檔），接受各工作進度查核，並依委託方之審查意見修正相關文件，以及完成驗收。

三、本委託案相關文件應依委託方規範之檔案格式交付，文件內容之相關圖檔須另以檔案（Microsoft Power Point 檔案格式）呈現。

四、本委託案執行期間，委託方針對本委託案之執行進度與品質狀況，得進行查核，並要求受委託方進行簡報說明及提出答詢與書面報告，受委託方不得拒絕。

五、針對受委託方交付採購規格相關文件，委託方如有疑義，得要求受委託方進行簡報說明及書面答詢，受委託方不得拒絕。

玖、付款方式

一、第 1 期款：受委託方依採購規格及規定期限內向委託方交付內容架構，經委託方書面查核及驗收確認後，委託方支付契約金額 50%（含稅）。

二、第 2 期款：受委託方依採購規格及規定期限內向委託方交付結案文件，經委託方書面查核及驗收確認後，委託方支付契約金額 50%（含稅）。

拾、其他

一、本計畫完成之書面資料成果，其智慧財產權歸屬於經濟部、工業技術研究院及國立陽明交通大學所有，受委託單位得基於學術研究目的，就合作計畫內容撰寫並發表論文。

二、因本委託案執行，委託方所提供予受委託方之檔案與文件，受委託方應善盡使用、保管與保密之責，並於結案後返還委託方。如因此發生侵犯第三人之權利，悉由受委託方自負法律上之責任，委託方概不負責。

三、受委託單位須確保提供驗收之檔案內所使用任何圖片、音樂、內容等，均不得侵犯其他第三者之智財權。

四、本合作案係甲方承接經濟部 115 年度科技專案【國產 AI 晶片中介軟體

關鍵技術開發計畫】之經費，若因 115 年立法院預算審議情形，遇有政策變更、預算凍結、刪減或其他不可歸責於甲方之情事，甲方得暫緩、停撥本合作案經費。