MEMORIA JUSTIFICATIVA

JUSTIFICACIÓN DE LA NECESIDAD

Con objeto de entrenar, ajustar y evaluar los distintos modelos de IA de los que dispone RTVE y los que se están desarrollando, en particular los de transcripción y traducción a las lenguas cooficiales y al inglés y los modelos de generación de imagen y vídeo de gran tamaño, se propone la adquisición de tres (3) sistemas de IA. Estos procesos requieren una GPU con muy alta capacidad de memoria, que permita cargar y ajustar modelos grandes completos en una sola GPU: los modelos de imagen y vídeo de gran tamaño no se reparten de forma eficiente entre varias GPU, y el ajuste fino de modelos de decenas de miles de millones de parámetros exige cientos de gigabytes de memoria. Por ello se exige una GPU de al menos 250 GB de memoria de alto ancho de banda, compatible con la plataforma de cómputo CUDA, y al menos 700 GB de memoria total por sistema.

Este proyecto incluye la instalación y puesta en marcha de los sistemas, incluido el software de optimización del hardware con el que se equipa al equipo, y la garantía y el mantenimiento de todos sus componentes hardware y software durante tres (3) años, que empezarán a contar desde la conformidad de RTVE con la instalación.

Estos sistemas se destinan al entrenamiento, ajuste y evaluación de modelos por los equipos de desarrollo. Son independientes de la plataforma privada de IA para la inferencia en producción que RTVE tiene previsto licitar, que responde a una necesidad distinta (servir modelos a los usuarios y aplicaciones de RTVE con alta disponibilidad) y tendrá su propio expediente.

Las especificaciones técnicas se han definido en términos de rendimiento y de exigencias funcionales (artículo 126 LCSP), sin referencia a marcas ni modelos, de modo que puedan concurrir sistemas de distintos fabricantes que cumplan los requisitos. La exigencia de compatibilidad binaria con la plataforma de cómputo CUDA se justifica por el objeto del contrato (artículo 126.6 LCSP): los modelos, bibliotecas y herramientas de entrenamiento que RTVE utiliza, y la infraestructura GPU que ya explota, se basan en esa plataforma, y otra tecnología obligaría a portar y volver a validar todo el software. La CPU podrá ser de arquitectura ARM o x86-64; en el primer caso los contenedores y herramientas deberán estar disponibles para dicha arquitectura, extremo que los licitadores deberán acreditar.

INFORME DE INSUFICIENCIA DE MEDIOS

En la actualidad la Dirección de Sistemas e Innovación no cuenta con el equipamiento objeto de este proyecto. RTVE dispone de tres servidores adquiridos en 2024, cada uno con dos GPU de 141 GB conectadas por bus PCIe 5.0. Ampliar estos servidores con más GPU no resuelve la necesidad, por un problema de escalabilidad del bus interno: cada GPU se comunica con las demás y con la

memoria del sistema a través de PCIe, a unos 64 GB/s por sentido, frente a unos 900 GB/s de los enlaces de interconexión directa. Al entrenar o ajustar un modelo repartido entre varias GPU, la sincronización constante entre ellas satura el bus, de modo que el rendimiento no crece en proporción al número de GPU añadidas y cae de forma acusada en los modelos de mayor tamaño. Los sistemas solicitados evitan ese límite por diseño: una única GPU de al menos 250 GB permite cargar y ajustar el modelo completo sin comunicación entre GPU, y la interconexión coherente entre CPU y GPU, de en torno a 900 GB/s (unas catorce veces el ancho de banda de PCIe 5.0 x16), da acceso a la memoria total del sistema, de al menos 700 GB, para modelos y conjuntos de datos que no caben en la memoria de la GPU. Por ello se exige en el pliego técnico una memoria total mínima por sistema y una única GPU de alta memoria, en lugar de varias GPU conectadas por PCIe.

LOTES DE LA CONTRATACIÓN

Este expediente se ha planteado en un único lote. El objeto consiste en tres sistemas idénticos que deben ser homogéneos en configuración, software y mantenimiento, para que los procesos de entrenamiento puedan ejecutarse indistintamente en cualquiera de ellos. Su división en lotes sólo multiplicaría proveedores e interlocutores de mantenimiento para un mismo equipamiento, sin ventaja para la concurrencia.

JUSTIFICACIÓN DEL PROCEDIMIENTO

Se propone la tramitación de este expediente por procedimiento ABIERTO, sujeto a regulación armonizada, por ser su valor estimado (459.000,00 €) superior al umbral comunitario aplicable a los contratos de suministro de RTVE.

Cualquier procedimiento abierto es garantía de transparencia e igualdad de trato entre todos los licitadores, y además garantiza mayor concurrencia.

JUSTIFICACION DE LOS CRITERIOS DE SOLVENCIA:

Solvencia económica:

Tener un volumen anual de negocios referido a alguno de los tres últimos ejercicios disponibles igual o superior al 75% del valor estimado del contrato, que equivale a 344.250,00€.

Se considera que un volumen anual de negocios de al menos el 75 % del valor estimado del contrato es suficiente para garantizar que el licitador tiene capacidad económica para ejecutar el suministro y no es un volumen de negocio excesivo, de forma que no se restrinja la concurrencia. Se da la

posibilidad de acreditar el cumplimiento por la cifra de negocios de alguno de los últimos tres ejercicios, lo que aporta una mayor flexibilidad. Teniendo en cuenta el presupuesto estimado y las características del sector, la cifra es proporcional al objeto del contrato.

Solvencia técnica:

Contar al menos con 1 referencia de trabajos similares al objeto del contrato, realizados en los últimos tres años cuyo importe sea igual o superior al 50% del valor estimado del contrato, que equivale a 229.500,00€.

RTVE considera que el mínimo indispensable para poder garantizar la ejecución del contrato en condiciones satisfactorias es de 1 referencia en trabajos similares. Una referencia es una cantidad asumible no restrictiva para los licitadores del sector y demostrará mayor experiencia en el objeto del contrato.

JUSTIFICACIÓN DE LOS CRITERIOS DE VALORACIÓN:

La aplicación de los criterios de valoración definidos en la cláusula 11 ª del PCAP, permitirá que la oferta que resulte adjudicada sea la más adecuada para cubrir las necesidades de RTVE.

Los criterios técnicos de valoración incluidos tienen como objetivo valorar la idoneidad de las ofertas presentadas.

Se asignan 40 puntos a criterios técnicos automáticos, sobre aspectos de interés directo para RTVE que suponen un valor añadido sobre las exigencias técnicas obligatorias:

- Capacidad y potencia de la GPU (20 puntos): memoria de GPU y rendimiento de cálculo en FP8 según la ficha técnica oficial del fabricante. Permite que cada licitador ofrezca el sistema de mayor capacidad disponible y que RTVE obtenga la mayor capacidad por euro, con valores objetivos y verificables documentalmente.
- Reducción del plazo de entrega, instalación y puesta en marcha (15 puntos): los proyectos de transcripción, traducción y generación de imagen y vídeo tienen un calendario ligado a la actividad de RTVE, y la disponibilidad de los sistemas es crítica en un mercado con plazos de entrega largos para esta gama de GPU.
- Soporte en sitio con sustitución de componentes en 4 horas (5 puntos): reduce el tiempo de indisponibilidad de unos equipos de uso intensivo.

Por último, se asignan los 60 puntos restantes a las condiciones económicas, es decir, al precio ofertado. La fórmula aplicada permite calificar adecuadamente las ofertas económicas asignando una puntuación inversamente proporcional al precio respecto de la oferta más ventajosa.

EXISTENCIA DE PRESUPUESTO

El presupuesto base de licitación es de 459.000,00 € (IVA no incluido), 555.390,00 € con un 21 % de IVA (96.390,00 €), correspondiente a tres sistemas con un precio unitario máximo de 153.000,00 €, que incluye el suministro, la instalación, la puesta en marcha y el mantenimiento durante tres años. El valor estimado coincide con el presupuesto base de licitación, al no preverse prórrogas ni modificaciones.

Esta inversión está aprobada en el comité de inversiones, correspondiendo a la orden de inversión I502140M3340 “Desarrollo de actividades en Inteligencia Artificial (I.A.)”.

El presupuesto se ha calculado a partir de la oferta recibida de un integrador para un sistema equivalente (151.000,00 € por unidad, con tres años de mantenimiento en horario laborable y sin plazo de entrega comprometido) y de los precios de catálogo publicados para sistemas de esta gama (en torno a 85.000–100.000 € sólo el equipo, sin plazo de entrega garantizado ni instalación). El precio máximo incluye además tres años de mantenimiento con sustitución en sitio, la instalación y la puesta en marcha, y un margen por la escasez actual de GPU de esta gama y el incremento del precio de la memoria. La adjudicación por procedimiento abierto, con 60 puntos asignados al precio, permitirá que el precio final se ajuste al mercado. No hay comparativa con un expediente anterior, puesto que las características técnicas de los sistemas son distintas.

El gasto se imputará a una única anualidad, la del ejercicio de la conformidad de los sistemas al facturarse en una sola vez tras la conformidad.