ARTICLE Open Access

Quantum-enhanced reconfigurable in-memory stochastic computing

Hong-Zhe Yang[1,2], Jian-Peng Dou[1,2], Feng Lu[1,2], Xiao-Wen Shang[1,2], Chao-Ni Zhang[1,2], Heng Zhou[1,2], Hao Tang[1,2] and Xian-Min Jin[1,2,3,4]✉

Abstract

In-memory computing, which enables computation directly within memory, represents an efficient approach to processing massively parallel computation tasks that are intractable for conventional computers. However, implementations of in-memory computing have been primarily limited to the classical regime, with its nonclassical counterpart yet to be fully explored. Quantum memory, with its unique capability to generate, preserve, and nontrivially operate on quantum states, offers spectacular quantum-enhanced advantages and is thus a promising candidate for in-memory computing. Here, leveraging a room-temperature quantum memory, we demonstrate a quantum-enhanced and reconfigurable in-memory stochastic computing system, where correlated photons, randomly produced in the quantum memory, serve as the computing resources. We show that addition and multiplication operations can be straightforwardly achieved by accumulating photon counts, and multiple computing tasks can be accelerated by mapping them into parallel accumulations of photon counts. Furthermore, the calculation results are obtained through stochastic processes, ensuring security in remote computation since no efficient information can be distinguished by eavesdropping on a small portion of the computation data. This in-memory computing system is enhanced by nonclassical correlations, which accelerate computing process and may stimulate future research and applications in the emerging field of quantum-enhanced computing architectures.

Introduction

With the rise of artificial intelligence alongside large amounts of data, the demand for efficiently processing massively parallel computation tasks becomes progressively prominent. However, it is inaccessible or challenging for conventional computers to meet this demand, due to intolerable energy cost and computing latency incurred by processor-memory dichotomy[1,2]. This motivates researches on new computing architectures. Non-von Neumann architecture, accompanied by advances of integrated photonic chips[3–7], is proposed to efficiently process conventionally intractable tasks in a way of physical-layer parallel computation[8–14]. As one promising non-von Neumann architecture, in-memory computing breaks the traditional concept of separate processing and memory units[15–17], and it even no longer requires deterministic storage, i.e., memory devices can be implemented in a way of stochastic computing[18–21]. Up to now, based on various memory mechanisms and materials, the implementations of in-memory computing have been realized in different systems ranging from electronic[22–27] to photonic platforms[5,28–32], and in-memory computing has been commonly recognized as a promising solution for efficiently processing massively parallel data[33–38], such as applications in accelerating matrix-vector multiplication. Section SA of the Supplementary Information provides a comprehensive introduction to the foundational concepts and key challenges in stochastic computing, in-memory computing, and quantum computing.

Despite significant advances in in-memory computing, these achievements typically rely on the macroscopic properties of memory units in a classical manner, such as the resistance of memristive memory or the transmittance of non-volatile phase-change memory. To date, a memory

Correspondence: Xian-Min Jin (xianmin.jin@sjtu.edu.cn)
[1]Center for Integrated Quantum Information Technologies (IQIT), School of Physics and Astronomy and State Key Laboratory of Photonics and Communications, Shanghai Jiao Tong University, Shanghai, China
[2]Hefei National Laboratory, Hefei, China
Full list of author information is available at the end of the article
These authors contributed equally: Hong-Zhe Yang, Jian-Peng Dou

system that operates intrinsically in the nonclassical regime has not been explored for demonstrating in-memory computing. Quantum memory[39,40], capable of preserving quantum states, plays a crucial role for scalable quantum technologies which promise to outperform their classical counterparts[41–46]. Over the last 20 years, a tremendous amount of work on quantum memory has been done based on various protocols, such as electromagnetically induced transparency[47–51], Raman memory[52–55], photon-echo memory[56–59], Duan-Lukin-Cirac-Zoller (DLCZ) protocol[60–63], off-resonant ladder (or cascaded) memory[64,65], optical loop memory[66,67], and so on. Besides, the wide applications of quantum memory include long-distance quantum communication[68–75], multiphoton synchronization[65,66,76], hybrid quantum networks[67] and single-photon sources[77–79]. Considerable efforts have been dedicated to make quantum memory practical for these suggested applications mentioned above. However, efforts dedicated to a full exploitation of the significance of quantum memories, such as those for constructing unconventional computing, are not enough, although applications of quantum memory are promising.

Here, leveraging a room-temperature quantum memory, we demonstrate an in-memory stochastic computing system that is intrinsically reconfigurable for various computing tasks, due to its controllable and programmable memory processes. We show that basic operations such as addition, scalar multiplication, and vector multiplication can be straightforwardly performed by accumulating photon counts. Furthermore, the computation of multiple tasks can be accelerated by mapping them to parallel photon count accumulations. Our experiment also offers a secure method for remote computing, thanks to the probabilistic generation of photons. This ensures that eavesdropping on a small portion of the computation data provides no meaningful information. Additionally, by utilizing correlated photons directly generated within the quantum memory, we demonstrate that the computing speed can be further enhanced, despite the quantum memory's retrieval efficiency being only around 0.3% (the intrinsic value of 4% multiplied by the total detection efficiency of 7%), which is much lower than the conventional standard for practical quantum memories. These results underscore the quantum-enhanced advantages of in-memory quantum computing and highlight the potential of quantum memories that can be realized with existing technology.

Results

The schematic diagram and physical interpretation of our quantum in-memory computing are shown in Fig. 1 and Fig. 2, respectively. Our in-memory computing is based on a room-temperature quantum memory involving billions of motional atoms, as is shown in Fig. 1a. We illustrate the computing process in Fig. 1b. The computing tasks, such as a computational formula, accepted by the interface unit, are encoded into a corresponding calculation configuration and further into the configuration of addressing pulses. Then, the addressing pulses enter the quantum memory (in-memory unit) and excite or retrieve correlated photons. In this way, the configuration of addressing pulses, as well as calculation tasks, are transferred to the generation probability of Stokes photon, anti-Stokes photon, and atomic spin wave (see Fig. 2). After leaving the quantum memory, the photons generated in a probabilistic manner are detected by single-photon detectors, and the photon counts are accumulated by an accumulator, according to a task-specific accumulation logic. Finally, the interface unit decodes the accumulated data with the information of the calculation configuration, and outputs the calculation results.

In the above calculation procedure, light-matter interaction in the in-memory unit is the key point. Fig. 2a is an illustration of the atomic states and single photons involved in the in-memory unit. To obtain these nonclassical states, three basic physical operations are adopted, as shown in Fig. 2b. A pump light resonant with the transition $|s\rangle \leftrightarrow |e\rangle$ is used to initialize the atoms. After the pump, almost all of the atoms populate the ground state $|g\rangle$. The other two basic physical operations are 'write in' and 'read out', with which the configuration of addressing pulses is encoded into the generation probability of Stokes photon and anti-Stokes photon. The configuration of these basic operations is task-specific. For example, addition operation can be realized by accumulating photon counts of Stokes photon, thus only the pump operation and write operation are involved. However, multiplication based on the product of two probabilities, is realized by detecting the coincidence of non-classically correlated photons, such as correlated Stokes photons and anti-Stokes photons, as is shown in Fig. 2c.

The quantum memory acts as a time-bin analog-to-digital converter and transfers analog information of the addressing pulses into the generation probability of digital-like information, i.e., single photons. In Fig. 3, we experimentally characterize the property of this analog-to-digital converter under different configurations, such as different settings of pulse width and pulse energy. Figure 3a demonstrates a linear dependence of the excitation probability of Stokes photons on write pulse width, and Fig. 3b shows a linear dependence on write pulse energy. As is shown in Fig. 3c, calculation tasks can be encoded by a bin of write pulses, such as an operation of summing can be encoded on the total generation probability of Stokes photons excited by two or more write pulses. The dependence of retrieval efficiency on the pulse width and pulse energy are shown in Fig. 3d, e,

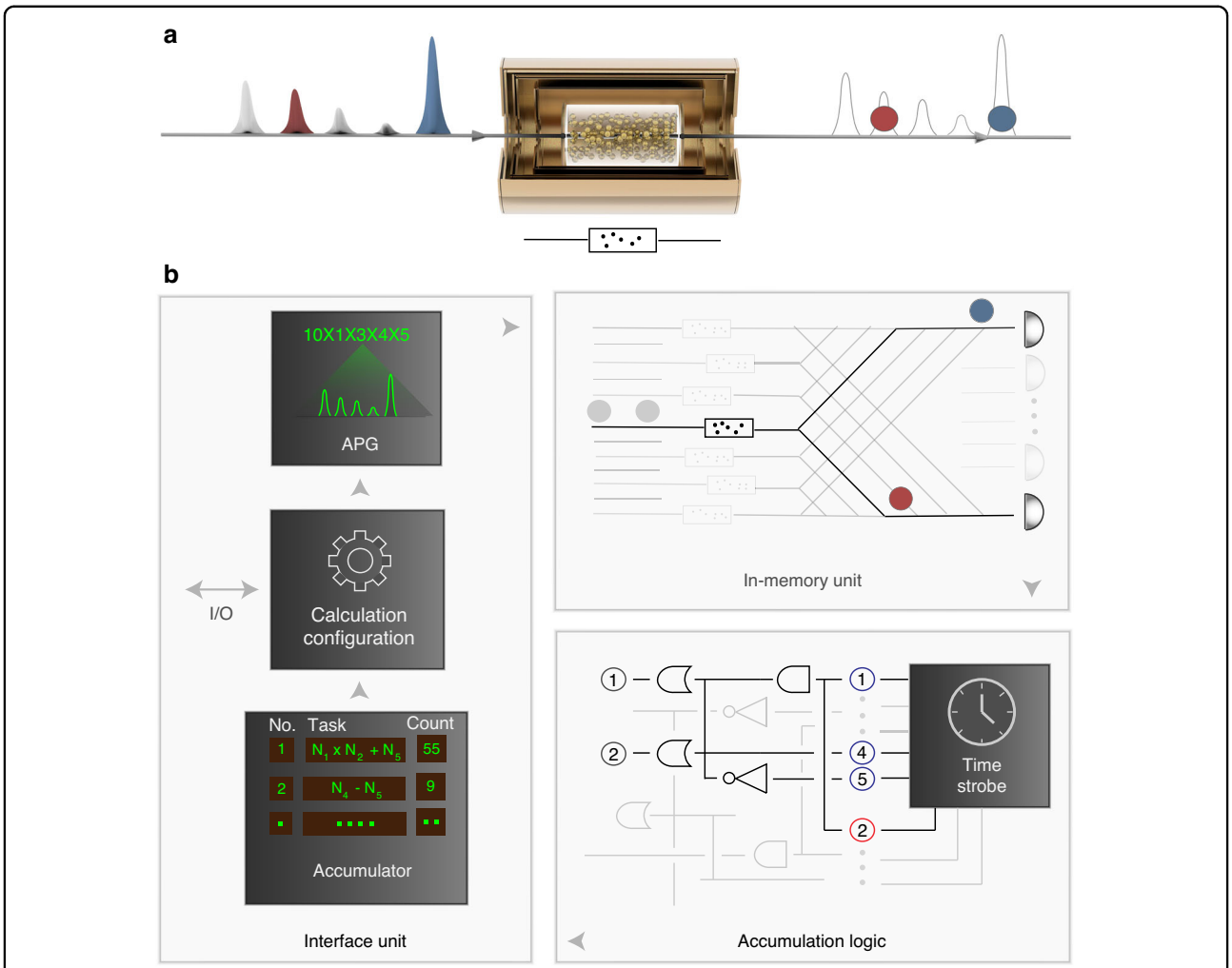

Fig. 1 Schematic of in-memory computing based on quantum memory. a Quantum memory, as the core of quantum in-memory computing, consists of a caesium atom vapour and a three-layer magnetic shielding. The addressing pulses interact with the atomic ensemble packaged in a glass cell, while generating quantum states, including collective atomic excitations and correlated photons. **b** Quantum enhanced in-memory computing process. Interface unit receives the calculation task and encodes it into the configuration of addressing pulses. In-memory unit converts the analog information of addressing pulses to the generation probability of correlated photons. The output from the single-photon detectors is processed and accumulated according to the accumulation logic designed or programmed by the experimenter. Then the accumulated data (i.e., the detected photon counts) is decoded by the interface unit, and the calculation results can be obtained

respectively. Figure 3f demonstrates a nonlinear influence of time delay on cross-correlation of Stokes photon and anti-Stokes photon. It is worth mentioning that linear responses of the quantum memory to the addressing pulses can be directly used in a straightforward and reconfigurable computation, while the nonlinear response can be used in some special scenarios, such as the simulation of a nonlinear process[80].

The Stokes photons and anti-Stokes photons, generated by the addressing pulses in a probabilistic manner, are detected by single-photon detectors. Then the photon counts are recorded by an accumulator. The accumulator is stopped when the accumulated photon counts reach a target count. For obtaining a same target count, different calculation results correspond to different numbers of trials. Therefore, we can obtain the calculation results by recording how many trials are used to achieve the target count. For example, it takes 10,000 trials for accumulating 10 photons, which means a result with a value of 1. Then, the result is 2 when it takes 5000 trials for accumulating 10 photons. However, in practice, there are fluctuations of the number of trials needed to obtain 10 photons, and there is a probability distribution of the number of trials. The width of the distribution is non-zero, which may lead to the overlap between different calculation results. In the overlap, one cannot distinguish different results. For example, the result should be 2, but the trial number may be 10,000 rather than 5000 for obtaining 10 photons

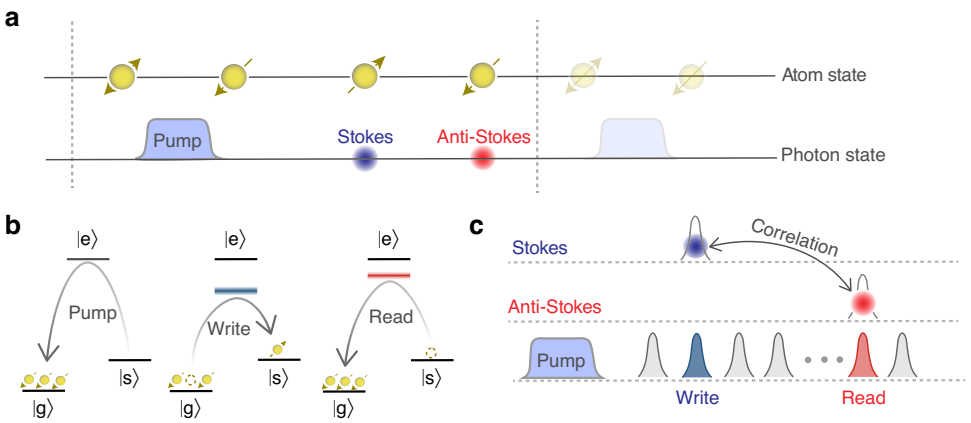

Fig. 2 Physical picture of quantum in-memory computing. a Time sequence of photons and atomic excitations. Yellow circles in upper line depict atomic state. The blue (red) circle in another line represents the Stokes (anti-Stokes) photon. **b** Physical operations that underpin the in-memory computing. There are three basic operations include pump, write-in and read-out operations. **c** Illustration for a multiplication. A multiplication of two numbers is represented by the joint probability of correlated Stokes photon and anti-Stokes photon. For example, the first (second) number is represented by the detection probability of Stokes (anti-Stokes) photon. A pulse train containing more than two pulses promises to process multiple computation tasks involving additions, scalar multiplications and even vector multiplications in a way of parallel accumulation, see sections SB and SC of Supplementary Information for details

under a slightly lower probability, consequently one mistakes this result for 1 under a non-negligible probability.

Both addition and multiplication share a common feature: the operands are encoded in the photon excitation probabilities (for Stokes photons) or the conditional readout efficiencies (for anti-Stokes photons), and the results are obtained by accumulating photon counts. In this sense, the two operations are conceptually similar, as they both rely on counting discrete stochastic events. Specifically, addition corresponds to the classical accumulation of independently generated Stokes photons. Multiplication is realized via the correlation between Stokes and anti-Stokes photons, where the first operand is encoded in the Stokes excitation probability, and the second operand in the conditional readout probability. Mathematically, if n_i^{S} and n_i^{AS} respectively denote the i-th Stokes and anti-Stokes events (0 or 1), the operations can be expressed as:

$$\textbf{Addition}: N_{\mathrm{add}} = \sum_{i=1}^{M} n_i^{\mathrm{S}},\ \langle N_{\mathrm{add}} \rangle = MP_{\mathrm{S}}$$

$$\textbf{Multiplication}: N_{\mathrm{mult}} = \sum_{i=1}^{M} n_i^{\mathrm{S}} n_i^{\mathrm{AS}},\ P_{\mathrm{coinc}} = \frac{\langle N_{\mathrm{mult}} \rangle}{M} = P_{\mathrm{S}} P_{\mathrm{R}}$$

where P_{S} is the Stokes photon excitation probability and P_{R} the retrieval efficiency. M denotes the average trial number required to obtain the photon counts N_{add} or N_{mult}.

For effectively decoding the accumulated data, i.e., for accurately distinguishing different results, one should carry out enough trials for accumulating enough photon counts, by which the fluctuation and the overlap can be suppressed or eliminated. Figure 4 shows the probability distributions of the trial numbers needed for accumulating a target count of Stokes photons. The peaks in different colors correspond to results with different central values of trial numbers due to the different pulse energies of addressing pulses. From Fig. 4a–d, the overlap between different results decreases with the increasing of target photon count and trial number. Regarding computational accuracy and fidelity, the final result is obtained by summing photon counts, with the target photon number determining the distinguishability of different outcomes. The required target photon number for reliable resolution is analysed in the section SD of the Supplementary Information.

The large overlap in Fig. 4a is an advantage in some sense, since it may secure the computation in a new way. Due to the large overlap, when the photon count is small, no one can eavesdrop an accurate result by eavesdropping a small portion of the photon counts. In a computing network consisting of remote in-memory units and remote accumulators, a secure remote computation may be guaranteed by virtue of this overlap. For example, if large amounts of data are intercepted, eavesdropping will be easily detected. Alternatively, communication is interrupted when the data rate falls below a certain threshold to ensure security. In Fig. 4, the pulse configurations with higher energies appear to correspond to narrower probability distributions. The physical origin of this effect is that a stronger pulse leads to a higher

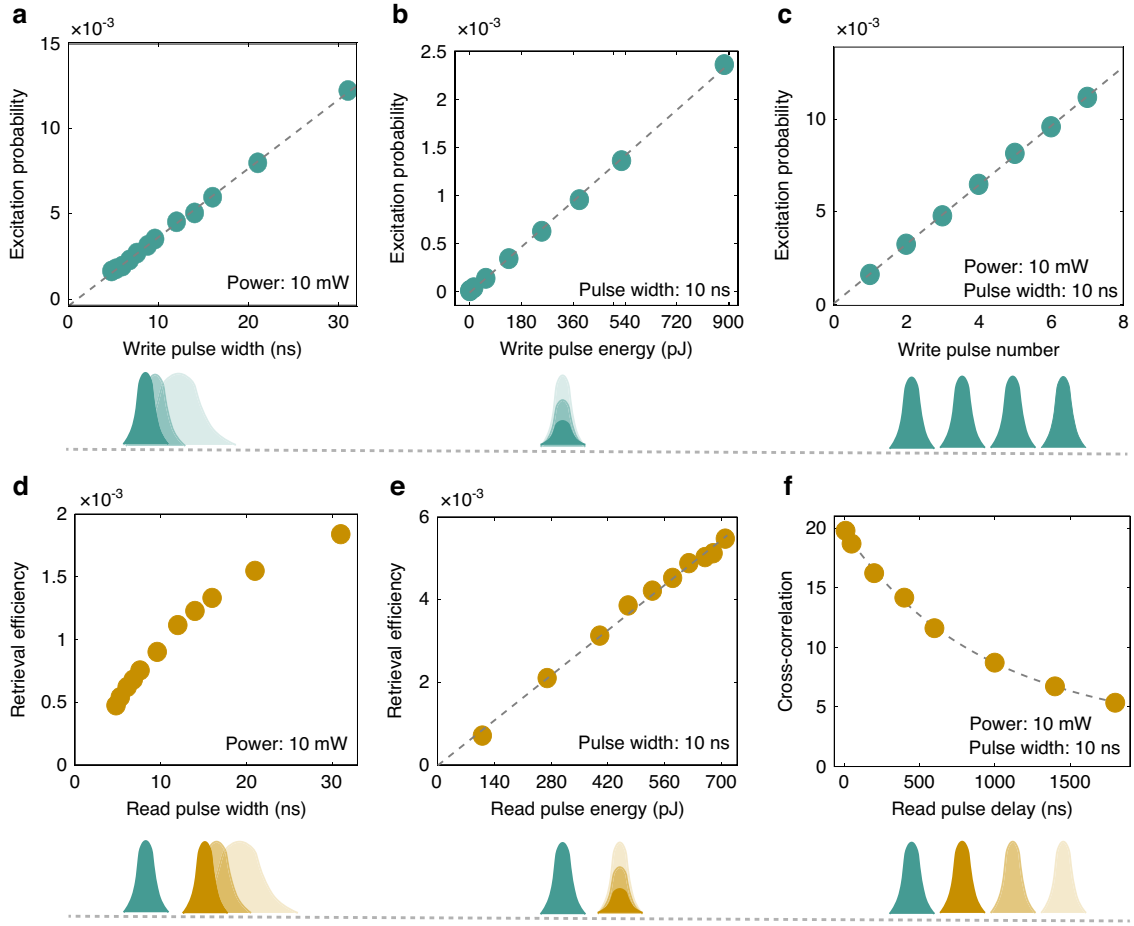

Fig. 3 Excitation probability, retrieval efficiency and cross-correlation under different configurations of the addressing pulse. Dark cyan circles represent excitation probability of Stokes photons with different pulse width (**a**), pulse energy (**b**), and number (**c**) of write pulses. Brown circles depict retrieval efficiency of anti-Stokes photons and cross-correlation with different read pulse width (**d**), read pulse energy (**e**), and time delay between write pulse and read pulse (**f**). The error bars are too small to be visible. Dashed grey lines are fitting curves of the experimental data. Dark cyan pulses and brown pulses represent write pulses and read pulses, respectively. The experimental parameters involving light power or pulse width are presented in the bottom right of each sub-figure. The above values of probability and efficiency represent raw data, rather than the intrinsic characteristics of the quantum memory. For instance, the reported retrieval efficiency of 0.3% in the sub-figure (**e**) corresponds to an intrinsic value of 4%, where the total detection efficiency of 7% is taken into account

excitation probability and a larger variance for generating Stokes photons. Note that the average number of trials for obtaining one target photon count is, in fact, the inverse of the detected excitation probability.

In quantum communication, security is designed to protect the transmission channel, ensuring that quantum information cannot be intercepted or cloned during propagation (as in quantum key distribution). In contrast, our approach focuses on the protection of the computational process itself. In our system, the computation proceeds through photon-count accumulation within distributed in-memory units and remote accumulators. During this process, the computational state is continuously evolving and has not yet been converted into any readable classical information. An external observer attempting to eavesdrop would only access a highly uncertain quantum state–a mixture of stochastic photon events with overlapping probability distributions–from which no meaningful intermediate results can be inferred. Although the underlying process involves photon exchange and may superficially resemble secure communication, its essential function is to guarantee the confidentiality of quantum computation, akin in spirit to blind quantum computing[81]. Thus, the proposed mechanism should be regarded as a secure quantum computing framework, rather than a communication protocol.

As our in-memory computing is based on the accumulation of photon counts, its computing speed is determined by the generation rate of photons. A high generation rate of photons means a high computing speed. Especially, the rate of coincidence counts of photons from different counting

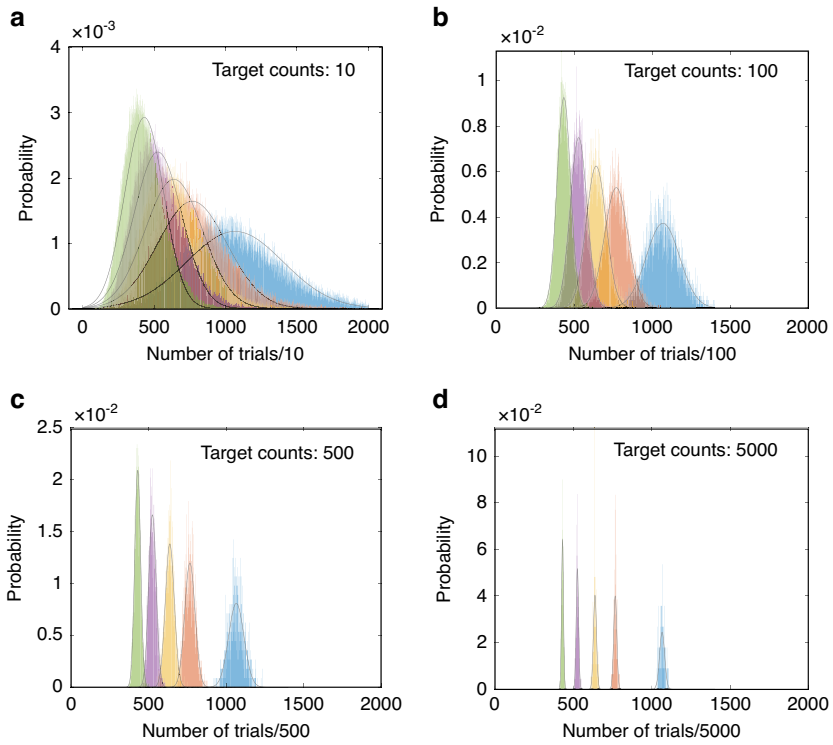

Fig. 4 Number of trials required for effectively decoding the accumulated data. The peaks represent the probability distributions of obtaining a target count of Stokes photons, as a function of the number of trials. For example, in (**a**), the probability of obtaining 10 target Stokes photons with 3900 trials is about 0.003, which is the maximum value of green peak. The peaks in different colors correspond to different pulse energy of write pulse, and the peak in green corresponds to the write pulse with the maximum pulse energy, while the peak in blue corresponds to the weakest write pulse. The envelop (black solid line) of each peak is a fitting curve in a form of Gaussian function. (**a**–**d**) correspond to different fixed target photon counts and trial number needed, while the horizontal scale, i.e., the ratio of fixed target photon counts and the corresponding trial number needed, is maintained. The overlap between different peaks decreases with the increasing of target photon counts and trial number, see section SD of Supplementary Information for details

channels (Stokes photon and anti-Stokes photon) is a key parameter for multiplication operations. In our experiment, the coincidence probability between two photons from different trials is lower than the coincidence probability between correlated photons generated in a same trial. For example, when the energy of write and read pulses is 400 pJ, the available retrieval efficiency is around 0.3%, much lower than unit efficiency, but is still larger than the probability (0.1%) of spontaneous Raman scattering in the write process, see Fig. 3b, e. This is the reason why computing speed dependent on the rate of coincidence events can be accelerated by using correlated photons. The acceleration factor (*AF*) is written as the ratio of quantum-enhanced coincidence probability to classically random coincidence probability,

$$AF = \frac{P_{Si-ASj}}{P_{Si}P_{Sj}} \tag{1}$$

where i indexes a write pulse and j a read pulse. Note that $i < j$ to ensure that the write pulse is prior to the read pulse. The acceleration factor with different i and j are shown in Fig. 5. The acceleration factor is about 3 when the Stokes photons and anti-Stokes photons are generated by a pair of adjacent write pulse and read pulse, i.e., $i = j - 1$. The acceleration factor reaches its maximum for single-memory operation because the read pulse is applied immediately after the write pulse (i.e., at the minimal delay), leading to the highest read-out efficiency and the strongest temporal correlation between the Stokes and anti-Stokes photons. When multiple read pulses follow a single write pulse, or when the read-out delay increases, the memory coherence and retrieval efficiency decrease. Consequently, $P_{Si}P_{ASj}$ approaches $P_{Si}P_{Sj}$, and the acceleration effect is diminished.

The inset shows the effects of pulse energy on acceleration factor (green circles) and cross-correlation $g^{(2)} = \frac{P_{Si-ASj}}{P_{Si}P_{ASj}}$ (red squares) when $i = 1$ and $j = 2$. The acceleration factor, on the other hand, characterises how this nonclassical correlation accelerates the computation under the same excitation probability (i.e., the same pulse energy). A high cross-correlation

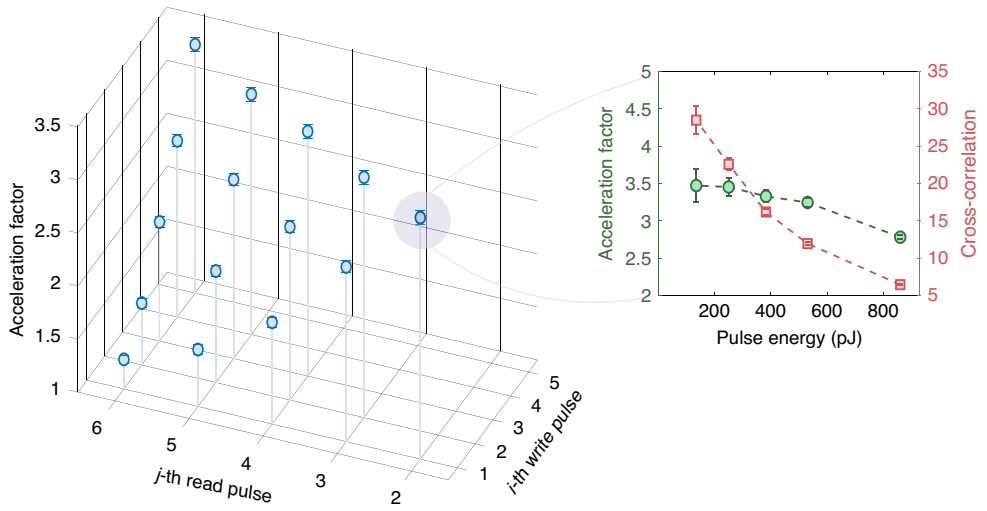

Fig. 5 Acceleration factor due to nonclassical correlation. The acceleration factor equals $P_{Si-ASj}/(P_{Si}P_{Sj})$. P_{Si} denotes the probability of Stokes photons generated by the *i*-th write pulse. $(P_{Si}P_{Sj})$ denotes the random coincidence probability of Stokes photons generated by the *i*-th and *j*-th write pulses. P_{Si-ASj} is the joint probability of Stokes photon and anti-Stokes photon, which are respectively excited and retrieved by the *i*-th write pulse and *j*-th read pulse in the same trial. The inset shows the effects of pulse energy on acceleration factor (green circles) and cross-correlation (red squares) when $i = 1$ and $j = 2$

corresponds to a high acceleration factor, which implies the advantage of nonclassical correlation for in-memory computing. The acceleration factor is maximized when the quantum correlation between the write-in and read-out processes are strongest–namely, in the single-memory, minimal-delay configuration. Therefore, the computing speed scales with both the excitation probability and the strength of the nonclassical correlation: higher excitation increases the photon generation rate, while stronger correlation enhances the effective accumulation rate, together determining the overall acceleration of computation.

Discussion

The logical operation takes place within the quantum memory through light-matter interaction, where correlated quantum states are generated and stored. The photon-number accumulation during detection merely maps the pre-existing quantum correlations into classical outcomes. Therefore, the operation on the quantum state (via coherent evolution inside the atomic ensemble) is physically distinct from the photon-number accumulation (a classical readout process). The multiplication operation inherently exhibits the key characteristics of in-memory computing. Specifically, during the write process, the joint excitation of a Stokes photon and its correlated atomic excitation encodes the first operand (A) in the excitation probability of the Stokes field. During the subsequent read process, a properly tuned read pulse converts the stored atomic excitation into an anti-Stokes photon, with the retrieval probability encoding the second operand (B). The coincidence probability between the Stokes and anti-Stokes photons thus represents the product $A \times B$, realized entirely through light-matter interaction within the same atomic ensemble.

It is worth emphasizing that the quantum in-memory computing system and classical in-memory computing address different computational paradigms–the former focuses on quantum randomness and correlation-enhanced computing, while the latter targets deterministic parallel computation[28]. Keeping this distinction in mind, it is insightful to make some comparisons between the two approaches. In a classical photonic in-memory computing platform based on phase-change materials (PCMs), the crystallization and amorphization processes (i.e., initialization) require optical powers on the order of 5–14 mW, while the energy per write pulse is typically a few hundred picojoules. In the quantum-memory-based in-memory computing system, the optical power required for initializing the atomic ensemble is of the same order (around 10 mW), and the energy of each write/read pulse is also a few hundred picojoules. Therefore, the energy consumed per operation in our system is comparable to that of the classical photonic in-memory computing approach. In addition, there are two key distinctions. First, in the quantum version, the signal photons (Stokes and anti-Stokes photons) are generated intrinsically within the quantum memory, without requiring external signal light input, which leads to an inherent energy-saving advantage compared to classical optical systems that rely on continuous input beams (typically tens of microwatts). Second, the computation demonstrated in this work is a stochastic quantum process based on the generation of

truly random single photons, with a typical generation probability of about 1%. Reliable computational outcomes are obtained by accumulating a few hundred photon-counting events. This operating mechanism is fundamentally different from conventional in-memory computing systems, which employ deterministic optical signals.

In summary, we have demonstrated an instance of quantum-enhanced in-memory stochastic computing by mapping computation tasks into two basic physical operations, including the write-in and read-out of a room-temperature quantum memory. Computation is completed through two steps. Firstly, encode the analogue information depicting computation tasks into the generation probability of digital-like photons. Then, after a period of accumulation of the photon counts, the accumulated data is decoded to the desired calculation results according to the target photon counts and used trial number. We analyse the required computing resources for completing addition and multiplication, as well as the corresponding physical processes in quantum memory device. By characterising the encoding and decoding process, we further analyse the performance of in-memory computing enabled by intrinsic randomness and non-classical correlation, which introduces two advantages of quantum in-memory computing: Our results promise a secure way for remote quantum computation[82–84]. The security of computing is guaranteed by quantum randomness, since the calculation results are obtained by accumulating randomly generated photons, and it is hard to infer a precise result based on a small portion of detection events, due to the overlap of probability distribution of needed trial number. Furthermore, in classical regime, in order to increase computing speed, one needs to raise the pulse energy to reach a high generation probability of photons. While in our experiment, non-classical correlation is shown to accelerate computing by higher coincidence probability with a same pulse energy as the classical situation. Sections SE and SF of Supplementary Information provide comprehensive comparisons with the classical in-memory computing systems and bit-stream generators.

In-memory computing is limited in classical regime due to the lack of efforts on exploring the significance of imperfect quantum memories in computing, and it is usually suggested that a quantum memory without a high memory efficiency is impractical for memory-based long-distance and large-scale quantum networks[85,86], while our results show that a quantum memory with a low memory efficiency can be utilized for constructing quantum-enhanced in-memory stochastic computing. As long as the coincidence probability of correlated photons is higher than that of uncorrelated photons, the computing involving additions and multiplications can be accelerated. Now is perhaps the time to consider the question: Is one imperfect quantum memory practical? Maybe, in addition to further efforts of exploring quantum memories with better performance, some efforts should be dedicated to have a more comprehensive knowledge and a full exploitation of the significance of imperfect quantum memories. A quantum memory that's not perfect for the generally expected applications may be just right for some others that are usually unnoticed. A notable example, quantum memory hasn't been fully exploited for constructing unconventional computing in a way of non-deterministic storage.

For real-life applications, the computing speed is dependent on the generation rate of photon counts. Therefore, the improvement of available memory efficiency (including efficiencies in write and read processes) enables the capability of rapidly processing computation tasks. There are many other aspects to further improve the performance of in-memory computing with quantum memory, such as a broad bandwidth memory for data-intensive works and a long-life memory for multiplications of large vectors. Combining with integrated photonic techniques[87–89] and spatial multiplexing techniques[90–92], which promise applications in quantum information processing[93,94], a compact in-memory quantum computing device may be realized. Furthermore, the coherence stored in quantum memory is an engine of quantum computing, which is urgently required in the construction of future in-memory quantum computing. In addition to in-memory computing, the potential applications of an imperfect quantum memory also include light-matter interface for investigating hybrid interference[95,96] and fundamental tests of quantum mechanics[97], as the quantum states stored in quantum memory can be nontrivially operated[98]. In a word, our work takes one step towards the in-memory quantum computing, as well as emphasizes the role of imperfect quantum memories in quantum computing and provides enlightenment to future researches on quantum memories.

Materials and methods

One external cavity diode laser acts as the source of pump light with sensitive temperature and frequency feedback, and provides frequency reference for other lasers. A distributed Bragg reflector laser locked to the reference laser with a frequency difference of 4 GHz is used as the source of write/read pulses. The continuous wave from the distributed Bragg reflector laser is chopped by an Electro Optic Modulator (EOM). Then, a tapered amplifier is utilized to boost the power of write/read pulses. Cesium atoms are packed in a 75 mm-long cylindrical glass cell with 10-Torr Ne buffer gas. The glass cell is placed in a three-layer magnetic shielding, and is

heated up to 61 °C. Before entering cesium cell, addressing pulses are horizontally polarized by a Glan-Taylor polarizer. The polarization of generated Stokes photons and anti-Stokes photons are vertical to that of write/read pulses, due to which we can use a Wollaston prism to basically filter out the write/read pulses. A frequency filter, consisting of six home-made cascaded cavities, separates the Stokes photons from anti-Stokes photons, and filters out the noise photons. The transmission rate of every cavity is higher than 90%, and the extinction ratio of every cavity is up to 500:1. Stokes photons and anti-Stokes photons are detected by different single-photon detectors, and the detected photon counts are recorded by a multi-channel counting system for further processing.

For realizing scalar multiplication, coincidence counts between correlated Stokes photons and anti-Stokes photons are used. For realizing vector multiplication, which is a fundamental operation for matrix manipulation, multiple pairs of write pulse and read pulse should be implemented. For example, $[N_1\ N_2\ N_3][M_1\ M_2\ M_3]^T$ can be realized by three pairs of write pulse and read pulse. The first pair of pulses encodes the product of N_1 and M_1, and the second pair encodes N_2M_2. Based on the ability of completing addition and multiplication in a predetermined accumulation logic, an acceleration is brought into the calculation process because of the intrinsic parallel computing mode. Each write operation probabilistically generates a correlated pair consisting of a Stokes photon and an atomic spin excitation through spontaneous Raman scattering. The intrinsic excitation probability per write pulse is typically around 1% (or even lower), meaning that, on average, only one successful excitation occurs in about one hundred trials. Consequently, different write processes are almost statistically independent, and no coherent interference between successive write pulses is expected. This provides a key strength of our approach: it inherently mitigates the accumulation of computational errors as the vector size increases. Based on our experimental observations, if one write operation successfully creates an atomic excitation and a Stokes photon, the probability of generating another excitation in the subsequent write pulse increases slightly–by approximately 0.5%. This minor effect arises from enhanced Raman scattering due to the pre-existing atomic excitation and can be compensated by a slight adjustment of the write-pulse energy. The stochastic and low-probability nature of the Raman write process ensures that no phase control across pulses is required, and quantum-state interference between different write-in pulses does not occur under our experimental conditions.

Acknowledgements
This research is supported by the National Key R&D Program of China (Grant No. 2024YFA1409300); National Natural Science Foundation of China (NSFC) (Grants No. 62235012, No. 12304342, No.12574549, No.12574542); Quantum Science and Technology-National Science and Technology Major Project (Grants No. 2021ZD0301500, and No. 2021ZD0300700); Science and Technology Commission of Shanghai Municipality (STCSM) (Grants No. 2019SHZDZX01, No. 24ZR1438700, No. 24ZR1430700 and No. 24LZ1401500); Startup Fund for Young Faculty at SJTU (SFYF at SJTU) (Grants No. 24X010502876 and No. 24X010500170); Frontier Technologies R&D Program of Jiangsu (Grant No.SBF20250000094). X.-M.J. acknowledges additional support from a Shanghai talent program and support from Zhiyuan Innovative Research Center of Shanghai Jiao Tong University. H. T. acknowledges additional support from Yangyang Development Fund.

Author details
[1]Center for Integrated Quantum Information Technologies (IQIT), School of Physics and Astronomy and State Key Laboratory of Photonics and Communications, Shanghai Jiao Tong University, Shanghai, China. [2]Hefei National Laboratory, Hefei, China. [3]TuringQ Co. Ltd., Shanghai, China. [4]Chip Hub for Integrated Photonics Xplore (CHIPX), Shanghai Jiao Tong University, Wuxi, China

Author contributions
X.-M.J. supervised the project. H.-Z.Y., J.-P.D., and X.-M.J. designed the experiment. H.-Z.Y., J.-P.D., F.L., X.-W.S., C.-N.Z., H.Z., and H.T. performed the experiment. J.-P.D., H.-Z.Y., and X.-M.J. analysed the data and wrote the paper. All the authors discussed the results and contributed to the writing of the paper.

Data availability
All study data are included in the article and/or supporting information.

Conflict of interest
The authors declare no competing interests.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41377-025-02181-6.

Received: 23 May 2025 Revised: 25 December 2025 Accepted: 25 December 2025
Published online: 18 March 2026

References
1. Horowitz, M. 1.1 Computing's energy problem (and what we can do about it). 2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC). San Francisco: IEEE, 10–14s, (2014).
2. Li, C. et al. The challenges of modern computing and new opportunities for optics. *PhotoniX* **2**, 20 (2021).
3. Kimovski, D. et al. Beyond Von Neumann in the computing continuum: Architectures, applications, and future directions. *IEEE Internet Comput.* **28**, 6–16 (2024).
4. Xu, X.-Y. & Jin, X.-M. Integrated photonic computing beyond the von Neumann architecture. *ACS Photonics* **10**, 1027–1036 (2023).
5. Feldmann, J. et al. Parallel convolutional processing using an integrated photonic tensor core. *Nature* **589**, 52–58 (2021).
6. Xu, X.-Y. et al. A scalable photonic computer solving the subset sum problem. *Sci. Adv.* **6**, eaay5853 (2020).
7. Zhang, X. M. & Yung, M. H. Low-depth optical neural networks. *Chip* **1**, 100002 (2022).
8. Pierangeli, D. et al. Scalable spin-glass optical simulator. *Phys. Rev. Appl.* **15**, 034087 (2021).
9. Borders, W. A. et al. Integer factorization using stochastic magnetic tunnel junctions. *Nature* **573**, 390–393 (2019).
10. Jouppi, N. P. et al. In-datacenter performance analysis of a tensor processing unit. 2017 ACM/IEEE 44th Annual International Symposium on Computer Architecture (ISCA). Toronto: IEEE, 1–12, (2017).
11. Monmasson, E. & Cirstea, M. N. FPGA design methodology for industrial control systems–a review. *IEEE Trans. Ind. Electron.* **54**, 1824–1842 (2007).
12. Keckler, S. W. et al. GPUs and the future of parallel computing. *IEEE Micro* **31**, 7–17 (2011).
13. Burstein, I. Nvidia data center processing unit (DPU) architecture. 2021 IEEE Hot Chips 33 Symposium (HCS). Palo Alto: IEEE: 1–20, (2021).

14. Xu, X.-Y. et al. Reconfigurable integrated photonic processor for NP-complete problems. *Adv. Photonics* **6**, 056011 (2024).
15. Liu, Y. T. et al. Cryogenic in-memory computing using magnetic topological insulators. *Nat. Mater.* **24**, 559–564 (2025).
16. Sun, Z. et al. A full spectrum of computing-in-memory technologies. *Nat. Electron.* **6**, 823–835 (2023).
17. Sebastian, A. et al. Memory devices and applications for in-memory computing. *Nat. Nanotechnol.* **15**, 529–544 (2020).
18. Alam, M. R. et al. Stochastic computing in beyond Von-Neumann era: Processing bit-streams in memristive memory. *IEEE Transac. Circ. Syst. II: Exp. Briefs* **69**, 2423–2427 (2022).
19. Banchi, L. Accuracy vs memory advantage in the quantum simulation of stochastic processes. *Mach. Learn.: Sci. Technol.* **5**, 025036 (2024).
20. Shim, Y. et al. Stochastic spin-orbit torque devices as elements for Bayesian inference. *Sci. Rep.* **7**, 14101 (2017).
21. Alaghi, A. & Hayes, J. P. Survey of stochastic computing. *ACM Trans. Embedded Comput. Syst. (TECS)* **12**, 92 (2013).
22. Le Gallo, M. et al. A 64-core mixed-signal in-memory compute chip based on phase-change memory for deep neural network inference. *Nat. Electron.* **6**, 680–693 (2023).
23. Traversa, F. L. et al. Memcomputing NP-complete problems in polynomial time using polynomial resources and collective states. *Sci. Adv.* **1**, e1500031 (2015).
24. Wei, S. T. et al. Trends and challenges in the circuit and macro of RRAM-based computing-in-memory systems. *Chip* **1**, 100004 (2022).
25. Chua, L. Resistance switching memories are memristors. *Appl. Phys. A* **102**, 765–783 (2011).
26. Wong, H. S. & Salahuddin, S. Memory leads the way to better computing. *Nat. Nanotechnol.* **10**, 191–194 (2015).
27. Wang, S. Y. et al. Two-dimensional ferroelectric channel transistors integrating ultra-fast memory and neural computing. *Nat. Commun.* **12**, 53 (2021).
28. Ríos, C. et al. In-memory computing on a photonic platform. *Sci. Adv.* **5**, eaau5759 (2019).
29. Dong, B. W. et al. Partial coherence enhances parallelized photonic computing. *Nature* **632**, 55–62 (2024).
30. Pintus, P. et al. Integrated non-reciprocal magneto-optics with ultra-high endurance for photonic in-memory computing. *Nat. Photonics* **19**, 54–62 (2025).
31. Sanz, M., Lamata, L. & Solano, E. Invited Article: Quantum memristors in quantum photonics. *APL Photonics* **3**, 080801 (2018).
32. Wu, C. M. et al. Programmable phase-change metasurfaces on waveguides for multimode photonic convolutional neural network. *Nat. Commun.* **12**, 96 (2021).
33. Lin, Y. D. et al. Deep Bayesian active learning using in-memory computing hardware. *Nat. Computational Sci.* **5**, 27–36 (2025).
34. Hoffmann, A. et al. Quantum materials for energy-efficient neuromorphic computing: Opportunities and challenges. *APL Mater.* **10**, 070904 (2022).
35. Kim, K. M. et al. Single-cell stateful logic using a dual-bit memristor. *Phys. Status Solidi - Rapid Res. Lett.* **13**, 1800629 (2019).
36. Tuma, T. et al. Detecting correlations using phase-change neurons and synapses. *IEEE Electron Device Lett.* **37**, 1238–1241 (2016).
37. Le Gallo, M. et al. Compressed sensing recovery using computational memory. 2017 IEEE International Electron Devices Meeting (IEDM), San Francisco: IEEE, 28.3.1–28.3.4, (2017).
38. Eleftheriou, E. et al. Deep learning acceleration based on in-memory computing. *IBM J. Res. Dev.* **63**, 7:1–7:16 (2019).
39. Lvovsky, A. I., Sanders, B. C. & Tittel, W. Optical quantum memory. *Nat. Photonics* **3**, 706–714 (2009).
40. Afzelius, M., Gisin, N. & de Riedmatten, H. Quantum memory for photons. *Phys. Today* **68**, 42–47 (2015).
41. O'Brien, J. L., Furusawa, A. & Vučković, J. Photonic quantum technologies. *Nat. Photonics* **3**, 687–695 (2009).
42. Gao, J. et al. Quantum advantage with Membosonsampling. *Chip* **1**, 100007 (2022).
43. Bharti, K. et al. Noisy intermediate-scale quantum algorithms. *Rev. Mod. Phys.* **94**, 015004 (2022).
44. Ladd, T. D. et al. Quantum computers. *Nature* **464**, 45–53 (2010).
45. Gisin, N. & Thew, R. Quantum communication. *Nat. Photonics* **1**, 165–171 (2007).
46. Li, Z.-M. et al. Fast correlated-photon imaging enhanced by deep learning. *Optica* **8**, 323–328 (2021).
47. Chanelière, T. et al. Storage and retrieval of single photons transmitted between remote quantum memories. *Nature* **438**, 833–836 (2005).
48. Eisaman, M. D. et al. Electromagnetically induced transparency with tunable single-photon pulses. *Nature* **438**, 837–841 (2005).
49. Zhang, H. et al. Preparation and storage of frequency-uncorrelated entangled photons from cavity-enhanced spontaneous parametric downconversion. *Nat. Photonics* **5**, 628–632 (2011).
50. Xu, Z. X. et al. Long lifetime and high-fidelity quantum memory of photonic polarization qubit by lifting Zeeman degeneracy. *Phys. Rev. Lett.* **111**, 240503 (2013).
51. Hsiao, Y. F. et al. Highly efficient coherent optical memory based on electromagnetically induced transparency. *Phys. Rev. Lett.* **120**, 183602 (2018).
52. Reim, K. F. et al. Multi-pulse addressing of a Raman quantum memory: Configurable beam splitting and efficient readout. *Phys. Rev. Lett.* **108**, 263602 (2012).
53. Ding, D. S. et al. Raman quantum memory of photonic polarized entanglement. *Nat. Photonics* **9**, 332–338 (2015).
54. Hosseini, M. et al. Unconditional room-temperature quantum memory. *Nat. Phys.* **7**, 794–798 (2011).
55. Guo, J. X. et al. High-performance Raman quantum memory with optimal control in room temperature atoms. *Nat. Commun.* **10**, 148 (2019).
56. Sparkes, B. M. et al. AC Stark gradient echo memory in cold atoms. *Phys. Rev. A* **82**, 043847 (2010).
57. Cho, Y. W. et al. Highly efficient optical quantum memory with long coherence time in cold atoms. *Optica* **3**, 100–107 (2015).
58. Tang, J. S. et al. Storage of multiple single-photon pulses emitted from a quantum dot in a solid-state quantum memory. *Nat. Commun.* **6**, 8652 (2015).
59. Fröwis, F. et al. Experimental certification of millions of genuinely entangled atoms in a solid. *Nat. Commun.* **8**, 907 (2017).
60. Zhao, B. et al. A millisecond quantum memory for scalable quantum networks. *Nat. Phys.* **5**, 95–99 (2009).
61. Radnaev, A. G. et al. A quantum memory with telecom-wavelength conversion. *Nat. Phys.* **6**, 894–899 (2010).
62. Dou, J.-P. et al. A broadband DLCZ quantum memory in room-temperature atoms. *Commun. Phys.* **1**, 55 (2018).
63. Dideriksen, K. B. et al. Room-temperature single-photon source with near-millisecond built-in memory. *Nat. Commun.* **12**, 3699 (2021).
64. Kaczmarek, K. T. et al. High-speed noise-free optical quantum memory. *Phys. Rev. A* **97**, 042316 (2018).
65. Finkelstein, R. et al. Fast, noise-free memory for photon synchronization at room temperature. *Sci. Adv.* **4**, eaap8598 (2018).
66. Kaneda, F. et al. Quantum-memory-assisted multi-photon generation for efficient quantum information processing. *Optica* **4**, 1034–1037 (2017).
67. Pang, X.-L. et al. A hybrid quantum memory-enabled network at room temperature. *Sci. Adv.* **6**, eaax1425 (2020).
68. Zhang, S. et al. Fast delivery of heralded atom-photon quantum correlation over 12 km fiber through multiplexing enhancement. *Nat. Commun.* **15**, 10306 (2024).
69. Duan, L. M. et al. Long-distance quantum communication with atomic ensembles and linear optics. *Nature* **414**, 413–418 (2001).
70. Sangouard, N. et al. Quantum repeaters based on atomic ensembles and linear optics. *Rev. Mod. Phys.* **83**, 33 (2011).
71. Yu, Y. et al. Entanglement of two quantum memories via fibres over dozens of kilometres. *Nature* **578**, 240–245 (2020).
72. Liu, X. et al. Heralded entanglement distribution between two absorptive quantum memories. *Nature* **594**, 41–45 (2021).
73. Yang, T.-H. et al. Time-bin entanglement built in room-temperature quantum memory. *Phys. Rev. A* **103**, 062403 (2021).
74. Pu, Y. F. et al. Experimental demonstration of memory-enhanced scaling for entanglement connection of quantum repeater segments. *Nat. Photonics* **15**, 374–378 (2021).
75. Li, H. et al. Heralding quantum entanglement between two room-temperature atomic ensembles. *Optica* **8**, 925–929 (2021).
76. Nunn, J. et al. Enhancing Multiphoton Rates with Quantum Memories. *Phys. Rev. Lett.* **110**, 133601 (2013).
77. Davidson, O. et al. Single-photon synchronization with a room-temperature atomic quantum memory. *Phys. Rev. Lett.* **131**, 033601 (2023).
78. Chen, S. et al. Deterministic and Storable Single-Photon Source Based on a Quantum Memory. *Phys. Rev. Lett.* **97**, 173004 (2006).

79. Dou, J.-P. et al. Direct observation of broadband nonclassical states in a room-temperature light-matter interface. *npj Quantum Inf.* **4**, 31 (2018).
80. Berdan, R. et al. Low-power linear computation using nonlinear ferroelectric tunnel junction memristors. *Nat. Electron.* **3**, 259–266 (2020).
81. Wei, Y.-C. et al. Universal distributed blind quantum computing with solid-state qubits. *Science* **388**, 509–513 (2025).
82. Zeuner, J. et al. Experimental quantum homomorphic encryption. *npj Quantum Inf.* **7**, 25 (2021).
83. Fitzsimons, J. F. Private quantum computation: an introduction to blind quantum computing and related protocols. *npj Quantum Inf.* **3**, 23 (2017).
84. Huang, H. L. et al. Experimental blind quantum computing for a classical client. *Phys. Rev. Lett.* **119**, 050503 (2017).
85. Razavi, M., Piani, M. & Lütkenhaus, N. Quantum repeaters with imperfect memories: Cost and scalability. *Phys. Rev. A* **80**, 032301 (2009).
86. Wu, Y. F., Liu, J. L. & Simon, C. Near-term performance of quantum repeaters with imperfect ensemble-based quantum memories. *Phys. Rev. A* **101**, 042301 (2020).
87. Xia, F. N., Sekaric, L. & Vlasov, Y. Ultracompact optical buffers on a silicon chip. *Nat. Photonics* **1**, 65–71 (2007).
88. Hummon, M. T. et al. Photonic chip for laser stabilization to an atomic vapor with 10^{-11} instability. *Optica* **5**, 443–449 (2018).
89. Liu, D. N. et al. Generation and dynamic manipulation of frequency degenerate polarization entangled Bell states by a silicon quantum photonic circuit. *Chip* **1**, 100001 (2022).
90. Tian, L. et al. Spatial multiplexing of atom-photon entanglement sources using feedforward control and switching networks. *Phys. Rev. Lett.* **119**, 130505 (2017).
91. Pu, Y. F. et al. Experimental realization of a multiplexed quantum memory with 225 individually accessible memory cells. *Nat. Commun.* **8**, 15359 (2017).
92. Jiang, N. et al. Experimental realization of 105-qubit random access quantum memory. *npj Quantum Inf.* **5**, 28 (2019).
93. Bluvstein, D. et al. Logical quantum processor based on reconfigurable atom arrays. *Nature* **626**, 58–65 (2024).
94. Bluvstein, D. et al. A quantum processor based on coherent transport of entangled atom arrays. *Nature* **604**, 451–456 (2022).
95. Wang, X. C. et al. Quantum interference between photons and single quanta of stored atomic coherence. *Phys. Rev. Lett.* **128**, 083605 (2022).
96. Du, W. et al. SU(2)-in-SU(1,1) nested interferometer for high sensitivity, loss-tolerant quantum metrology. *Phys. Rev. Lett.* **128**, 033601 (2022).
97. Proietti, M. et al. Experimental test of local observer independence. *Sci. Adv.* **5**, eaaw9832 (2019).
98. Hammerer, K., Sørensen, A. S. & Polzik, E. S. Quantum interface between light and atomic ensembles. *Rev. Mod. Phys.* **82**, 1041–1093 (2010).